



Population Genetics

P. Clote

Spring semester 2019-20
Boston College



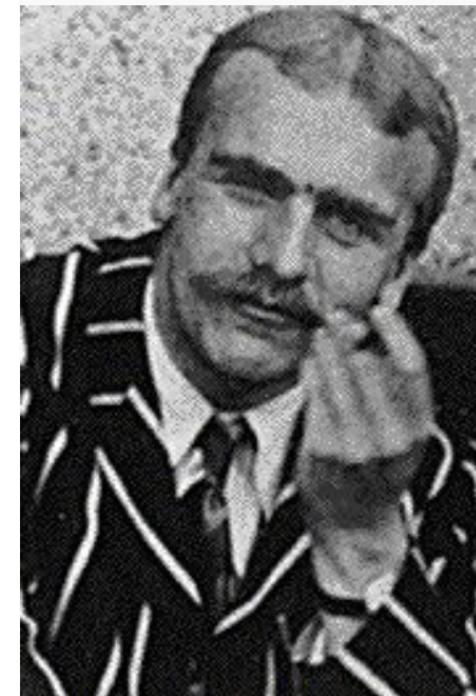
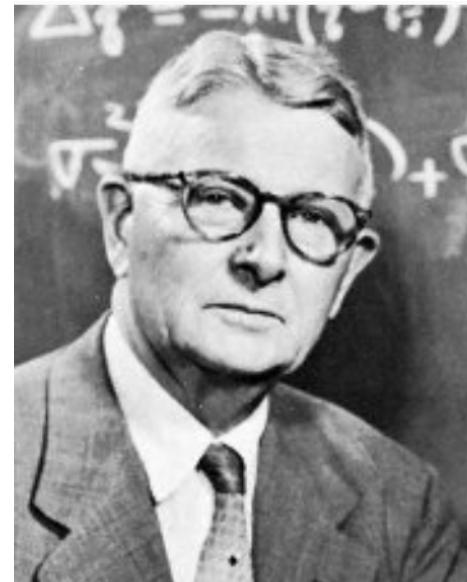
Chapter 1: Genetic Variation

Disclaimer: Uncredited images are taken from Population Genetics by J.H. Gillespie, or Molecular Population Genetics by Matthew W. Hahn, or are generated by code I have written.



Population genetics was founded by Fisher, Wright and Haldane

- ▶ Sir Ronald Aylmer Fisher FRS (17 February 1890 – 29 July 1962), a British statistician and geneticist (maximum likelihood, ANOVA)
- ▶ Sewall Green Wright (December 21, 1889 – March 3, 1988), an American geneticist, who worked on evolutionary theory
- ▶ John Burdon Sanderson Haldane FRS (5 November 1892 – 1 December 1964), a British-Indian scientist, who worked in physiology, genetics, evolutionary biology, and mathematics



(Image credit: Wikipedia)

- early experimental work in **molecular population genetics** done on **allozymes** (allele + enzyme), distinguishing protein variants by migration times using electrophoresis) – see Harris (1966) and Lewontin and Hubby (1966)
 - ▷ H. Harris (1966), “Enzyme polymorphisms in man”, Proceedings of the Royal Society B: Biological Sciences 164: 298-310
 - ▷ R.D.C. Lewontin and J.L. Hubby (1966), “A molecular approach to the study of genic heterozygosity in natural populations. II Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*”, Genetics 54: 595-609
- **Goal of population genetics:** develop mathematical models that quantitatively explain how polymorphisms arose in current populations, and how different mutation rates can result in different numbers of polymorphisms
- **polymorphism:** nucleotide or amino acid difference between individuals of same species. Locus is called **segregating site**
- **single nucleotide polymorphism (SNP):** nucleotide acid difference between individuals of same species. At a **di-allelic** site, there is a **major allele** and **minor allele**.
- **haplotype:** set of alleles found on a single DNA sequence – later, the term is used in slightly different sense in **“haplotype block”**
- **substitution:** nucleotide or amino acid difference between different species
- mathematical models of mutation
 - ▷ nucleotide mutates with equal probability to any of 3 other nucleotides
 - ▷ nucleotide transitions occur more often than nucleotide transversions
 - * transition: $A \leftrightarrow G$ (purines to purines) and $C \leftrightarrow T$ (pyrimidines to pyrimidines)
 - * transversion of purine to pyrimidine: $A \rightarrow C, T$ and $G \rightarrow C, T$
 - * transversion of pyrimidine to purine: $C \rightarrow A, G$ and $T \rightarrow A, G$
 - ▷ mutation rates may be 10-fold higher in CpG regions

- variation is detected by a **multiple sequence alignment** (see next slide), produced either manually by an expert, or computationally by software such as **Clustal Omega** (<https://www.ebi.ac.uk/Tools/msa/clustalo/>)
- in eukaryotes, the mutation rate is low, approximately 10^{-8} to 10^{-9} per site per generation, according to Lynch (2010)

M. Lynch (2010)
 Evolution of the mutation rate
 Trends in Genetics 26:345–352

- since mutation rate ρ is tiny and number of potential mutation sites N is enormous, the **number of mutations** within a given genomic region can be modeled by a **Poisson process**, and the **number of nucleotides between two successive mutations** can be modeled by the **geometric** (discrete) or **exponential** (continuous) distribution
- In Probability Appendix, it is shown that a **Poisson** random variable with mean μ satisfies

$$\begin{aligned} \text{Prob}[X = k] &= e^{-\mu} \cdot \frac{\mu^k}{k!} \\ E[X] &= \mu \\ V[X] &= \mu \end{aligned}$$

and that if X is a **geometrically** distributed random variable with success probability p , so that X represents the number of times a coin (with heads probability p) must be flipped until the first heads is obtained, then

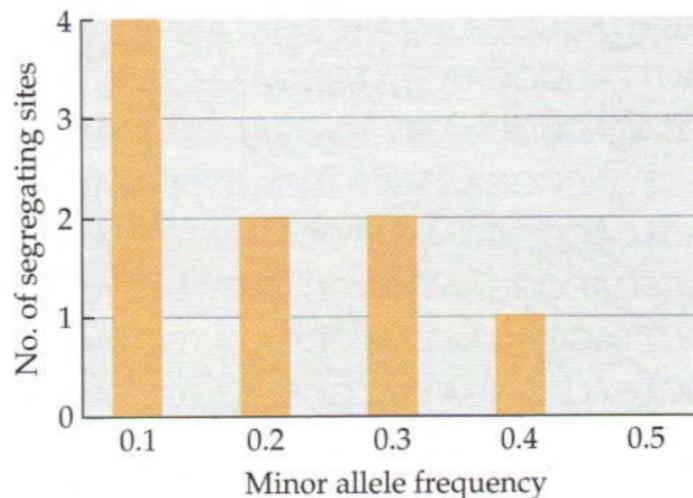
$$\begin{aligned} \text{Prob}[X = k] &= (1 - p)^{k-1} \cdot p \\ E[X] &= \frac{1}{p} \\ V[X] &= \frac{q}{p^2} \end{aligned}$$

Multiple sequence alignment with minor alleles indicated

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | C | C | G | A | T | A | T | G | C | T | A | G | T | A | | |
| G | C | T | T | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | G | T | T | G | T | A | | |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | G | T | A | G | A | A | | |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | C | T | A | G | A | A | | |
| G | C | T | C | A | C | C | G | G | G | A | T | T | A | T | G | G | A | T | A | T | G | C | T | A | G | T | A | | |
| G | C | T | C | A | C | C | G | G | A | A | T | T | A | T | G | C | G | A | T | A | T | G | C | T | A | G | A | | |
| G | C | T | T | A | C | C | G | G | A | A | T | T | A | T | C | C | G | A | T | A | T | G | C | T | A | G | A | | |
| G | C | T | C | A | C | C | G | G | A | A | T | T | A | T | G | C | G | C | T | A | T | G | G | T | A | G | A | | |
| G | C | T | C | A | C | C | G | G | A | A | T | T | A | T | C | C | G | A | T | A | T | G | C | T | A | G | T | A | |
| Total | | 2 | | 1 | | 2 | | 1 | | 3 | | 1 | | | | | | | 3 | | 1 | | 4 | | | | | | |
| N=10, L=29, S=9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| segregating sites: 4, 7, 10, 13, 16, 19, 24, 26, 28 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

- if an **outgroup** sequence is available from a related but distinct species, then the SNPs can be characterized as either **derived** or **ancestral**
- an ancestral SNP is one that is shared with the outgroup
- a derived SNP is one that is not shared with the outgroup
- multiple alignment above is of 10 genomic regions, each of length 29, with no **indels** (insertion or deletion, as indicated by a gap symbol —), 9 **segregating sites**: 4 sequences have 1 SNP, 2 sequences have 2 SNPs, 2 sequences have 3 SNPs, and 1 sequence has 4 SNPs
- since an outgroup sequence is not given, we show the **minor allele frequency** below the alignment

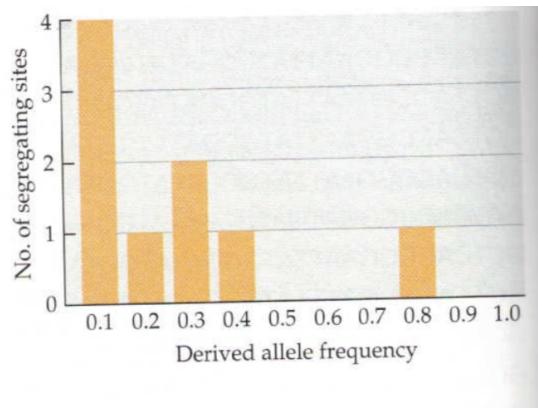
Minor allele frequency spectrum (“folded” spectrum)



- **x-axis:** number (or proportion) of sequences having a minor allele at segregating site
y-axis: number (or proportion) of segregating sites for which there are x sequences with the minor allele
- minor allele vector $(x_1, x_2, x_3, x_4, x_5)$ is $(4, 2, 2, 1, 0)$ where x_i is the number of segregating sites, for which exactly i many aligned sequences have a minor allele
- If S denotes the total number of segregating sites on a given chromosome (say chromosome 21, the smallest human chromosome containing ≈ 48 million bp), then between any 2 humans we expect about $\frac{1}{1000} \cdot 48 \times 10^6 = 48,000$ differences. It is for this reason that the x-axis of the allele frequency spectrum often corresponds to the **proportion** of SNPs rather than the number of SNPs.

Multiple sequence alignment with derived alleles indicated

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | # |
|-------|---|---|----------|---|---|----------|----------|----------|----------|----------|----|----------|----|----------|----|----|----|----|----------|----|----|----------|----|----------|----|----------|----|----|---|
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | C | C | G | A | T | A | T | G | C | T | A | G | T | A | | |
| G | C | T | T | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | G | T | T | G | T | A | | |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | G | T | A | G | A | A | | |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | C | T | A | G | A | A | | |
| G | C | T | C | A | C | C | G | G | G | A | T | G | A | T | G | C | G | A | T | A | T | G | C | T | A | G | T | A | |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | C | T | A | G | A | A | | |
| G | C | T | T | A | C | C | G | G | A | T | T | A | T | C | C | G | A | T | A | T | G | C | T | A | G | T | A | | |
| G | C | T | C | A | C | C | A | G | G | G | A | T | T | A | T | G | C | G | C | T | A | T | G | C | T | A | G | T | A |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | G | C | G | A | T | A | T | G | G | T | A | G | A | A | | |
| G | C | T | C | A | C | C | G | G | A | T | T | A | T | C | C | G | A | T | A | T | G | C | T | A | G | T | A | | |
| Total | | 8 | | 1 | | 2 | | 1 | | 3 | | 1 | | | | | | | | 3 | 1 | 4 | | | | | | | |



- **derived allele frequency spectrum** – in contrast to minor allele frequency spectrum
- **x-axis:** number (or proportion) of sequences having a derived allele at segregating site
y-axis: number (or proportion) of segregating sites for which there are x sequences with the derived allele
- derived allele vector $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})$ is $(4, 1, 2, 1, 0, 0, 0, 1, 0, 0)$ where x_i is the number of segregating sites, for which exactly i many aligned sequences have a derived allele

Variation within species (*D. melanogaster*)

1 g
 atg.tcg.ttt.act.ttg.acc.aac.aag.aac.gtg.att.ttc.gtt.gcc.ggt.ctg.gga.ggc.att.ggt
 Met.Ser.Phe.Thr.Leu.Thr.Asn.Lys.Asn.Val.Ile.Phe.Val.Ala.Gly.Leu.Gly.Gly.Ile.Gly
 61
 ctg.gac.acc.agc.aag.gag.ctg.ctc.aag.cgc.gat.ctg.aag.aac.ctg.gtg.atc.ctc.gac.cgc
 Leu.Asp.Thr.Ser.Lys.Glu.Leu.Leu.Lys.Arg.Asp.Leu.Lys.Asn.Leu.Val.Ile.Leu.Asp.Arg
 121
 att.gag.aac.ccg.gct.gcc.att.gcc.gag.ctg.aag.gca.atc.aat.cca.aag.gtg.acc.gtc.acc
 Ile.Glu.Asn.Pro.Ala.Ala.Ile.Ala.Glu.Leu.Lys.Ala.Ile.Asn.Pro.Lys.Val.Thr.Val.Thr
 181 t
 ttc.tac.ccc.tat.gat.gtg.acc.gtg.ccc.att.gcc.gag.acc.acc.aag.ctg.ctg.aag.acc.atc
 Phe.Tyr.Pro.Tyr.Asp.Val.Thr.Val.Pro.Ile.Ala.Glu.Thr.Thr.Lys.Leu.Leu.Lys.Thr.Ile
 241
 ttc.gcc.cag.ctg.aag.acc.gtc.gat.gtc.ttg.atc.aac.gga.gct.ggt.atc.ctg.gac.gat.cac
 Phe.Ala.Gln.Leu.Lys.Thr.Val.Asp.Val.Leu.Ile.Asn.Gly.Ala.Gly.Ile.Leu.Asp.Asp.His
 301
 cag.atc.gag.cgc.acc.att.gcc.gtc.aac.tac.act.ggc.ctg.gtc.aac.acc.acg.acg.gcc.att
 Gln.Ile.Glu.Arg.Thr.Ile.Ala.Val.Asn.Tyr.Thr.Gly.Leu.Val.Asn.Thr.Thr.Thr.Ala.Ile
 361 t a
 ctg.gac.ttc.tgg.gac.aag.cgc.aag.ggc.ggt.ccc.ggt.ggt.atc.atc.tgc.aac.att.gga.tcc
 Leu.Asp.Phe.Trp.Asp.Lys.Arg.Lys.Gly.Gly.Pro.Gly.Gly.Ile.Ile.Cys.Asn.Ile.Gly.Ser
 421 a
 gtc.act.gga.ttc.aat.gcc.atc.tac.cag.gtg.ccc.gtc.tac.tcc.ggc.acc.aag.gcc.gcc.gtg
 Val.Thr.Gly.Phe.Asn.Ala.Ile.Tyr.Gln.Val.Pro.Val.Tyr.Ser.Gly.Thr.Lys.Ala.Ala.Val
 481 a c g t
 gtc.aac.ttc.acc.agc.tcc.ctg.gcg.aaa.ctg.gcc.ccc.att.acc.ggc.gtg.acc.gct.tac.acc
 Val.Asn.Phe.Thr.Ser.Ser.Leu.Ala.Lys.Leu.Ala.Pro.Ile.Thr.Gly.Val.Thr.Ala.Tyr.Thr
 541 c
 gtg.aac.ccc.ggc.atc.acc.cgc.acc.acc.ctg.gtg.cat.aag.ttc.aac.tcc.tgg.ttg.gat.gtt
 Val.Asn.Pro.Gly.Ile.Thr.Arg.Thr.Thr.Leu.Val.His.Lys.Phe.Asn.Ser.Trp.Leu.Asp.Val
 601 t c c
 gag.ccc.cag.gtt.gct.gag.aag.ctc.ctg.gct.cat.ccc.acc.cag.cca.tcg.ttg.gcc.tgc.gcc
 Glu.Pro.Gln.Val.Ala.Glu.Lys.Leu.Leu.Ala.His.Pro.Thr.Gln.Pro.Ser.Leu.Ala.Cys.Ala
 661 a
 gag.aac.ttc.gtc.aag.gct.atc.gag.ctg.aac.cag.aac.gga.gcc.atc.tgg.aaa.ctg.gac.ctg
 Glu.Asn.Phe.Val.Lys.Ala.Ile.Glu.Leu.Asn.Gln.Asn.Gly.Ala.Ile.Trp.Lys.Leu.Asp.Leu
 721
 ggc.acc.ctg.gag.gcc.atc.cag.tgg.acc.aag.cac.tgg.gac.tcc.ggc.atc.
 Gly.Thr.Leu.Glu.Ala.Ile.Gln.Trp.Thr.Lys.His.Trp.Asp.Ser.Gly.Ile.

 **Representative** (genome information for reference and representative genomes)

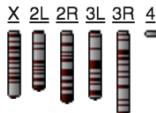
Reference genome:

-  *Drosophila melanogaster Release 6 plus ISO1 MT*

Submitter: The FlyBase Consortium/Berkeley Drosophila Genome Project/Celera Genomics

| Loc | Type | Name | RefSeq | INSDC | Size (Mb) | GC% | Protein | rRNA | tRNA | Other RNA | Gene | Pseudogene |
|-----|------|------|-------------|------------|-----------|------|---------|------|------|-----------|-------|------------|
| | Chr | X | NC_004354.4 | AE014298.5 | 23.54 | 42.5 | 5,390 | 16 | 26 | 596 | 2,675 | 68 |
| | Chr | 2L | NT_033779.5 | AE014134.6 | 23.51 | 41.8 | 5,694 | - | 41 | 962 | 3,501 | 47 |
| | Chr | 2R | NT_033778.4 | AE013599.5 | 25.29 | 42.6 | 6,051 | 100 | 100 | 721 | 3,628 | 46 |
| | Chr | 3L | NT_037436.4 | AE014296.5 | 28.11 | 41.7 | 5,880 | - | 50 | 786 | 3,466 | 34 |
| | Chr | 4 | NC_004353.4 | AE014135.4 | 1.35 | 35.4 | 295 | - | - | 38 | 111 | 6 |
| | Chr | 3R | NT_033777.3 | AE014297.3 | 32.08 | 42.6 | 7,205 | - | 80 | 894 | 4,202 | 36 |
| | Chr | Y | NC_024512.1 | CP007106.1 | 3.67 | 40.2 | 25 | - | - | 28 | 113 | 62 |
| | | MT | NC_024511.2 | KJ947872.2 | 0.02 | 17.8 | 13 | 2 | 22 | - | 37 | - |
| | Un | - | - | - | 6.16 | 41.1 | 6 | - | - | - | 5 | - |

 **Chromosomes**



Click on chromosome name to open Genome Data Viewer

- coding region (CDS with 768 bases, hence 256 amino acids) of alcohol dehydrogenase (ADH) from reference genome of *D. melanogaster*

<https://www.ncbi.nlm.nih.gov/genome?term=vih&cmd=DetailsSearch>

- Kreitman (1983) sequenced ADH gene in 11 fruit flies (*D. melanogaster*) from various places (Florida, Washington, Africa, Japan, France)
- 14 variable sites within coding region; i.e. 14 of the 768 columns in the multiple alignment of ADH **coding sequence (CDS)** showed variation. Variation at a DNA position within a species is called a single nucleotide polymorphism (SNP), for which there is usually a **major** allele and **minor** allele.
- 13 of 14 SNPs silent (same amino acid), and 1 replacement or non-silent SNP, in which nucleotide 578 of coding region either lysine AAG or threonine ACG.

Allele differences may be either **polymorphisms** (same species) or mutations that have been **fixed** (different species).

- SNPs may occur in both coding and non-coding portions of a protein-coding gene (i.e. 5'-UTR, 3'-UTR, introns), and elsewhere (“junk” DNA, or non protein-coding RNA genes, such as microRNA genes)
- On average there is SNP difference between 2 humans every 1000 nucleotides
— see

<https://ghr.nlm.nih.gov/primer/genomicresearch/snp>

for more information.

Summary of SNP locations in ADH of *D. melanogaster*

| Allele | 39 | 226 | 387 | 393 | 441 | 513 | 519 | 531 | 540 | 578 | 606 | 615 | 645 | 684 |
|-----------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Reference | T | C | C | C | C | C | T | C | C | A | C | T | A | G |
| Wa-S | . | T | T | . | A | A | C | . | . | . | . | . | . | . |
| Fl-1S | . | T | T | . | A | A | C | . | . | . | . | . | . | . |
| Af-S | . | . | . | . | . | . | . | . | . | . | . | . | . | A |
| Fr-S | . | . | . | . | . | . | . | . | . | . | . | . | . | A |
| Fl-2S | G | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Ja-S | G | . | . | . | . | . | . | . | T | . | T | . | C | A |
| Fl-F | G | . | . | . | . | . | . | G | T | C | T | C | C | . |
| Fr-F | G | . | . | . | . | . | G | T | C | T | C | C | . | . |
| Wa-F | G | . | . | . | . | . | G | T | C | T | C | C | . | . |
| Af-F | G | . | . | . | . | . | G | T | C | T | C | C | . | . |
| Ja-F | G | . | . | A | . | . | G | T | C | T | C | C | . | . |

Table from Gillespie book summarizing data from Kreitman 1983

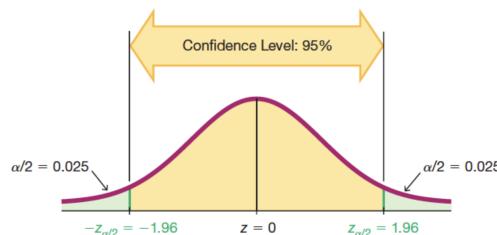
Confidence intervals for allele frequencies

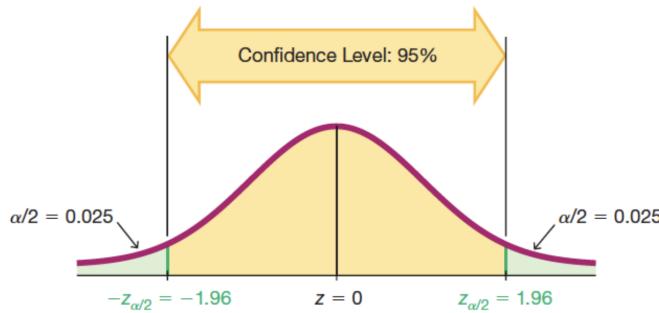
- Let E denote **margin of error** in **pointwise estimate of proportion** p , given a sample of size n .
- Kreitman's sample size: $n = 11$, where 6 of 11 flies code lysine (Lys=K, a positively charged amino-acid) at position 192 of amino acid sequence, while remaining 5 flies code threonine (Thr=T, a neutral and polar, or hydrophylic, amino-acid)
- in statistics, \hat{p} is a **pointwise estimate** for the (unknown) probability p in the population
- $\hat{p} = 6/11 = 0.55$ allele frequency for lysine at position 192 of *D. melanogaster* ADH protein
- $\sigma = \sqrt{\frac{p(1-p)}{n}}$ denotes the standard deviation, estimated in this case by $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- $z_{\alpha/2}$ denotes the value for which the area under the standard normal distribution to the right of $z_{\alpha/2}$ is equal to $\alpha/2$

$$\alpha/2 = \frac{1}{\sqrt{2\pi}} \int_{z_{\alpha/2}}^{\infty} \exp(-\frac{z^2}{2}) dz = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2}} \exp(-\frac{z^2}{2}) dz$$

$$z_{\alpha/2} = NORM.S.INV(1 - \alpha/2) \quad (\text{using Excel})$$

The Excel function *NORM.S.INV(x)* returns that value z such that the area under the standard normal density function from $-\infty$ to z equals x . If $\alpha/2$ is the area under the standard normal density function from $z_{\alpha/2}$ to $+\infty$, then $1 - \alpha/2$ is the area from $-\infty$ to $z_{\alpha/2}$. This explains the last line of the equation above.





■ $(1 - \alpha)$ confidence interval

$$E = z_{\alpha/2} \cdot \sigma = z_{\alpha/2} \cdot \sqrt{\hat{p}(1 - \hat{p})/n}$$

$$z_{\alpha/2} = 1.96 \text{ for } \alpha = 0.05$$

$$E = 1.96 \sqrt{0.55(0.45)/11} = 1.96 \cdot 0.15 = 0.294$$

with 95% confidence, lysine allele frequency $p = 0.55 \pm 0.294$

- $p = 0.55 \pm 0.294$ means that $p \in (\hat{p} - 0.294, \hat{p} + 0.294)$, i.e. $p \in (0.256, 0.844)$
- $p = 0.55 \pm 0.294$ with 95% confidence means that if we were to repeatedly sequence 11 different fruit flies (*D. melanogaster*), determine the lysine allele frequency \hat{p} in each sample, then the real but unknown (theoretical) lysine allele frequency p of ALL fruit flies will lie in 95% of the confidence intervals constructed
- Since

$$E = z_{\alpha/2} \cdot \sqrt{\hat{p}(1 - \hat{p})/n}$$

$$n = \frac{z_{\alpha/2}^2}{E^2} \cdot \hat{p}\hat{q}$$

one needs LARGE sample sizes for accurate allele frequency estimation (e.g. for $E = 0.01$, $\hat{p} = 1/2$, $n = 9604$)

McDonald-Kreitman test whether allele variation neutral

| | polymorphic | fixed | | |
|---------------|-------------|-------|---------|------------|
| nonsynonomous | Pn | Dn | | |
| synonomous | Ps | Ds | | |
| | | | | |
| | polymorphic | fixed | row sum | |
| replacement | 1 | 10 | 11 | |
| silent | 13 | 26 | 39 | |
| column sum | 14 | 36 | 50 | value of N |

- 2 × 2 contingency table for silent (synonomous) and replacement (nonsynonomous) mutations, where polymorphism indicates SAME species, and fixed indicates DIFFERENT species. There are 36 mutation sites when comparing *D. melanogaster* and closely related species *D. erecta*, using the data from Kreitman 1983.
- McDonald-Kreitman test is simply a **G likelihood-ratio test** of whether the ratio $D_n/D_s = P_n/P_s$, where equality is the **null hypothesis** – this test is an alternative to χ^2 test
- null hypothesis $H_0: P_n/P_s = D_n/D_s$
i.e. null hypothesis asserts that ratio of nonsynonymous to synonymous variation within a species $P_n/P_s = 1/13$ (0.07692) ratio of nonsynonymous to synonymous variation between species $D_n/D_s = 10/26$ (0.38462).
- χ^2 statistic $\sum_i \frac{(O_i - E_i)^2}{E_i} = 2.5012$
- p-value for χ^2 test with $df = 1$ is 0.114, so can NOT REJECT the null hypothesis for Gillespie's table – i.e. there is **no strong statistical evidence** in favor of positive selection for ADH in *D. melanogaster*

- however, the paper by McDonald and Kreitman

J.H. McDonald and M. Kreitman
 Adaptive protein evolution at the ADH locus in Drosophila
 Nature 351:652–654 (1991)

presents statistical evidence **in favor of positive selection** for ADH in *D. melanogaster* (tables shown in the next slide; computations done in my Excel spreadsheet)

- web server for McDonald-Kreitman test

<http://mkt.uab.es/mkt/mkt.asp>

- G-statistic

$$G = 2 \sum_{i=1}^4 O_i \ln \left(\frac{O_i}{\widehat{E}_i} \right)$$

$$G = 2 \left\{ P_n \ln \left(\frac{P_n}{\widehat{P}_n} \right) + D_n \ln \left(\frac{D_n}{\widehat{D}_n} \right) + P_s \ln \left(\frac{P_s}{\widehat{P}_s} \right) + D_s \ln \left(\frac{D_s}{\widehat{D}_s} \right) \right\}$$

- number of degrees of freedom (*df*) for G-test in this case is 3
- p*-value for G-test is computed from χ^2 distribution with *df* = 3
- Sequencing data from McDonald and Kreitman:

GenBank: X57361.1 - X57364.1 for *D. simulans*
 GenBank: X57365.1 - X57366.1 for *D. yakuba*

TABLE 1 Variable nucleotides from the coding region of the *Adh* locus in *D. melanogaster*, *D. simulans* and *D. yakuba*

| Con. | <i>D. melanogaster</i> | | | | | | | | | | <i>D. simulans</i> | | | | | | <i>D. yakuba</i> | | | | | | | | | | Rep1. | Fixed | | |
|------|------------------------|---|---|---|---|---|---|---|---|---|--------------------|---|---|---|---|---|------------------|---|---|---|---|---|---|---|---|-------|---------|---------|-------|-------|
| | a | b | c | d | e | f | g | h | i | j | k | l | a | b | c | d | e | f | a | b | c | d | e | f | g | h | i | j | k | l |
| 781 | G | T | T | T | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Repl. | Fixed |
| 789 | T | - | - | - | - | - | - | - | - | - | - | - | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | Syn. | Fixed | |
| 808 | A | - | - | - | - | - | - | - | - | - | - | - | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | Repl. | Fixed | |
| 816 | G | T | T | T | T | - | - | - | - | - | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 834 | T | - | - | - | - | - | - | - | - | - | C | C | - | - | C | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 859 | C | - | - | - | - | - | - | - | - | - | - | - | G | G | G | G | G | G | G | G | G | G | G | G | G | G | Repl. | Fixed | | |
| 867 | C | - | - | - | - | - | - | - | - | - | - | - | G | G | G | G | G | G | G | G | G | G | G | G | G | G | Syn. | 2 Poly. | | |
| 870 | C | T | T | T | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 950 | G | - | - | - | - | - | - | - | - | - | - | - | - | A | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 974 | G | - | - | - | - | - | - | - | - | - | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 983 | T | - | - | - | - | - | - | - | - | - | - | - | C | C | C | C | C | C | C | C | C | C | C | C | C | C | Syn. | Fixed | | |
| 1019 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1031 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1034 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1043 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1068 | C | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1089 | C | - | - | - | - | - | - | - | - | - | - | - | A | A | A | A | A | A | - | - | - | - | - | - | - | - | Repl. | Fixed | | |
| 1101 | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1127 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1131 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1160 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1175 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1178 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1184 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1190 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1196 | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1199 | C | - | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1202 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1203 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1229 | T | - | - | C | C | C | C | C | C | C | C | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1232 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1235 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1244 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1265 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1271 | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1277 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1283 | C | A | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1298 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | |
| 1304 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1316 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1425 | C | A | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1431 | T | C | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1443 | C | - | - | - | G | G | G | G | G | G | G | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1452 | C | - | - | - | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | |
| 1490 | A | - | - | - | - | C | C | C | C | C | C | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Repl. | Fixed | | |
| 1504 | C | T | T | T | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | | |
| 1518 | C | - | - | - | - | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1524 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | | |
| 1527 | C | T | T | T | T | T | T | T | T | T | T | T | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1530 | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1545 | T | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | | |
| 1548 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1551 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1555 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Repl. | Poly. | | | |
| 1557 | C | A | A | A | A | A | A | A | A | A | A | A | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1560 | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1573 | G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | C | Fixed | | | |
| 1581 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | | |
| 1584 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1590 | C | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | Syn. | Poly. | | | |
| 1596 | G | - | - | A | A | - | - | - | - | - | - | - | T | T | T | T | T | T | T | T | T | T | T | T | T | Syn. | Poly. | | | |
| 1611 | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Fixed | | | |
| 1614 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | 2 Poly. | | | |
| 1635 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |
| 1657 | A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Repl. | Fixed | | | |
| 1665 | C | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | Syn. | Poly. | | | |

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

| | Fixed | Polymorphic |
|-------------|-------|-------------|
| Replacement | 7 | 2 |
| Synonymous | 17 | 42 |

A G-test of independence (with the Williams correction for continuity)¹ was used to test the null hypothesis, that the proportion of replacement substitutions is independent of whether the substitutions are fixed or polymorphic. $G=7.43$, $P=0.006$.

- χ^2 statistic $\sum_i \frac{(O_i - E_i)^2}{E_i} = 8.1978$
- p -value for χ^2 test with $df = 1$ is 0.0042, so REJECT the null hypothesis for McDonald-Kreitman Nature 1991 data – i.e. there is **significant statistical evidence** in favor of positive selection for ADH in *D. melanogaster*
- as explained in my Excel spreadsheet, McDonald and Kreitman actually use the G-test for likelihood ratio with William's correction, which involves a test statistic

$$G = 2 \sum_{i=1}^4 O_i \cdot \ln \left(\frac{O_i}{E_i} \right)$$

which **should** follow a χ^2 distribution with $df = 3$ (not 1!)

- McDonald and Kreitman erroneously use $df = 1$ ($df = 1$ in χ^2 test, but $df = 3$ in G-test), hence they claim to have a p -value of 0.006. In point of fact, their p -value is 0.06!

Computing expected values for contingency table

Expected contingency table for silent (synonomous) and replacement (nonsynonomous) mutations, for polynomorphisms and fixed mutations, for use in χ^2 test

- Probability of being in first row is $p = \text{first row sum}/n$. Probability of being in first column is $q = \text{first col sum}/n$. Probability of being in first row and first col is $p \cdot q$
Expected number in first row and first col is $pq \cdot n$



$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 2.5$$

Computation of χ^2 test statistic is given in Excel demo.

- With a 2×2 contingency table, number df of degrees of freedom is $(\text{number rows} - 1) \times (\text{number columns} - 1)$, so $df = 1$.

Neutrality index (NI), positive and negative selection

- Neutrality index: $NI = \frac{P_n/P_s}{D_n/D_s}$
- **Negative** (or **purifying**) selection: when $\frac{D_n}{D_s} < \frac{P_n}{P_s}$
- **Positive** selection: when $\frac{D_n}{D_s} > \frac{P_n}{P_s}$
- Proportion of mutations driven by positive selection: $1 - NI = 1 - \frac{P_n/P_s}{D_n/D_s}$

If there is not sufficient statistical evidence to REJECT the null hypothesis $H_0 : P_n/P_s = D_n/D_s$, then we conclude that there is no selection, hence that mutations occurring in different species have been fixed by chance, not evolution. In a finite population, this is called **balancing selection**, where there is a balance between the introduction of mutations, due to errors in the polymerization of DNA, polymerase slippage, etc. and the removal of mutations by genetic drift.

Expected values in contingency table for χ^2 test using Kreitman 1983 data

| | | | | |
|-------------------|-------------|-------------------|-------------|--|
| Prob[row 1]=11/50 | 0.22 | Prob[col 1]=14/50 | 0.28 | |
| Prob[row2]=39/50 | 0.78 | Prob[col2]=36/50 | 0.72 | |
| | | | | |

EXPECTED contingency table

| | polymorphic | fixed | row sum |
|--------------------|--------------------|--------------|----------------------|
| replacement | 3.08 | 7.92 | 11 |
| silent | 10.92 | 28.08 | 39 |
| column sum | 14 | 36 | 50 value of N |

Table of values $(O-E)^2/E$

| | polymorphic | fixed |
|--------------------|--------------------|--------------|
| replacement | 1.404675325 | 0.546262626 |
| silent | 0.396190476 | 0.154074074 |

$$\chi^2 = 2.501202501$$

$$df = 1$$

$$p - \text{value} = 0.113759407$$

- It follows that the Adh mutation (in particular the non-silent lysine-threonine mutation) in *D. melanogaster* appears to be neutral, when using data from Kreitman 1983.

χ^2 distribution for a given df

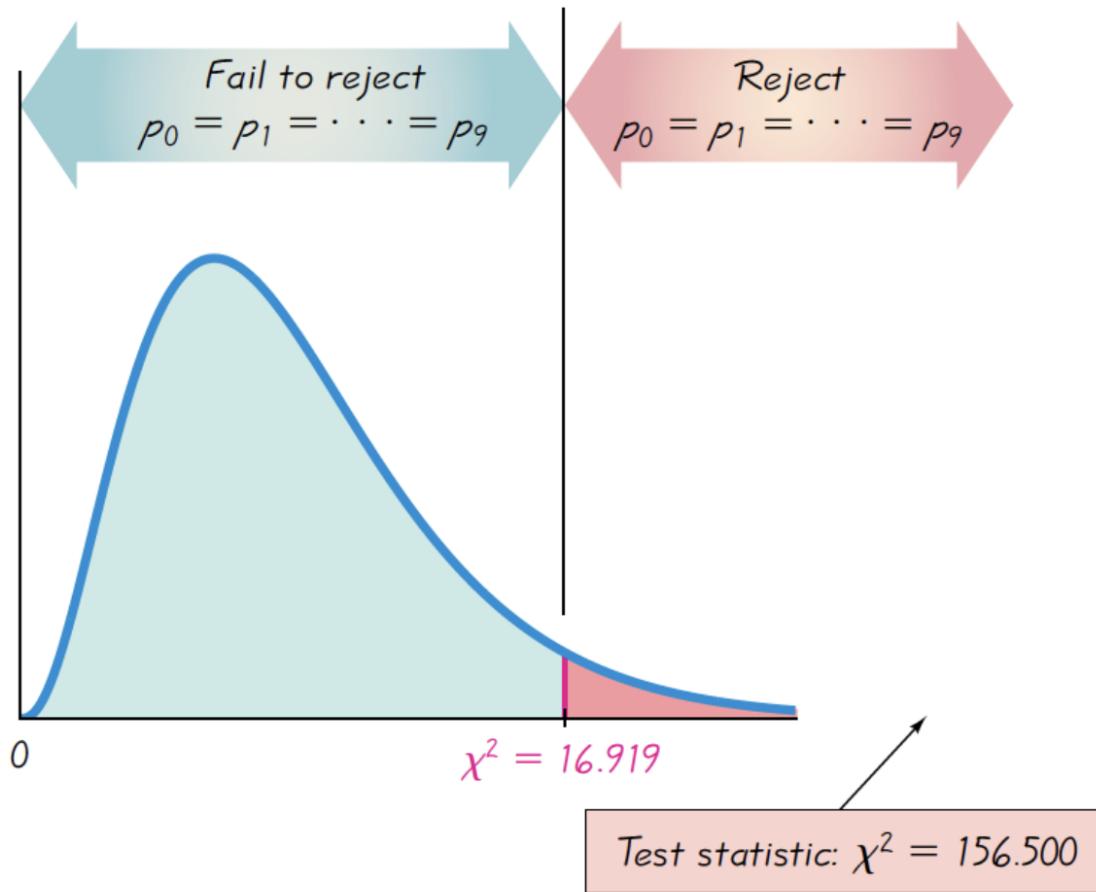


Image from Biostatistics by Triola and Triolo.

χ^2 table

TABLE A-4

Chi-Square (χ^2) Distribution

| Degrees of Freedom | Area to the Right of the Critical Value | | | | | | | | | |
|--------------------|---|-------|-------|-------|-------|--------|--------|--------|--------|--------|
| | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
| 1 | — | — | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.071 | 12.833 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |

Table from Biostatistics by Triola and Triolo. For Kreitman 1983 data (Gillespie book), test statistic $\chi^2 = 2.5$ with $df = 1$. From table, area to right of 3.841 is $0.05 = 5\%$. In order to REJECT null hypothesis H_0 , test statistic χ^2 must be greater than 3.841. Here we have insufficient evidence to reject the null hypothesis, hence conclude that, $P_n/P_s = D_n/D_s$, i.e. mutations have been fixed without natural selection, or evolution, but are instead neutral mutations.

Various notions of identical alleles

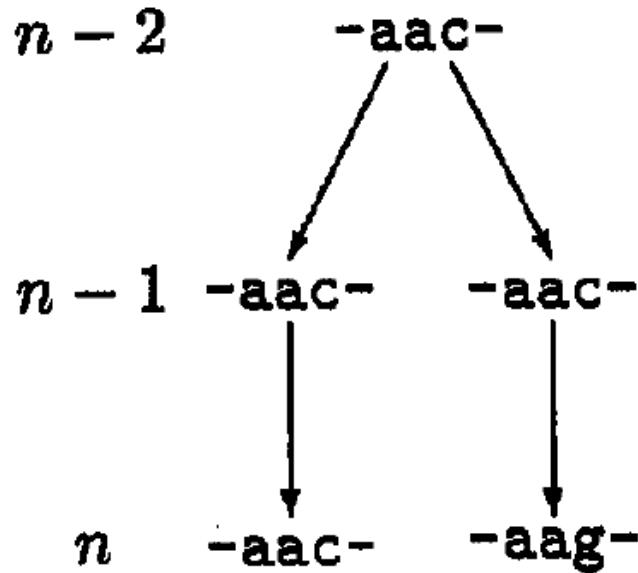
Alleles at the same locus may be identical or different with respect to (1) origin, (2) state, (3) descent:

- different by origin: alleles from different chromosomes (either different individuals or different chromosomes of same individual).
- different by state: alleles have different sequences (depends whether one is working with nucleotide or amino acid sequence data)
- identical by descent: alleles that share a (recent) common ancestor, i.e. the alpha-chain hemoglobin from you and from your cousin – since in theory any two alleles derive from a potentially ancient common ancestor, a recent common ancestor is understood for this case

In this context, it is useful to recall that homologous genes may be either **orthologous** or **paralogous**.

- orthologous** genes: common ancestor, e.g. horse alpha-chain hemoglobin and human alpha-chain hemoglobin
- paralogous** genes: gene duplication in species, e.g. human alpha-chain hemoglobin and human (or horse) myoglobin. The gene duplication event occurred in or before common ancestor of human and horse. Myoglobin and hemoglobin are paralogues.

Identical by origin, state and descent



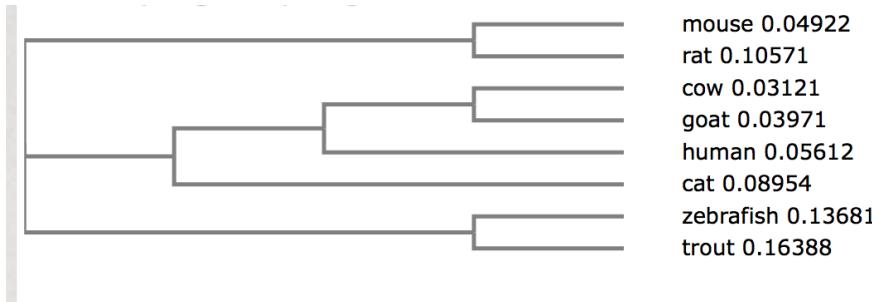
Example of two alleles that are identical by descent but not identical by state.

Allele and genotype frequencies

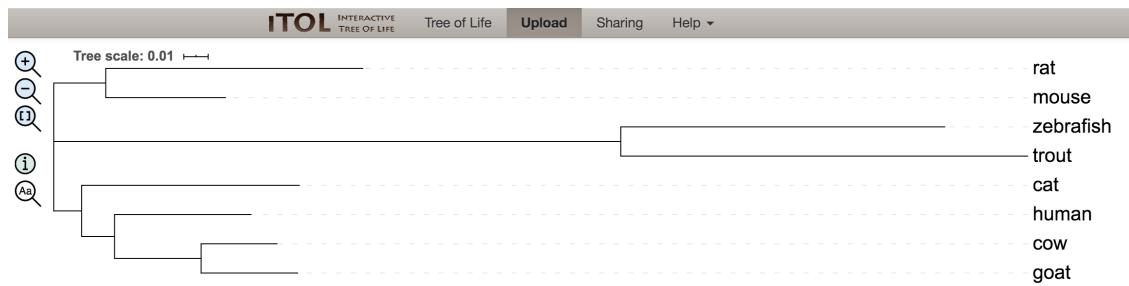
| | | |
|-----------|--|-----|
| mouse | MVLSGEDKSNIKAAWGKIGGGHGAELYGAEALERMFASFPTTKTYFPHFD-VSHGSAQVKGH | 59 |
| rat | MVL SADDKTNIKNCWGKIGGGHGYGEYEEALQRMFAAFPTTKTYFSHID-VSPGSAQVKAH | 59 |
| cow | -VLSAADKGKGVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFD-LSHGSAQVKGH | 58 |
| goat | MVL SAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFD-LSHGSAQVKGH | 59 |
| human | MVL SPADKTNVKAAWGKVGHAHGEYGAEALERMFLSFPTTKTYFPHFD-LSHGSAQVKGH | 59 |
| cat | -VLSAADKSNVKAACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFD-LSHGSAQVKAH | 58 |
| zebrafish | MSLSDTDKAVVKAIWAKISPKADEIGAEALARMLTVYPQTCKTYFSHWADLSPGSGPVKKH | 60 |
| trout | XSLSAKDKANVKAIWKGKILPKSDEIGEQALSRLMLVVYPQTAKYFSHWASVAPGSAPVKHH | 60 |
| | ** * * : * *.*: .: * :** * : :* * :*.* : : * *. ** * | |
| mouse | GKKVADALANAAGHLDLPGALSA SDLHAKL RVD PVNF KLLSHC LLVT LASHHPADFT | 119 |
| rat | GKKVADALAKAADHV E DLP G A L S T L S D L H A H K L R V D P V N F K L F L S H C L L V T L A C H H P G D F T | 119 |
| cow | GAKVAAA ALT K A V E H L D D L P G A L S E L S D L H A H K L R V D P V N F K L L S H S L L V T L A C H L P S D F T | 118 |
| goat | GEK VAAA ALT K A V G H L D D L P G T I L S D L H A H K L R V D P V N F K L L S H S L L V T L A C H L P N D F T | 119 |
| human | GKKVADAL TNAVAH VDDMPN A L S A L S D L H A H K L R V D P V N F K L L S H C L L V T L A A H L P A E F T | 119 |
| cat | GQKVADAL T Q A V A H M D D L P T A M S A L S D L H A Y K L R V D P V N F K L F L S H C L L V T L A C H H P A E F T | 118 |
| zebrafish | GKTIMGAVGEAVSKIDDLVGGI A L S E L H A F K L R V D P A N F K I L S H N V I V V I A M L F P A D F T | 120 |
| trout | GITIMNQI DDCVGHMDDLFGLK LSELHATKL R V D P T N F K I L A H N L I V V I A A Y F P A E F T | 120 |
| | * .. : ... ::* : :: * :*** * *****,* :* :* :* :* :* :* :* :* : | |
| mouse | PAVHASLDKFLASVSTVLTSKYR | 142 |
| rat | PAMHASLDKFLASVSTVLTSKYR | 142 |
| cow | PAVHASLDKFLANVSTVLTSKYR | 141 |
| goat | PAVHASLDKFLANVSTVLTSKYR | 142 |
| human | PAVHASLDKFLASVSTVLTSKYR | 142 |
| cat | PAVHASLDKFFSAVSTVLTSKYR | 141 |
| zebrafish | PEVHVSVDKFFFNNLALASEKYR | 143 |
| trout | PEIHL SVDKFLQQLALALA EKYR | 143 |
| | * :* * :*** : :: .*:.* | |

Multiple alignment of the amino acid sequences for 8 alpha-chain hemoglobins. Data obtained from NCBI, multiple alignment produced by Clustal Omega, a web server of the European Bioinformatics Institute (EBI) <https://www.ebi.ac.uk/Tools/msa/clustalo/>.

Allele and genotype frequencies



- Phylogeny of 8 alpha-chain hemoglobins produced by Clustal Omega, a web server of the European Bioinformatics Institute (EBI) <https://www.ebi.ac.uk/Tools/msa/clustalo/>.



- Phylogenetic tree Image produced by

<https://itol.embl.de/upload.cgi>

from the Newick format tree obtained from the “Phylogenetic tree” selection from Clustal Omega

Mendel's three laws

1. Law of segregation: For each trait, every individual has 2 alleles. During meiosis, these alleles separate (segregate); each gamete then contains exactly one of these alleles, so the offspring (zygote) contains one allele from each parent
2. Law of independent assortment: alleles for distinct traits are passed to offspring independently, e.g. eye color independent of height
3. Law of dominance: recessive alleles are masked by dominant alleles, e.g. eye color is brown when the individual contains both an allele for blue and brown eye color

These laws imply that in the absence of mutation and selection, allele frequencies in a population are constant. Compute genotype probabilities from allele probabilities;

$$p = \text{Prob}[A_1]$$

$$q = 1 - p = \text{Prob}[A_2]$$

$$\text{Prob}[A_1 A_1] = p^2$$

$$\text{Prob}[A_1 A_2] = 2pq$$

$$\text{Prob}[A_2 A_2] = q^2 = (1 - p)^2$$

Compute offspring generation allele probabilities from parent generation genotype probabilities:

$$p' = p^2 + pq = p(p + q) = p(1) = p$$

$$q' = q^2 + pq = q(q + p) = q(1) = q$$

Hardy-Weinberg theorem

- **Allele frequency** $p = \text{Prob}[A_1]$ is the proportion of A_1 alleles in the collection of $2N$ chromosomes for a population of size N . If the locus is **bi-allelic**, then allele frequency $q = \text{Prob}[A_2] = 1 - p$ is the proportion of A_2 alleles; if the locus has multiple alleles A_1, \dots, A_k , then p_i is the proportion of A_i alleles, and $\sum_{i=1}^k p_i = 1$. Below we focus on the bi-allelic case.
- **Genotype frequency** $\text{Prob}[A_1A_1]$ is the proportion of chromosome pairs homozygous for A_1 , and similarly $\text{Prob}[A_2A_2]$ is the proportion of chromosome pairs homozygous for A_2 . Genotype frequency $\text{Prob}[A_1A_2]$ is the proportion of chromosome pairs heterozygous for A_1 and A_2 (combining individuals whose paternal chromosome contains A_1 and whose maternal chromosome contains A_2 , and vice-versa).
- **assumptions** of Hardy-Weinberg theorem:
 - ▷ infinite population with random mating
 - ▷ no mutation
 - ▷ no selection
- computing genotype frequencies from allele frequencies
 - $\text{Prob}[A_1A_1] = p^2$
 - $\text{Prob}[A_1A_2] = 2pq$
 - $\text{Prob}[A_2A_2] = q^2$
- computing allele frequencies from genotype frequencies

$$p = \text{Prob}[A_1A_1] + \frac{\text{Prob}[A_1A_2]}{2}$$
$$q = \text{Prob}[A_2A_2] + \frac{\text{Prob}[A_1A_2]}{2}$$

■ **Hardy-Weinberg Theorem** states

- ▶ allele frequency never changes
- ▶ after the first generation, genotype frequencies never change, and are given by

$$\text{Prob}[A_1A_1] = p^2$$

$$\text{Prob}[A_1A_2] = 2pq$$

$$\text{Prob}[A_2A_2] = q^2$$

- For instance, if generation 0 has genotype frequencies $\text{Prob}[A_1A_1] = 1/4$ and $\text{Prob}[A_1A_2] = 3/4$, then $p = 1/4 + 3/8 = 5/8$ and $q = 3/8$, and beginning with generation 1, genotype frequencies are constant and equal to

$$\text{Prob}[A_1A_1] = p^2 = \frac{25}{64} = 0.390625$$

$$\text{Prob}[A_1A_2] = 2pq = 2 \cdot \frac{5}{8} \cdot \frac{3}{8} = \frac{15}{32} = 0.46875$$

$$\text{Prob}[A_2A_2] = q^2 = \frac{9}{64} = 0.140625$$

Hardy-Weinberg theorem

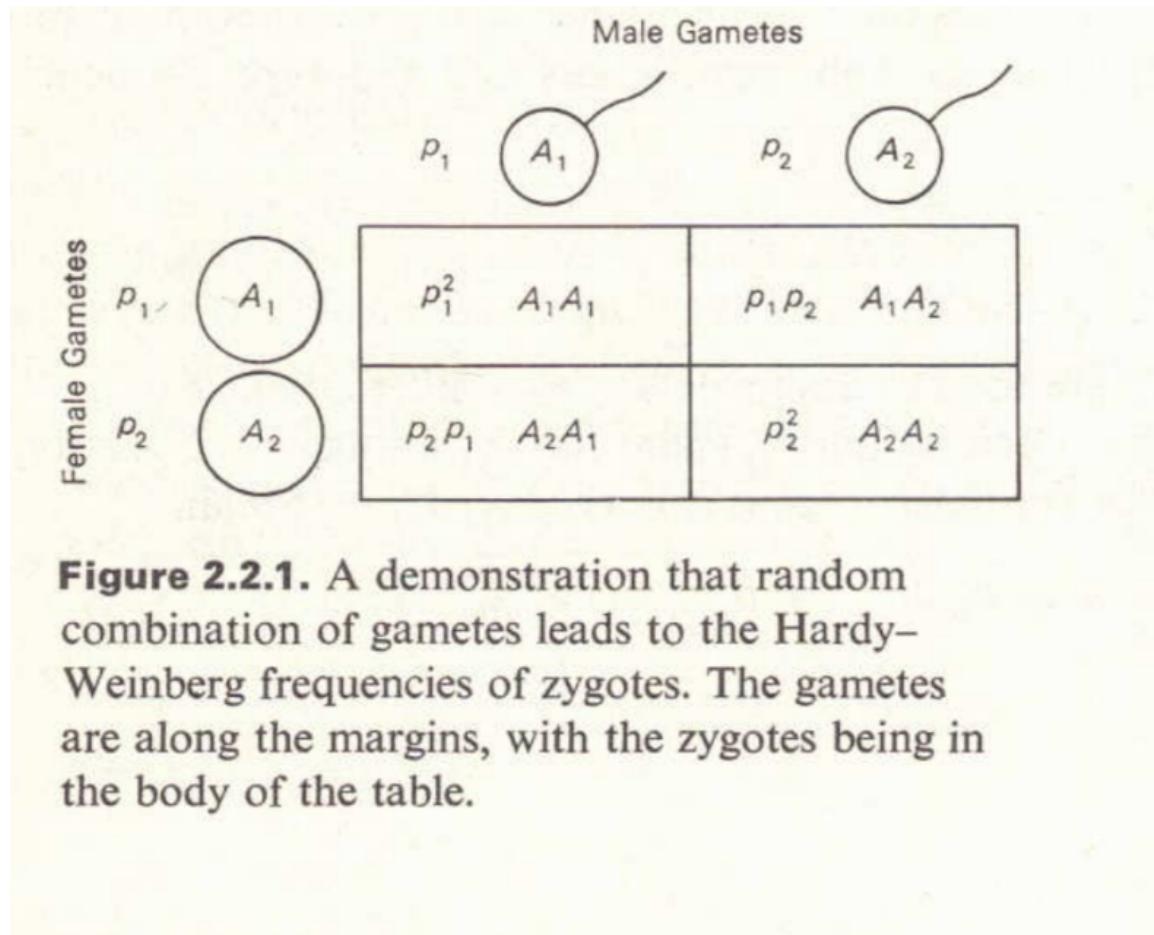


Figure 2.2.1. A demonstration that random combination of gametes leads to the Hardy-Weinberg frequencies of zygotes. The gametes are along the margins, with the zygotes being in the body of the table.

The Hardy-Weinberg theorem states the following. **In one generation, genotype frequencies obtain the same proportions as given by random combinations of gametes.**

(Image from **Introduction to Population Genetics Theory** by Crow and Kimura.)

More formally, assume that a gene has 2 alleles, A_1, A_2 , located on an autosome, and that a population has infinite size with allele probabilities $p = \text{Prob}[A_1]$, $q = \text{Prob}[A_2]$, where $p + q = 1$. Moreover, assume that both $p = \text{Prob}[A_1]$ in males and in females. If the population satisfies the random mating hypothesis, then in the absence of mutation and selection, the genotype probabilities of the offspring (and offspring of offspring, etc.) satisfies

$$\begin{aligned}x_{1,1} &= \text{Prob}[A_1A_1] = p^2 \\x_{1,2} &= \text{Prob}[A_1A_2] = 2pq \\x_{2,2} &= \text{Prob}[A_2A_2] = q^2 = (1 - p)^2\end{aligned}$$

Note that $p^2 + 2pq + q^2 = (p + q)^2 = 1$.

It can happen that initially, the A_1 allele frequency is initially different in males and females, i.e. $p_m \neq p_f$. If A_1 is located on one of the 22 pairs of autosomal chromosomes, then

$$\begin{aligned}p &= p_m p_f + \frac{p_m q_f}{2} + \frac{q_m p_f}{2} \\q &= q_m q_f + \frac{q_m p_f}{2} + \frac{p_m q_f}{2}\end{aligned}$$

so after one generation, males and females have the same A_1 allele frequency p , and after the second generation, we obtain Hardy-Weinberg equilibrium.

Example: After hordes of Martian males and Saturnian females invade earth, the blue-eared Martian males randomly mate with green-eared Saturnian females, where prior to mating both diploid extraterrestrials were homozygous for ear color. The A_1 [resp. A_2] allele for blue [resp. green] ears is not X-linked. Let p_m [resp. p_f] denote A_1 allele frequency for among Martian males [resp. Saturnian females], so $p_m = 1$, $p_f = 0$.

$$\begin{aligned} p &= p_m p_f + \frac{p_m q_f}{2} + \frac{q_m p_f}{2} \\ &= 1 \cdot 0 + \frac{1 \cdot 1}{2} + \frac{0 \cdot 0}{2} = 1/2 \end{aligned}$$

so $p = \frac{1}{2} = q$ in the F1 cross, while the F2 generation reaches Hardy-Weinberg genotype equilibrium of

$$Prob[A_1 A_1] = p^2 = 1/4$$

$$Prob[A_1 A_2] = 2pq = 1/2$$

$$Prob[A_2 A_2] = q^2 = 1/4$$

Weinberg's proof

Table 2.2.1. Derivation of the Hardy–Weinberg principle by combining parental genotypes in random proportions.

| MATING | FREQUENCY OF THIS MATING | PROGENY | | |
|------------------------|--|------------------------------------|--|------------------------------------|
| | | A_1A_1 | A_1A_2 | A_2A_2 |
| $A_1A_1 \times A_1A_1$ | $(P_{11})^2$ | P_{11}^2 | | |
| $A_1A_1 \times A_1A_2$ | $2(P_{11})(2P_{12})$ | $2P_{11}P_{12}$ | $2P_{11}P_{12}$ | |
| $A_1A_1 \times A_2A_2$ | $2(P_{11})(P_{22})$ | | $2P_{11}P_{22}$ | |
| $A_1A_2 \times A_1A_2$ | $(2P_{12})^2$ | P_{12}^2 | $2P_{12}^2$ | P_{12}^2 |
| $A_1A_2 \times A_2A_2$ | $2(2P_{12})(P_{22})$ | | $2P_{12}P_{22}$ | $2P_{12}P_{22}$ |
| $A_2A_2 \times A_2A_2$ | $(P_{22})^2$ | | | P_{22}^2 |
| Total | $(P_{11} + 2P_{12} + P_{22})^2$ $= 1$ | $(P_{11} + P_{12})^2$ $= p_1^2$ | $2(P_{11} + P_{12})(P_{12} + P_{22})$ $= 2p_1p_2$ | $(P_{12} + P_{22})^2$ $= p_2^2$ |

(Image from **Introduction to Population Genetics Theory** by Crow and Kimura.)

Example of blood type

Table 2.1.1. Numbers of persons of blood types M , MN , and N in a sample from a British population. Data from Race and Sanger (1962).

| PHENOTYPE GENOTYPE | M MM | MN MN | N NN | TOTAL |
|-----------------------|-------------|--------------|-------------|-------|
| NUMBER | 363 | 634 | 282 | 1,279 |
| FREQUENCY | 0.284 | 0.496 | 0.220 | 1.000 |

$$p_1 \text{ or } p_M = .284 + \frac{1}{2}(.496) = .532$$

$$p_2 \text{ or } p_N = .220 + \frac{1}{2}(.496) = .468$$

Table 2.2.2. Comparison of the observed proportion of MN blood types with the proportions expected with random mating. The data are from Table 2.1.1; $p_M = .532$ and $p_N = .468$.

| GENOTYPE | MM | MN | NN | TOTAL |
|---------------------|-------|-------|-------|--------|
| Expected proportion | .283 | .498 | .219 | 1.000 |
| Expected number | 362.0 | 636.9 | 280.1 | 1279.0 |
| Observed number | 363 | 634 | 282 | 1279 |

$$\chi^2 = 0.029, \text{ Prob} = .87$$

The χ^2 method is described in the appendix. In this instance there is only one degree of freedom, rather than the two that might have been expected. This is because the gene frequency has been estimated from the data, thereby reducing the number of degrees of freedom by one.

Hardy-Weinberg, when X-linked and $p_m \neq p_f$

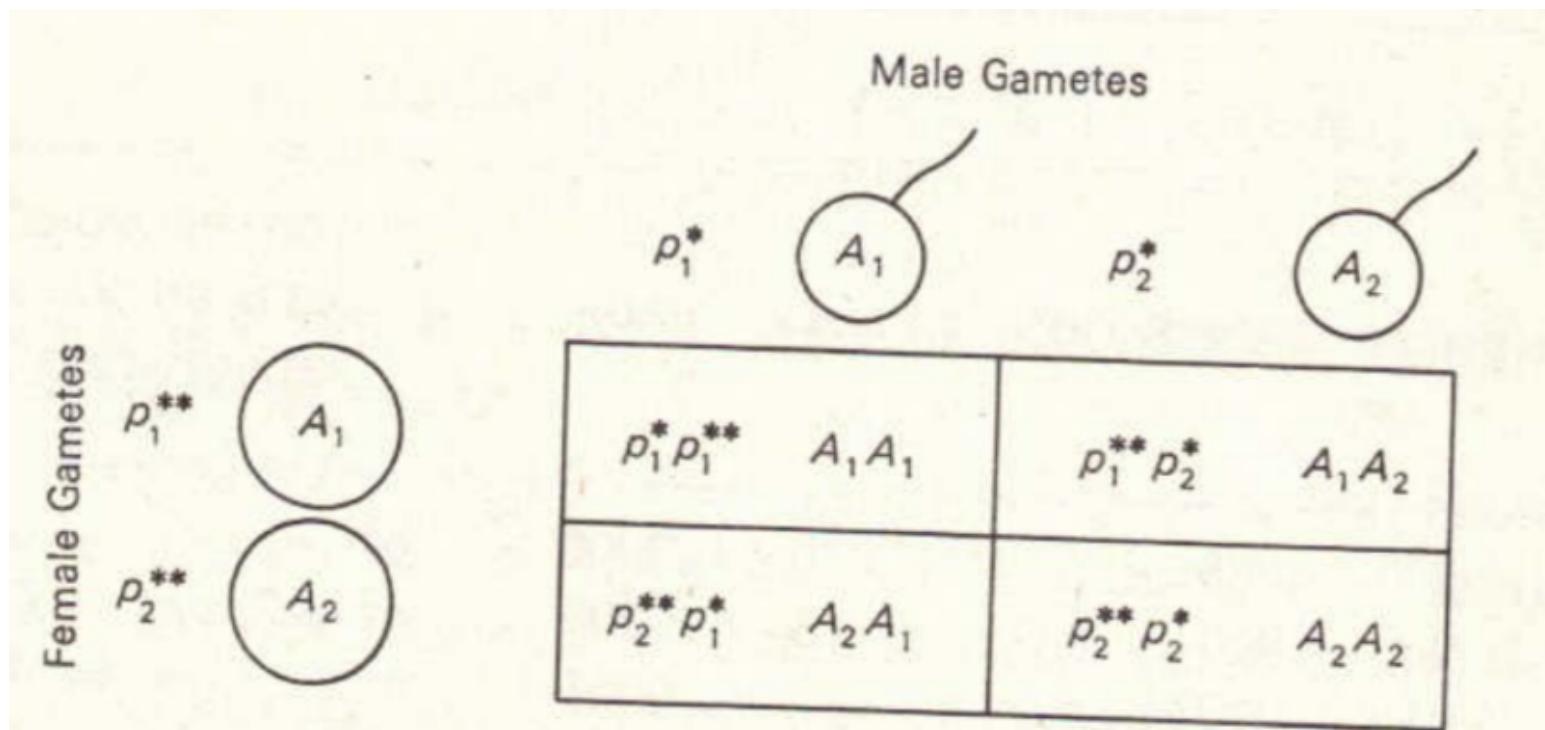


Figure 2.5.1. Random combination of gametes when the allele frequencies are different in the two sexes. The single and double asterisks refer to male and female frequencies, respectively.

$$\lim_{t \rightarrow \infty} |p_m - p_f| = 0$$

Let $p_m(t)$ [resp. $p_f(t)$] denote allele frequency of A_1 in males [resp. females] at time t , where initial frequencies given at $t = 0$. Note that

1. Males obtain copy of A_1 on X-chromosome from mother, so

$$p_m(t) = p_f(t - 1)$$

2. Females obtain copy of A_1 on X-chromosome from mother or father, so

$$p_f(t) = \frac{p_f(t - 1) + p_m(t - 1)}{2}$$

3. The average frequency of allele A_1 per generation is determined by contribution of one copy per male and two copies per female (males are XY, females are XX).

$$\bar{p}(t) = E[p(t)] = \frac{1}{3}p_m(t) + \frac{2}{3}p_f(t)$$

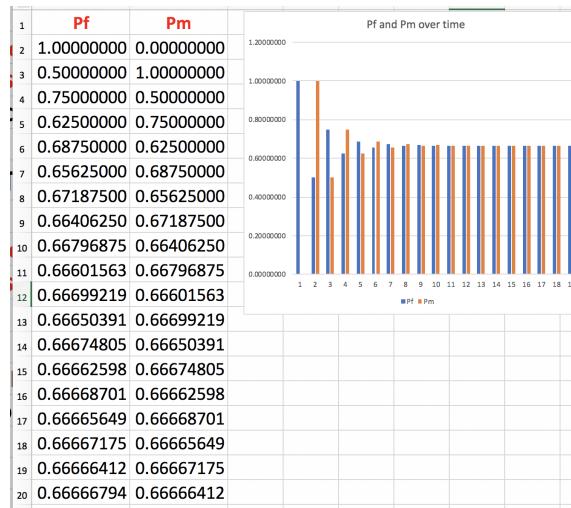
- **Claim:** Expected allele frequency is constant, i.e. for all $t, t' \geq 0$, $\bar{p}(t) = \bar{p}(t')$

- **Proof:** Using previous properties (1,2,3), prove by induction that

$$\begin{aligned}
 \bar{p}(t) &= \frac{1}{3}p_m(t) + \frac{2}{3}p_f(t) \\
 &= \frac{1}{3}p_f(t-1) + \frac{2}{3}\frac{p_f(t-1) + p_m(t-1)}{2} \\
 &= \frac{2}{3}p_f(t-1) + \frac{1}{3}p_m(t-1) \\
 &= \frac{1}{3}p_m(t-1) + \frac{2}{3}p_f(t-1) = \bar{p}(t-1)
 \end{aligned}$$

It follows that in the absence of mutation and selection, autosomal allele frequency $p = \text{Prob}[A_1]$ is constant, while for X-linked allele A_1 , the expected allele frequency \bar{p} of A_1 is constant.

- For a gene located on the X-chromosome, the graph below shows male A_1 allele frequencies p_M and female A_1 allele frequencies p_f over several generations for a population, where initially $p_f = 1$ (all females are homozygous A_1A_1 in the initial population) and $p_m = 0$ (all male X-chromosomes contain A_2 in the initial population).



Claim: $p_f(t) - p_f(t-1) = -\left[\frac{p_f(t-1) - p_f(t-2)}{2}\right]$

Proof: Using previous properties (1,2)

$$\begin{aligned} p_f(t) - p_f(t-1) &= \frac{p_m(t-1) + p_f(t-1)}{2} - p_f(t-1) \\ &= \frac{p_f(t-2)}{2} + \frac{p_f(t-1)}{2} - p_f(t-1) \\ &= -\left[\frac{p_f(t-2)}{2} - \frac{p_f(t-1)}{2}\right] = -\left[\frac{p_f(t-1) - p_f(t-2)}{2}\right] \end{aligned}$$

It follows that the sign alternates every generation, and that

$$\begin{aligned} |p_f(t) - p_f(t-1)| &= \left| \frac{p_f(t-1) - p_f(t-2)}{2} \right| \\ &= \left| \frac{p_f(t-k) - p_f(t-k-1)}{2^k} \right| \end{aligned}$$

so convergence is fast. Since $|p_f(1) - p_f(0)| \leq 1$, we have $|p_f(t) - p_f(t-1)| \leq (\frac{1}{2})^{t-1}$. Since $2^{10} = 1024 \approx 1000$, by the 10-th generation, male and female frequencies of A_1 are within 0.001 of each other, and both converge to the limit $\bar{p} = \frac{p_m(0)}{3} + \frac{2p_f(0)}{3}$.

Does the data on color-blind persons fit model?

Table 2.4.1. Frequencies of color blindness among school children in Oslo. Data from Waaler (1927).

| MALES | | | FEMALES | | |
|-------------|-----------------|----------|------------|--|-------|
| NUMBER | PROPORTION | OBSERVED | | EXPECTED ON BASIS OF MALE FREQUENCY | |
| | | NUMBER | PROPORTION | | |
| Color-blind | .0801 = p | 725 | .0801 | 40 | .0044 |
| Normal | .9199 = $1 - p$ | 8324 | .9199 | 9032 | .9956 |

(Image from **Introduction to Population Genetics Theory** by Crow and Kimura.)

- color-blindness allele is recessive and located on X-chromosome
- The proportion of men in the sample who are color-blind is $\frac{725}{725+8324} = 0.0801$. If this is the color-blindness allele frequency p in the population, then the proportion of color-blind women should be

$$p^2 \approx 0.0801^2 = 0.0064$$

However, the proportion of women in the sample who are color blind is $\frac{40}{40+9032} = 0.0044$. Do the data fit the model of a single color-blind allele on the X-chromosome with allele frequency p ?

- We want to apply the χ^2 goodness-of-fit test to data from the table, but this requires **expected values**.
- How can one compute the **expected** values for the proportion m of males and allele frequency p in the population (not just in the sample)?

Maximum likelihood

- note that the estimate \hat{p} completely ignores the female data, which is more complicated since color-blindness in females is a recessive trait on the X-chromosomes
- better approach: compute **maximum likelihood** estimates \hat{m} for proportion of males and \hat{p} for color-blindness allele frequency in population (not just in sample)
- suppose a coin has (unknown) heads probability p , and the result from flipping it n times is

$$x_1, \dots, x_n$$

where $x_i = 1$ for *heads* and $x_i = 0$ for *tails*

- the **maximum likelihood** estimate \hat{p} for the unknown heads probability p is determined by **maximizing** the **likelihood** $L(p)$, given by

$$L(p) = p^{x_1}(1-p)^{(1-x_1)} \cdots p^{x_n}(1-p)^{(1-x_n)}$$

- since the logarithm is monotonic, meaning that for positive x, y

$$x \leq y \Leftrightarrow \ln(x) \leq \ln(y)$$

it follows that the likelihood $L(p)$ obtains a maximum at the same position p where the **log likelihood** $\ln L(p)$ obtains a maximum

- for the coin flipping data, determine the value p which maximizes $\ln L(p)$ by setting the derivative $\frac{d}{dp} \ln L(p)$ to zero

$$\ln L(p) = x_1 \ln p + (1-x_1) \ln(1-p) + \cdots + x_n \ln p + (1-x_n) \ln(1-p)$$

$$\frac{d}{dp} \ln L(p) = \frac{x_1}{p} - \frac{1-x_1}{1-p} + \cdots + \frac{x_n}{p} - \frac{1-x_n}{1-p}$$

$$\frac{x_1}{p} + \cdots + \frac{x_n}{p} = \frac{1-x_n}{1-p} + \cdots + \frac{1-x_n}{1-p}$$

$$\frac{1}{p} \cdot \sum_{i=1}^n x_i = \frac{1}{1-p} \cdot \sum_{i=1}^n (1-x_i) = \frac{1}{1-p} \cdot \left(n - \sum_{i=1}^n x_i \right)$$

- if there were $m = \sum_{i=1}^n x_i$ heads and $n - m = n - \sum_{i=1}^n x_i$ tails, then

$$\frac{m}{p} = \frac{n-m}{1-p}$$

$$(1-p)m = p(n-m)$$

$$m - pm = pn - pm$$

$$p = \frac{m}{n}$$

- in this case, the maximum likelihood estimate is what you should expect – namely, just the proportion of heads obtained in n coin flips
- however, most applications of maximum likelihood are far from evident, and this method is of great importance in population genetics and computational biology

Maximum likelihood estimate of color-blindness allele frequency p and proportion of males in the population

$A = 725$ number of color-blind males

$B = 8324$ number of normal males

$C = 40$ number of color-blind females

$D = 9032$ number of normal females

p = (unknown) allele frequency for color-blindness in population

m = (unknown) frequency of males in population

$$L(m, p) = [mp]^A \cdot [m(1 - p)]^B \cdot [(1 - m)p^2]^C \cdot [(1 - m)(1 - p^2)]^D$$

Warning: p^2 and $1 - p^2$ for females. **Note:** $1 - p^2$ and **not** $(1 - p)^2$

$$\begin{aligned} \ln L(m, p) &= A \ln[mp] + B \ln[m(1 - p)] + C \ln[(1 - m)p^2] + D \ln[(1 - m)(1 - p)(1 + p)] \\ &= [A \ln m + B \ln m + C \ln(1 - m) + D \ln(1 - m)] + \\ &\quad [A \ln p + B \ln(1 - p) + 2C \ln p + D \ln(1 - p) + D \ln(1 + p)] \end{aligned}$$

$$\frac{\partial}{\partial m} \ln L(m, p) = \frac{A + B}{m} - \frac{C + D}{1 - m} \quad \text{in } \mathbf{\text{partial differentiation}}, \text{ other variables assumed constant}$$

$$\frac{\partial}{\partial m} \ln L(m, p) = 0 \Leftrightarrow (1 - m)(A + B) = m(C + D) \Leftrightarrow A + B = m(A + B + C + D)$$

$$\Leftrightarrow m = \frac{A + B}{A + B + C + D} = \frac{9049}{18121} = 0.4994 \quad (\text{proportion of males in sample, as expected})$$

$$\frac{\partial}{\partial p} \ln L(m, p) = \frac{A + 2C}{p} - \frac{B + D}{1 - p} + \frac{D}{1 + p}$$

$$\frac{\partial}{\partial p} \ln L(m, p) = 0 \Leftrightarrow p = \frac{-B + \sqrt{B^2 + 4(A + 2C)[A + B + 2(C + D)]}}{2[A + B + 2(C + D)]} = 0.0772$$

Example of color blindness

Table 2.4.2. Statistical analysis by the χ^2 method of the data in Table 2.4.1. The expected numbers are computed on the assumption of random mating by the method of maximum likelihood.

| | OBSERVED NUMBER | EXPECTED NUMBER | DEVIATION | DEVIATION ² EXPECTED |
|---------------------|--------------------|--------------------|-----------|------------------------------------|
| Color-blind males | 725 | 698.6 | 26.4 | 1.00 |
| Normal males | 8,324 | 8,350.4 | -26.4 | 0.08 |
| Color-blind females | 40 | 54.1 | -14.1 | 3.67 |
| Normal females | 9,032 | 9,017.9 | 14.1 | 0.02 |
| Total | 18,121 | 18,121.0 | | 4.77 |
| | $\chi^2_1 = 4.77$ | | $P = .03$ | |

(Table from **Introduction to Population Genetics Theory** by Crow and Kimura.)

- a p -value of 0.03 means that the **null hypothesis** should be rejected at 95% confidence level, where the null hypothesis is that the data support color-blindness being due to a single recessive allele on the X-chromosome
- explanation for the deviation from the model is due to fact that there are 4 types of color-blindness:
 - ▷ **protanopia** or red color-blindness (red looks green),
 - ▷ **protanomaly** or partial red color-blindness,
 - ▷ **deutanopia** or green color-blindness (green looks red),
 - ▷ **deutanomaly** or partial green color-blindness.
- note that from the sample data,

$$\hat{p}_f^2 = \frac{40}{40 + 9032} = 0.0044$$

$$\hat{p}_f = \sqrt{\hat{p}_f^2} = 0.0664$$

$$\hat{p}_m = \frac{725}{725 + 8324} = 0.0801$$

$$\bar{p} = \frac{\hat{p}_m}{3} + \frac{2\hat{p}_f}{3} = 0.07097$$

- thus the (weighted) average allele frequency p is **not** very close to the maximum likelihood value of 0.0772, which latter is “correct” manner of estimating the allele frequency p for the X-linked color-blindness trait. The reason these two values are not close is because color-blindness is not due to a **single** X-linked allele – since Hardy-Weinberg equilibrium has certainly been reached, if a single X-linked allele were responsible for color blindness, then we would have

$$p_m = \bar{p} = \frac{p_m}{3} + \frac{2p_f}{3}$$

$$p_f = \bar{p} = \frac{p_m}{3} + \frac{2p_f}{3}$$

which is not the case for our data.

Consequences and extensions of Hardy-Weinberg

A novel mutation (**rare** allele) A_2 is most often found in **heterozygotic** form, since ratio of A_2A_2 homozygotes to A_1A_2 heterozygotes is

$$\frac{q^2}{2pq} = \frac{q}{2p} \approx \frac{q}{2}$$

since $p \approx 1$. If $q = 0.0001 = 10^{-4}$, then A_2 is 20,000 times more likely to be found in a heterozygote than homozygote.

With multiple alleles A_1, \dots, A_k , Hardy-Weinberg has trivial extension, where **heterozygosity** H satisfies $H = 1 - G$, where G denotes homozygosity, and

$$H = 1 - G = 1 - \sum_{i=1}^k \text{Prob}[A_iA_i] = 1 - \sum_{i=1}^k p_i^2$$

$$p_i = \text{Prob}[A_iA_i] + \frac{1}{2} \cdot \sum_{j \neq i} \text{Prob}[A_iA_j] = p_i^2 + \sum_{j \neq i} p_i p_j$$

HW equilibrium for 3 alleles of alkaline phosphatase

| Genotype | Number | Frequency | Expected |
|----------|--------|-----------|----------|
| SS | 141 | 0.4247 | 0.4096 |
| SF | 111 | 0.3343 | 0.3507 |
| FF | 28 | 0.0843 | 0.0751 |
| SI | 32 | 0.0964 | 0.1101 |
| FI | 15 | 0.0452 | 0.0471 |
| II | 5 | 0.0151 | 0.0074 |
| Total | 332 | 1.0000 | 1.0000 |

Table 1.2: The frequencies of alkaline phosphatase genotypes in a sample from the English people. The expected Hardy-Weinberg frequencies are given in the fourth column. The data are from Harris (1966).

(Table from **Population Genetics** by Gillespie.)

χ^2 test for fit of Hardy-Weinberg model

| Genotype | Number | Frequency | Expected freq | Expected num | $(O-E)^2/E$ |
|-------------|------------|-----------|---------------|--------------|----------------|
| SS | 141 | 0.4247 | 0.4097 | 136.0128 | 0.1829 |
| SF | 111 | 0.3343 | 0.3509 | 116.4910 | 0.2588 |
| FF | 28 | 0.0843 | 0.0751 | 24.9428 | 0.3747 |
| SI | 32 | 0.0964 | 0.1099 | 36.4834 | 0.5510 |
| FI | 15 | 0.0452 | 0.0471 | 15.6235 | 0.0249 |
| II | 5 | 0.0151 | 0.0074 | 2.4465 | 2.6651 |
| Sums | 332 | 1 | 1 | 332 | 4.05733 |

From observed genotype frequencies for slow (S), intermediate (I) and fast (F) electrophoresis speeds, estimate allele frequencies S, I, F

$$S = 0.6401 \text{ computed by } S = SS + SI/2 + SF/2$$

$$F = 0.2741 \text{ computed by } F = FF + FI/2 + SF/2$$

$$I = 0.0858 \text{ computed by } I = II + FI/2 + FI/2$$

| | | | |
|----------|---|--|--|
| 4.057328 | Chisquare test statistic | | |
| 0.05 | alpha | | |
| 3 | degrees of freedom | | |
| 7.814728 | critical value to reject null hypothesis H0 obtained by CHISQ.INV.RT(alpha,df) | | |
| 0.25534 | p-value obtained by CHISQ.DIST.RT(testStatistic,df) | | |

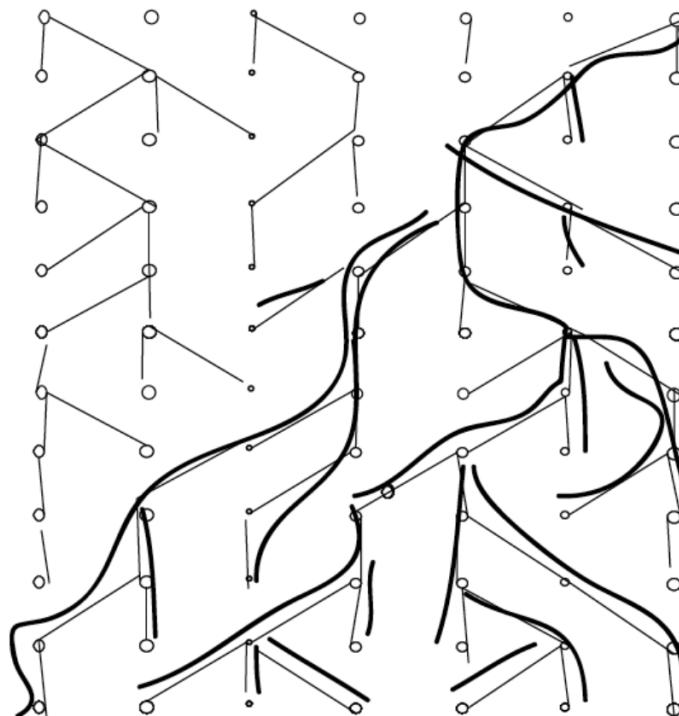
- Compute expected frequencies $E[SS] = S^2$, $E[SI] = SI$, etc.
- Compute expected numbers $N \cdot E[SS]$, $N \cdot E[SI]$, etc.,
- Compute test statistic $\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i}$ where there are $6-3=3$ degrees of freedom, since the 3 allele frequencies were estimated from the (observed) data
- Null hypothesis H_0 is that Hardy-Weinberg fits data. Since test statistic χ^2 is NOT GREATER than critical value for $\alpha = 0.05$, do NOT REJECT H_0 .
- Alternatively, since p -value NOT LESS than $\alpha = 0.05$, do NOT REJECT H_0 . Recall that p -value is area to right of test statistic under χ^2 density function with $df = 3$.
- we thus conclude that the data fit the Hardy-Weinberg equilibrium



Chapter 2: Genetic drift

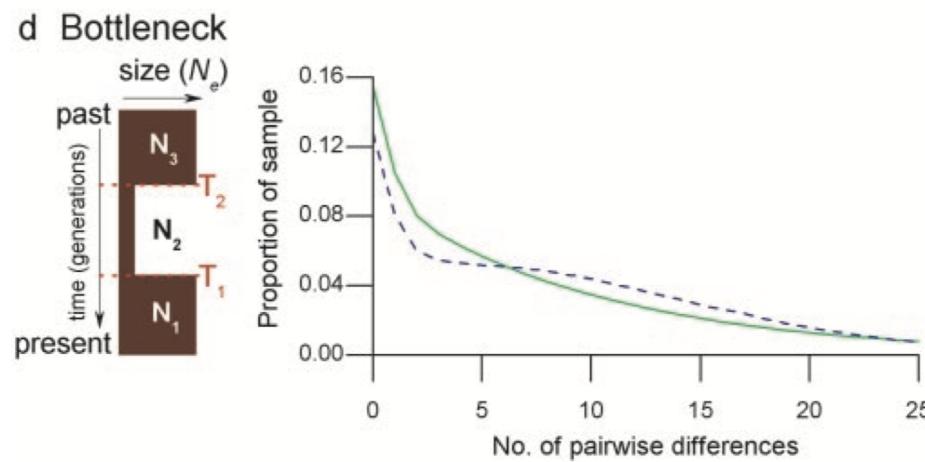
Loss of genetic diversity due to constant (finite) population size

- In absence of mutation and selection, Hardy-Weinberg law, which assumes the random mating hypothesis in an infinite population, implies that allele frequencies **do not change** over time
- In absence of mutation and selection, real populations may have (finite) constant size due to balance with the environment, and this implies that alleles drift towards **fixation** or **extinction**
- small population sizes may give rise to fixation of deleterious alleles and loss of advantageous alleles, caused by genetic drift



Neutral theory and genetic drift

- Genetic drift in a (finite) constant-size population is caused by random factors, such as
 - stochastic variation of number of offspring
 - variations in segregation (for diploid species)
- neutral theory of Kimura and others states that evolution due primarily to genetic drift, rather than natural selection (theory largely accepted for silent mutations in coding regions, but not accepted for replacement (non-silent) mutations in coding regions)



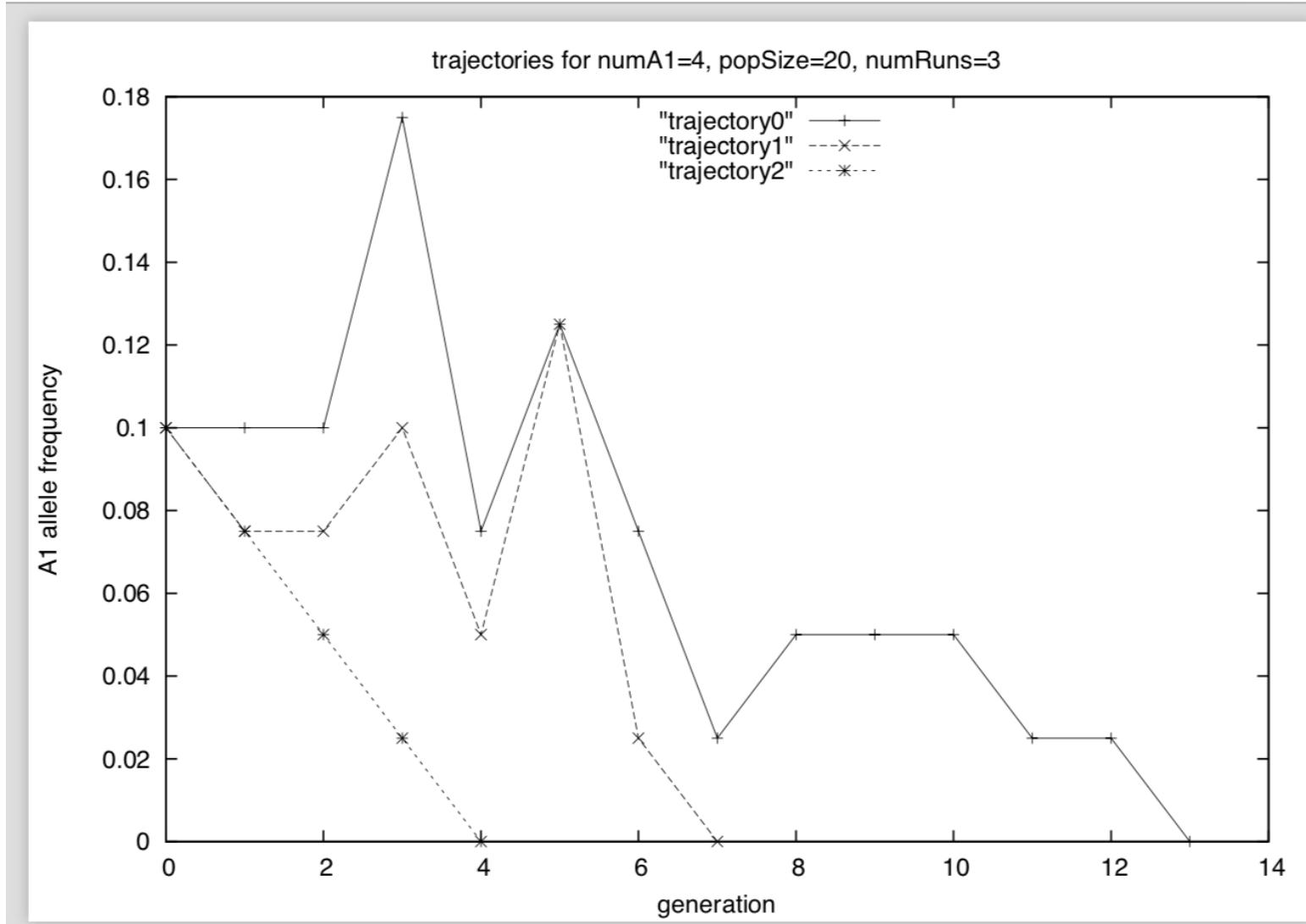
Model of population bottleneck of *H. sapiens* around 40K years ago provides a better fit for human SNP data than stationary, expansion, and collapse models. Image from Marth et al. PNAS 100(1): 376-381 (2003).

Simulation of genetic drift in diploid monoecious population of size N

- Wright-Fisher model for **diploid** monoecious (hermaphroditic) population of size N corresponds to the following simulation program
- 5 independent runs of geneticDrift.py, an implementation of the program described in Gillespie with the following pseudocode. Here the population consists of $N = 5$ hermaphroditic individuals, together containing $2N = 10$ alleles. In the initial population, there are $A = 2$ copies of the allele A_1 , and $2N - A = 8$ copies of the A_2 allele.

```
TRAJECTORIES = []
for k = 1 to M
    prevPop = [1,1,2,2,2,2,2,2,2,2] #alleles in population of size 2N
    numGen = 0                         #keep track of current generation
    repeat until prevPop either (does not contains 1) or (does not contain 2)
        nextPop = []
        for i = 1 to 2N
            x = randomly selected allele from prevPop
            nextPop.append(x)
        prevPop = nextPop; numGen = numGen+1
        append numGen to StoppingTimes
        append nextPop to TRAJECTORIES
```

- when $A = 8$ and $N = 20$, the proportion of A_1 alleles in the population of $2N = 40$ chromosomes is $8/40 = 0.2$, and when running my program for $M = 100$ repetitions, we have mean stopping time of 39.3 ± 47.92 , proportion of fixations of A_1 in 100 repetitions is 0.22, while the initial proportion $A/2N$ is 0.2
- three subsequent runs of my program with $A = 8$, $N = 20$, $M = 100$ yields mean stopping times 36.83 ± 33.99 , 38.14 ± 35.43 and 43.26 ± 42.22 while the proportion of fixations of A_1 were 0.16, 0.29, 0.17.
- for $M = 10,000$ repetitions, we obtain mean stopping time of 39.4707 ± 38.3772 and proportion of fixations of A_1 of 0.1971



- average stopping time over 10^4 runs returned the value mean stopping time 25.35 ± 33.37 and proportion of fixations of A1 of 0.098
- average stopping time over 10^5 runs returned the value mean stopping time 25.94 ± 33.01 and proportion of fixations of A1 of 0.10036

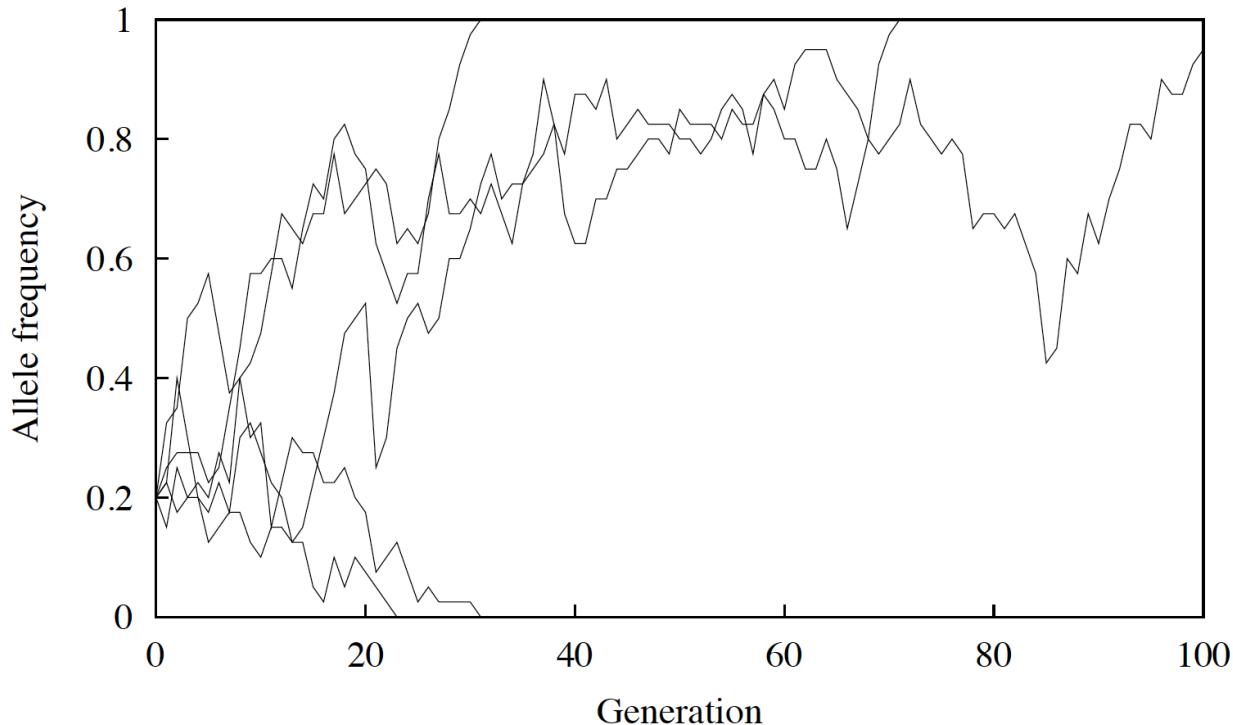


Figure 2.1: A computer simulation of genetic drift. The frequency of the A_1 allele, p , is graphed for 100 generations in five replicate populations each of size $N = 20$ and with initial allele frequency $p = 0.2$.

Figure 2.1 from Gillespie – note that by averaging 5 trajectories together, one obtains **very** different output!

Python code to simulate Wright-Fisher model

```
def main(A,N,M):  
    D = #dictionary of size M, D[k] is k-th trajectory (of variable len)  
    p0 = float(A)/(2*N) #initial allele relative frequency in (0,1)  
    A0 = A #initial allele frequency  
    for k in range(M):  
        D[k] = [p0] #initialize k-th trajectory to start with p0  
        numA1 = A #initial allele frequency  
        p = p0 #initial relative freq of A1 in initial population of size 2N  
        while numA1 not in [0,2*N]:  
            numA1 = 0 #numA1 will count number of A1 alleles in next population  
            for i in range(2*N):  
                z = random.random()  
                if z<p:  
                    numA1 += 1  
            p = float(numA1)/(2*N) #current allele freq of A1  
            D[k].append(p)  
    StoppingTimes = [] #list of all M stopping times
```

continuation of Python code

```
for k in range(M):
    StoppingTimes.append(len(D[k]))
    filename = "trajectory%s.txt" % k
    file = open(filename, 'w')
    for p in D[k]: file.write("%s\n" % p)
    file.close()

#compute mean and stdev of StoppingTimes
sum    = 0.0
sumSq = 0.0
for k in range(M):
    sum    += StoppingTimes[k]
    sumSq += (StoppingTimes[k])**2
mean   = sum/float(M)
var    = sumSq/float(M) - mean**2
stdev = math.sqrt(var)
print "mean stopping time: %s ± %s" % (mean, stdev)
```

Proportion of trajectories terminating with A_1 monomorphism

- Suppose that $p = \text{Prob}[A_1] = \frac{k}{2N}$ in a diploid population of constant size N , and that mutation and selection is not present.
- Running the previous program many times (see demo program `geneticDrift.py`), you find that the **proportion** of trajectories that terminate in a population that contains only the A_1 allele is approximately p . The following theorem shows equality in the limit, as the number of runs approaches ∞ . Let $\text{num}(A_1, M)$ denote the number of trajectories in which the final population contains only the A_1 allele (here each trajectory is run until convergence to allele monomorphism).

THEOREM:

$$\lim_{M \rightarrow \infty} \frac{\text{num}(A_1, M)}{M} = \text{Prob}[A_1] = p$$

Once heterogeneity has vanished (allele monomorphism has been achieved), the probability that any particular one of the $2N$ alleles from the initial population is the ancestor is $\frac{1}{2N}$, so if there are k A_1 alleles in the initial population, the probability that the ancestor of common alleles in the terminal population is A_1 is $\frac{k}{2N} = p$.

For finite population, with no mutation or selection, heterozygosity decreases over time

- Suppose that in a population of N individuals, the A_1 allele probability is $p = \frac{k}{2N}$, and the A_2 allele probability is $q = 1 - p = \frac{2N-k}{2N}$
- homozygosity G is defined as probability of choosing 2 alleles that are identical by state from a collection of $2N$ alleles
- \mathcal{G} is defined as (conditional) probability of choosing 2 alleles that are identical by state from a collection of $2N$ alleles, given that the alleles are different by origin
- homozygosity G is the probability of choosing 2 balls of the same color from an urn with $2N$ balls, with replacement
- \mathcal{G} is the probability of choosing 2 balls of the same color from an urn with $2N$ balls, without replacement
- Assume that in a population of size N , there are k copies of allele A_1 , for $0 \leq k \leq 2N$ and $2N - k$ copies of allele A_2 – i.e. $p = \frac{k}{2N}$, and $q = \frac{2N-k}{2N}$.

- homozygosity given by

$$G = p^2 + q^2 = \left(\frac{k}{2N}\right)^2 + \left(\frac{2N-k}{2N}\right)^2$$

- heterozygosity \mathcal{H} is defined by $1 - G$, where

$$\mathcal{G} = \frac{\binom{k}{2}}{\binom{2N}{2}} + \frac{\binom{2N-k}{2}}{\binom{2N}{2}}$$

We have

$$\begin{aligned} \mathcal{G} &= \frac{k(k-1)/2!}{2N(2N-1)/2!} + \frac{(2N-k)(2N-k-1)/2!}{2N(2N-1)/2!} = \frac{k(k-1) + (2N-k)(2N-k-1)}{2N(2N-1)} \\ &= \frac{k^2 - k + (2N)^2 - 2kN - 2N - 2kN + k^2 + k}{2N(2N-1)} = \frac{2k^2 + (2N)^2 - 4kN - 2N}{2N(2N-1)} \\ &= \frac{2N(2N-1) - 2k(2N-k)}{2N(2N-1)} = 1 - 2 \cdot \frac{k(2N-k)}{2N(2N-1)} = 1 - 2 \cdot \frac{k}{2N} \cdot \frac{2N-k}{2N-1} \approx 1 - 2pq = p^2 + q^2 = G \end{aligned}$$

$$\begin{aligned} G - \mathcal{G} &= \left[1 - 2 \cdot \frac{k}{2N} \cdot \frac{2N-k}{2N-1}\right] - \left[1 - 2 \cdot \frac{k}{2N} \cdot \frac{2N-k}{2N-1}\right] \\ &= \left[2 \cdot \frac{k}{2N} \cdot \frac{2N-k}{2N-1}\right] - \left[2 \cdot \frac{k}{2N} \cdot \frac{2N-k}{2N}\right] = \left(2 \cdot \frac{k}{2N}\right) \cdot \left[\frac{2N-k}{2N-1} - \frac{2N-k}{2N}\right] \\ &= \frac{k}{N} \cdot \frac{4N^2 - 2kN - (4N^2 - 2kN - 2N + k)}{2N(2N-1)} = \frac{k}{N} \cdot \frac{2N-k}{2N(2N-1)} = \frac{1}{N} \cdot \frac{k}{2N} \frac{2N-k}{2N-1} \approx \frac{pq}{N} \end{aligned}$$

- Look at Problem 2.6

- In the case that there are 3 alleles, N_1 of which are A_1 , N_2 of which are A_2 and N_3 of which are A_3 , where $n_1 + n_2 + n_3 = 2N$, the probability of choosing two alleles of the same state is given by G (with replacement) and \mathcal{G} (without replacement) where

$$G = \left(\frac{N_1}{2N}\right)^2 + \left(\frac{N_2}{2N}\right)^2 + \left(\frac{N_3}{2N}\right)^2$$

$$\mathcal{G} = \frac{\binom{N_1}{2} + \binom{N_2}{2} + \binom{N_3}{2}}{\binom{2N}{2}}$$

- Binomial distribution: describes probability of drawing a sample of $n = r + b$ balls, r of which are red and b of which are black, taken from an urn of $N = R + B$ balls, R of which are red and B of which are black with replacement

$$\text{binom prob} = \binom{n}{r} \cdot \left(\frac{R}{N}\right)^r \cdot \left(\frac{N-R}{N}\right)^{n-r}$$

$$= \binom{n}{r} p^r q^{n-r}$$

- Hypergeometric distribution: describes probability of drawing a sample of $n = r + b$ balls, r of which are red and b of which are black, taken from an urn of $N = R + B$ balls, R of which are red and B of which are black without replacement

$$\text{hypergeom prob} = \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}}$$

- It is known that for large N , the binomial and hypergeometric probabilities are approximately equal, so we expect $G \approx \mathcal{G}$

Decay of heterozygosity

- heterozygosity $\mathcal{H} = 1 - \mathcal{G}$
- One generation later, two distinct, randomly chosen alleles are identical (i.e. identical by state, but different by origin) is given by probability

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

since 2 distinct alleles can be copies of the same ancestral gene with probability $\frac{1}{2N}$, or this is not the case with probability $1 - \frac{1}{2N}$ and the two alleles were already identical by state yet different in origin in the previous generation.

- There is thus an exponentially fast decay in heterozygosity caused by genetic drift in a fixed size population

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

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

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

$$\mathcal{H}^{(t)} = \mathcal{H}^{(t-1)} \left(1 - \frac{1}{2N}\right) = \cdots = \mathcal{H} \left(1 - \frac{1}{2N}\right)^{(t)}$$

- Time $t_{1/2}$ necessary to reduce initial heterozygosity $\mathcal{H}^{(0)}$ by 50%

$$\frac{\mathcal{H}^{(0)}}{2} = \mathcal{H}^{(0)} \cdot \left(1 - \frac{1}{2N}\right)^{(t_{1/2})}$$

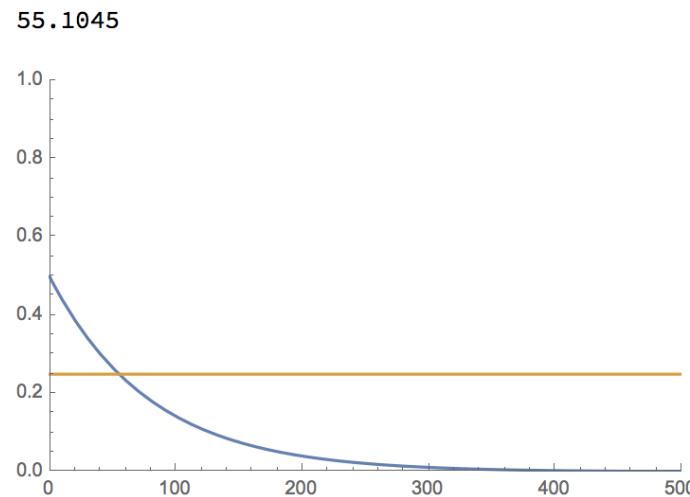
$$\ln(1/2) = t_{1/2} \cdot \ln\left(1 - \frac{1}{2N}\right)$$

$$t_{1/2} = \frac{-\ln(2)}{\ln\left(1 - \frac{1}{2N}\right)}$$

$$\ln(1 - x) = - \sum_{n=1}^{\infty} \frac{x^n}{n} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots \approx -x$$

$$t_{1/2} \approx 2N \ln 2$$

Plot of heterozygosity over time



- Plot of heterozygosity over time, for initial heterozygosity 0.5 and population size $N = 40$. Note that $t_{1/2} = 55.1045 \approx 2N \ln 2 = 80 \cdot \ln 2 = 55.45$.

Mathematica code to produce plot above

```
H0 = 0.5;
num = 40;
H[n_] := H0*(1 - 1/(2 num))^n;
Tonehalf = N[-Log[2]/Log[1 - 1/(2 num)]]
H0div2 = H0/2;
G = Plot[{H[n], H0div2}, {n, 0, 500}, PlotRange -> {{0, 500}, {0, 1}}]
```

Mutation and genetic drift

- Genetic drift reduces diversity, while mutation increases diversity
- in absence of natural selection, allele frequencies gravitate to an equilibrium value where mutation and drift balance each other
- u is mutation rate of neutral alleles (since we currently neglect natural selection), for instance given as a probability such as 10^{-8} for nucleotide mutation per site per generation
- $2Nu$ is expected number of mutations introduced in the population in a single generation
- Since we have

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

it follows that genetic drift removes diversity at rate of $\frac{1}{2N}$

- In absence of mutation and selection,

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

At equilibrium,

$$\begin{aligned}\mathcal{G}^* &= \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \mathcal{G}^* \\ 0 &= \frac{1}{2N} - \frac{\mathcal{G}^*}{2N} \\ \mathcal{G}^* &= 1\end{aligned}$$

- In the presence of mutation (but no selection), the probability that 2 alleles different in origin yet identical in state is given by multiplying the usual formula by $(1 - u)^2$, since this assumes both chosen alleles have not mutated. Note that this is a bit sloppy, since both alleles could have mutated the same genomic site to the same (mutated) nucleotide; however, not much is lost, since the genomic site in the coding sequence of a gene (i.e. allele) is so large that in all likelihood different genomic sites will be mutated. This assumption is the infinite alleles hypothesis. Thus

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

$$(1 - u)^2 = 1 - 2u + u^2 \approx 1 - 2u$$

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

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

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

$$\begin{aligned} \mathcal{H}' &= 1 - \mathcal{G}' \approx 1 - \frac{1}{2N} - \mathcal{G} + \frac{\mathcal{G}}{2N} + 2u\mathcal{G} = (1 - \mathcal{G}) - \frac{1 - \mathcal{G}}{2N} + 2u\mathcal{G} \\ &= (1 - \mathcal{G}) \left[1 - \frac{1}{2N} \right] + 2u\mathcal{G} = \mathcal{H} \left[1 - \frac{1}{2N} \right] + 2u(1 - \mathcal{H}) \end{aligned}$$

$$\Delta\mathcal{H} = \mathcal{H}' - \mathcal{H} = \mathcal{H} - \frac{\mathcal{H}}{2N} + 2u(1 - \mathcal{H}) - \mathcal{H} = -\frac{\mathcal{H}}{2N} + 2u - 2u\mathcal{H}$$

$$\Delta\mathcal{H} = 0 \Leftrightarrow -\frac{\mathcal{H}}{2N} + 2u - 2u\mathcal{H} = 0 \Leftrightarrow \frac{\mathcal{H}}{2N} + 2u\mathcal{H} = 2u \Leftrightarrow \mathcal{H} \left(2u + \frac{1}{2N} \right) = 2u \Leftrightarrow \mathcal{H} \left(\frac{4Nu + 1}{2N} \right) = 2u \Leftrightarrow \mathcal{H} = \frac{4Nu}{4Nu + 1}$$

$$\Delta\mathcal{H} = 0 \Leftrightarrow \mathcal{H} = \frac{4Nu}{4Nu + 1}$$

- Equilibrium between reduction of diversity due to drift and increase of diversity due to mutation occurs when

$$\mathcal{H}^* = \frac{4Nu}{4Nu + 1}$$

$$\mathcal{G}^* = 1 - \mathcal{H}^* = \frac{1}{4Nu + 1}$$

- Book uses notation $\widehat{\mathcal{H}}$ and $\widehat{\mathcal{G}}$, but this is poor notation, since it is standard in statistics to indicate an estimate of a frequency by the hat. Hence we'll stick to using * to denote equilibrium values.
- the value $4Nu$ is of utmost importance in population genetics, so it is usually denoted by $\theta = 4Nu$, and called the **mutation parameter**, which is different from the **mutation rate** u
- Using Clustal Omega to align human and horse alpha-chain hemoglobin, I found 109 substitutions in 480 aligned residues (ignoring gaps). Assuming 60 million years between human and horse, we have 109 substitutions in 2×60 million years, so the substitution rate ρ is

$$\rho = \frac{109}{1.2 \times 10^8} = 9.08333333333334e - 07 \approx 9.08 \times 10^{-7}$$

and so the mutation rate per residue is

$$\rho = \frac{9.08 \times 10^{-7}}{480} = 1.89 \times 10^{-9}$$

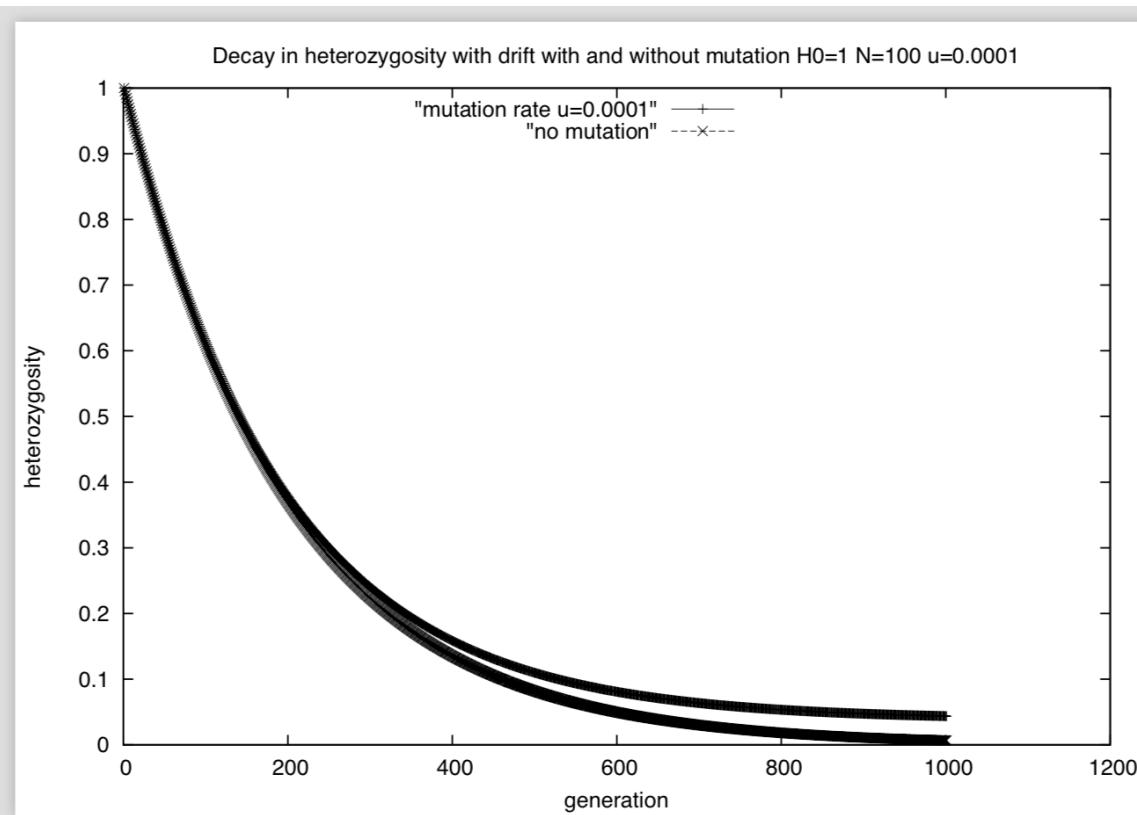
- Surprisingly, it is possible to show that **substitution rate** ρ equals **mutation rate** u . The expected number of (new) mutations in a diploid species of constant population size N is clearly $2Nu$, while we earlier showed that the fixation probability π of an allele A_1 with allele probability p is p , i.e.

$$\pi(p) = p$$

The allele frequency of a new mutation is $\frac{1}{2N}$, which is thus its fixation probability (proportion of trajectories that termination with all individuals having the new allele). The **substitution rate** ρ is the product of mutation rate and probability of fixation, so

$$\rho = 2Nu \cdot \pi\left(\frac{1}{2N}\right) = 2Nu \cdot \frac{1}{2N} = u$$

Decay of heterozygosity with and without mutation



- Graph showing **decay of heterozygosity** as a function of generation number. Equilibrium heterozygosity under drift and mutation for parameters $N = 100$, $u = 0.0001$ (upper curve) is

$$\mathcal{H}^* = \frac{4Nu}{1 + 4Nu} = 0.038461538461538464 \approx 0.038$$

Of course, with no mutation, $u = 0$ and $\mathcal{H}^* = 0$.

Coalescent

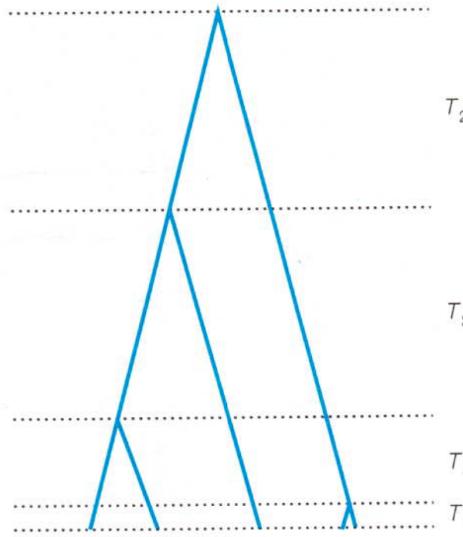
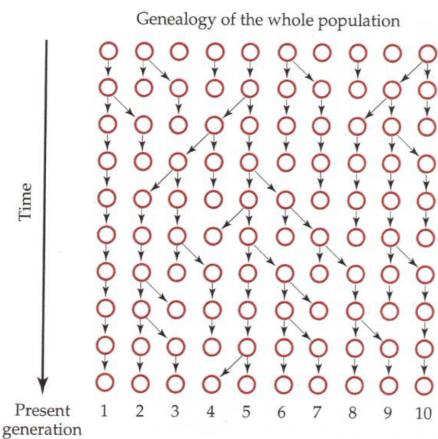


FIGURE 6.2 Example of a coalescent genealogy for a sample size of $n = 5$. The times between coalescent events, T_i , are equivalent to the amount of time that there are i lineages in the genealogy.

- the **coalescent** is an approximate **backward** simulation of the (forward) Wright-Fisher process, which samples **potential** genealogies for the ancestors of a given current population of individuals (alleles)
- The **coalescent** was introduced by J.F.C. Kingman (1982), and has become a workhorse of population genetics – see, for instance, “Natural Selection and Coalescent Theory” by J. Wakeley (Biology, Harvard)
- a **coalescence** occurs when two individuals (alleles) from generation k have the same parent in generation $k - 1$

- given a sample of k chromosomes from a population of size N , we will show that the **probability that a coalescence occurs** is

$$\frac{\binom{k}{2}}{N} = \frac{k(k-1)}{2N} \quad (\text{haploid case})$$

$$\frac{\binom{k}{2}}{2N} = \frac{k(k-1)}{4N} \quad (\text{diploid case})$$

- the coalescent has many applications, since mutations can be “thrown” at coalescent-generated genealogies, recombination can be accommodated, etc.
- to understand the simple algorithm behind the coalescent, we need first to revisit the **exponential distribution**, which is the continuous analogue of the **geometric distribution**
- exponential distribution:** random variable X is exponentially distributed with parameter $\lambda > 0$ if its **probability density** function is

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and the **cumulative distribution function** (CDF) is

$$F(x) = \int_{-\infty}^x p(t)dt = \int_0^x \lambda e^{-\lambda t} dt$$

$$= \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- Why does this generalize the geometric distribution? Recall that the geometric distribution corresponds to the number of coin flips before obtaining the first heads. If X denotes the number of flips before appearance of the first heads for a coin with heads probability p , then

$$\text{Prob}[X = k] = p(1 - p)^{k-1}$$

Now

$$\begin{aligned} (1 - p)^n &= (-p + 1)^n = \sum_{i=0}^n \binom{n}{i} \cdot (-p)^i \cdot 1^{n-i} \\ &= 1 - np + \frac{n(n-1)}{2}p^2 - \frac{n(n-1)(n-2)}{3 \cdot 2}p^3 + \dots + \frac{n(n-1)(n-2)(n-3) \cdots (n-(n-1))}{n!}p^n \\ (\text{by the binomial theorem } (a+b)^n &= \sum_{i=0}^n \binom{n}{i} a^i b^{n-i}) \\ &\approx 1 - np + \frac{(np)^2}{2!} - \frac{(np)^3}{3!} + \frac{(np)^4}{4!} - \dots = \sum_{i=0}^{\infty} \frac{(-np)^i}{i!} = e^{-np} \end{aligned}$$

$$p(1 - p)^t \approx pe^{-pt} \quad (\text{by replacing } n \text{ by } t \text{ and multiplying both sides by } p)$$

- Mean of the geometric distribution with parameter p is $\frac{1}{p}$, and mean of the exponential distribution with parameter λ is similarly $\frac{1}{\lambda}$
- Variance of the geometric distribution with parameter p is $\frac{1}{p^2} - \frac{1}{p} \approx \frac{1}{p^2}$, provided heads probability p is close to 0.
Variance of the exponential distribution with parameter λ is $\frac{1}{\lambda^2}$
- While the **Poisson** distribution with parameter μ has **mean** and **variance** of μ , we see that the **exponential** distribution with parameter λ has **mean** and **standard deviation** of $\frac{1}{\lambda}$
- Poisson** distribution models the **number** of events in a given time or space interval, while the **exponential** models the **time or distance** between successive events, which are counted by the Poisson in a given time or space interval

Computing the probability of a coalescence

- given an allele (leaf of coalescent tree), the probability that another given allele has a different ancestor allele from a population of constant size N is $1 - \frac{1}{2N}$, since we recall that there are $2N$ alleles (gametes) for a fixed locus
- given an allele (leaf of coalescent tree), the probability that two other given alleles have different ancestors that differ from each other and from the first is $1 - \frac{2}{2N}$, etc.
- probability that all n taxa have different ancestors is

$$\begin{aligned}
 q &= \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right) \cdots \left(1 - \frac{n-1}{2N}\right) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{2N}\right) \\
 &= 1 - \frac{1}{2N} - \frac{2}{2N} - \cdots - \frac{n-1}{2N} + \sum_{1 \leq i \neq j \leq 2N} \left(\frac{i}{2N} \cdot \frac{j}{2N}\right) - \sum_{1 \leq i \neq j \neq k \leq 2N} \left(\frac{i}{2N} \cdot \frac{j}{2N} \cdot \frac{k}{2N}\right) + \cdots
 \end{aligned}$$

$$q \approx 1 - \frac{1}{2N} - \frac{2}{2N} - \cdots - \frac{n-1}{2N} = 1 - \frac{1}{2N} \cdot \sum_{i=1}^{n-1} i = 1 - \frac{1}{2N} \cdot \frac{n(n-1)}{2} = 1 - \frac{n(n-1)}{4N}$$

$$p = 1 - q \approx \frac{1}{2N} \cdot \frac{n(n-1)}{2} = \frac{n(n-1)}{4N}$$

- Consider p the probability of heads (coalescence), with $q = 1 - p$ the probability of tails (non-coalescence), when flipping a coin an indefinite number of times – i.e. geometric distribution with heads probability p . The expected number of coin flips necessary to obtain the first heads is

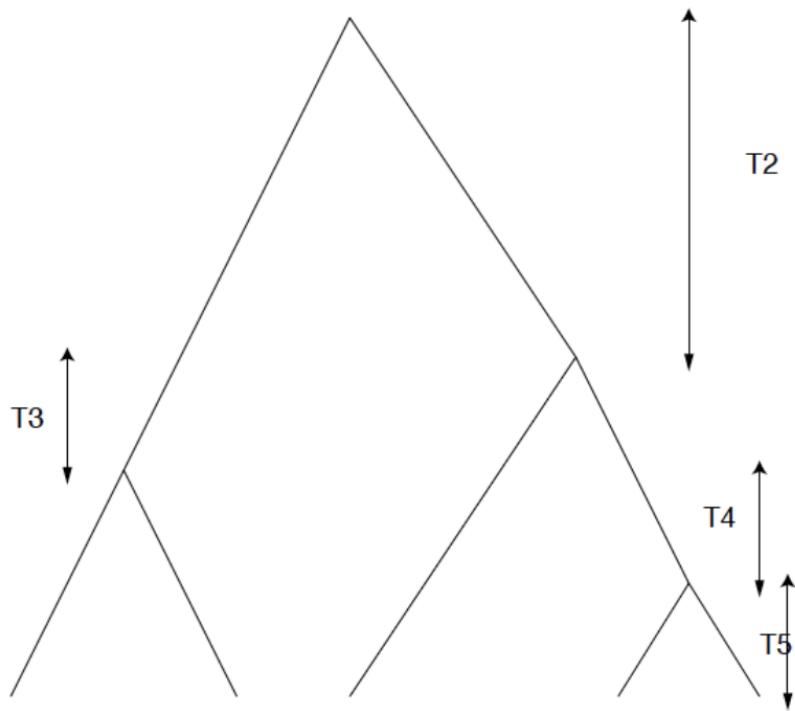
$$E[T_n] = \frac{1}{p} \approx \frac{1}{\frac{n(n-1)}{4N}} = \frac{4N}{n(n-1)}$$

- after a coalescence event (where we assume that exactly two of the n taxa have coalesced into $n - 1$ taxa (i.e. we do not consider the possibility that 3 or more alleles have the same ancestor at the coalescence event))
- by induction, it follows that

$$\begin{aligned} E[T_n] &\approx \frac{4N}{(n)(n-1)} \\ E[T_{n-1}] &\approx \frac{4N}{(n-1)(n-2)} \\ E[T_{n-2}] &\approx \frac{4N}{(n-2)(n-3)} \\ &\vdots \\ E[T_2] &\approx \frac{4N}{2 \cdot 1} = \frac{4N}{2} = 2N \end{aligned}$$

- Here is an alternate, shorter proof that the probability of a coalescence is $\frac{k(k-1)}{4N}$ if there are k samples from a diploid population of size N . For any two **fixed**, distinct alleles, the probability that these two alleles have the same parent is $\frac{1}{2N}$. Since there are $\binom{k}{2} = \frac{k(k-1)}{2}$ distinct pairs of alleles, the probability that at least one pair of alleles has the same parent is

$$\binom{k}{2} \cdot \frac{1}{2N} = \frac{k(k-1)}{2} \cdot \frac{1}{2N} = \frac{k(k-1)}{2N}$$



- For the example depicted above, if T_c denotes the **sum of all branch lengths in coalescent tree** then

$$T_c = 5T_5 + 4T_4 + 3T_3 + 2T_2$$

- more generally, the total sum of all branch lengths in any given coalescent tree is

$$T_c = nT_n + (n-1)T_{n-1} + \cdots + 2T_2 = \sum_{i=2}^n i \cdot T_i$$

$$E[T_c] = E[nT_n + (n-1)T_{n-1} + \cdots + 2T_2] = \sum_{i=2}^n E[i \cdot T_i] = \sum_{i=2}^n i \cdot E[T_i] = \sum_{i=2}^n i \cdot \frac{4N}{i(i-1)} = 4N \cdot \sum_{i=2}^n \frac{1}{i-1}$$

$$E[T_c] = 4N \cdot \sum_{i=1}^{n-1} \frac{1}{i} \approx 4N \cdot \ln(n-1)$$

- expected number of mutations appearing on coalescent tree is

$$uE[T_c] = 4Nu \cdot \sum_{i=1}^{n-1} \frac{1}{i}$$

where u denotes rate of mutation of DNA for the particular protein family

- Since DNA mutation rate is very small, it is unlikely (though possible) that the locus of each mutation event is distinct (i.e. assume no back-mutations and that it cannot happen that an A mutates to C, then later to a T), we assume that each mutation event corresponds to a distinct locus on the coalescent tree
- Letting S_n denote the number of segregating sites determined from the multiple alignment of n DNA sequences,

$$E[S_n] = uE[T_c] = 4Nu \cdot \sum_{i=1}^{n-1} \frac{1}{i} = \theta \cdot \sum_{i=1}^{n-1} \frac{1}{i}$$

$$\hat{\theta} = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

- Expected time $E[T_{MRCA}]$ to the most recent common ancestor in a population of constant size N

$$T_{MRCA} = T_n + T_{n-1} + \cdots + T_2$$

$$E[T_{MRCA}] = E[T_n + T_{n-1} + \cdots + T_2] = E[T_n] + E[T_{n-1}] + \cdots + E[T_2]$$

$$= \sum_{i=2}^n \frac{4N}{i(i-1)} = 4N \sum_{i=1}^{n-1} \frac{1}{i(i+1)} = 4N \left(\left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{n-1} - \frac{1}{n}\right) \right) = 4N\left(1 - \frac{1}{n}\right)$$

- Note that

$$T_2 = \frac{4N}{2(1)} = 2N$$

Thus half the time taken for allele fixation or extinction is spent in going from 2 alleles to 1!

Estimating $\theta = 4Nu$ for *D. melanogaster* ADH

| Allele | 39 | 226 | 387 | 393 | 441 | 513 | 519 | 531 | 540 | 578 | 606 | 615 | 645 | 684 |
|-----------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Reference | T | C | C | C | C | C | T | C | C | A | C | T | A | G |
| Wa-S | . | T | T | . | A | A | C | . | . | . | . | . | . | . |
| Fl-1S | . | T | T | . | A | A | C | . | . | . | . | . | . | . |
| Af-S | . | . | . | . | . | . | . | . | . | . | . | . | . | A |
| Fr-S | . | . | . | . | . | . | . | . | . | . | . | . | . | A |
| Fl-2S | G | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Ja-S | G | . | . | . | . | . | . | . | T | . | T | . | C | A |
| Fl-F | G | . | . | . | . | . | . | G | T | C | T | C | C | . |
| Fr-F | G | . | . | . | . | . | . | G | T | C | T | C | C | . |
| Wa-F | G | . | . | . | . | . | . | G | T | C | T | C | C | . |
| Af-F | G | . | . | . | . | . | . | G | T | C | T | C | C | . |
| Ja-F | G | . | . | A | . | . | . | G | T | C | T | C | C | . |

- Table 1.1 above shows $S_n = 14$ segregating sites, 13 of which involve silent mutations, in a multiple alignment of DNA from the coding region of the alcohol dehydrogenase (ADH) genes of $n = 11$ fruit flies (*D. melanogaster*) from Florida (Fl), Washington (Wa), Africa (Af), Japan (Ja), and France (Fr).
- it follows that

$$S_n = 14$$

$$n = 11$$

$$2.928968254 = \sum_{i=1}^{10} \frac{1}{i}$$

$$\hat{\theta} = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}} = \frac{14}{2.928968254} = 4.77984013$$

- neutral mutation is only accepted in the case of silent mutations, since non-silent mutations could involve replacement of a hydrophobic amino acid in the core by a hydrophilic amino acid, or replacement of a negatively charged amino acid (aspartic or glutamic acid) by a positively charged amino acid (asparagine, lysine, or histidine, the latter depending on pH)
- number of silent segregating sites is $S_n = 13$, so

$$S_n = 13$$

$$n = 11$$

$$2.928968254 = \sum_{i=1}^{10} \frac{1}{i}$$

$$\hat{\theta} = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}} = \frac{13}{2.928968254} = 4.438422978$$

- In a population of fixed size N with n alleles, a similar analysis shows that

$$T_{MRCA} = T_n + T_{n-1} + \cdots + T_2$$

$$E[T_{MRCA}] = E[T_n + T_{n-1} + \cdots + T_2] = E[T_n] + E[T_{n-1}] + \cdots + E[T_2] = \sum_{i=2}^n E[T_i]$$

$$= \sum_{i=2}^n \frac{2N}{i(i-1)} = 2N \sum_{i=1}^{n-1} \frac{1}{i(i+1)} = 2N \left[\left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{n-1} - \frac{1}{n}\right) \right]$$

$$E[T_{MRCA}] = 2N \left(1 - \frac{1}{n}\right)$$

Here $2N$ occurs in place of $4N$, since we consider instead the probability

$$p = \left(1 - \frac{1}{N}\right) \cdot \left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{n-1}{N}\right)$$

of choosing different ancestors

Some applications of the coalescent

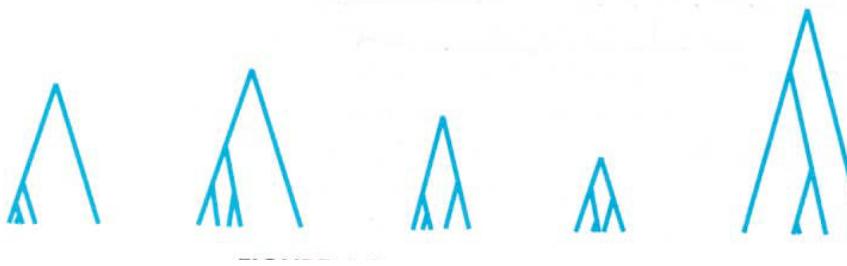
- to estimate the **mutation parameter** $\theta = 4Nu$, where u is nucleotide mutation rate, given the number S_n of segregating sites in a multiple alignment of n DNA sequences

$$\widehat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}$$

- to determine expected time $E[T_{MRCA}]$ to the most recent common ancestor in a population of fixed size N , given a sample of n sequences:

$$\begin{aligned}T_{MRCA} &= T_n + T_{n-1} + \cdots + T_2 \\E[T_{MRCA}] &= E[T_n + T_{n-1} + \cdots + T_2] = E[T_n] + E[T_{n-1}] + \cdots + E[T_2] \\&= 4N\left(1 - \frac{1}{n}\right)\end{aligned}$$

Sampling waiting times from exponential distribution



- For computer simulations, an exponentially distributed random variable for a given mean μ can be computed as follows. Let X be a uniformly distributed continuous random variable with $0 \leq X \leq 1$. (In Python, this can be done by calling function `random()`.) Note that $Pr[X > x] = 1 - Pr[X \leq x]$. Then

$$\begin{aligned}Pr[\text{there is an arrival in time } t] &= 1 - e^{-\frac{t}{\mu}} \\&= Pr[X > e^{-\frac{t}{\mu}}] = Pr\left[\ln X > -\frac{t}{\mu}\right] = Pr[-\mu \ln X < t]\end{aligned}$$

Thus repeatedly evaluating $-\mu \ln X$ for uniformly distributed random real numbers $0 \leq X \leq 1$ yields a sequence of sample interarrival times with mean μ .

- the stochastic variation in sampled times T_k , for $k = n, \dots, 2$ explains the difference in heights of sampled geneologies shown at the top of this slide

Pseudocode for 2 programs to compute coalescence times

- simple but slow program

```
A = present population {1,...,n} with indicated alleles
t = 0    #time backwards from present time
k = n    #number in sample at time t
T = {}   #T is dictionary of times T[n],T[n-1],...,T[2]
while k>1
    t = t+1 #update time
    each individual in 1,...,n chooses a parent from {1,...,n}
    if two individuals choose same parent
        T[k] = t
        k    = k-1 #decrease set of ancestors
        t    = 0 #reset time until next coalescence
return dictionary T
```

- more efficient (standard) program which samples from the **exponential** distribution, each time with different mean

```
A = present population {1,...,n} with indicated alleles
t = 0    #time backwards from present time
k = n    #number of samples at time t
T = {}   #T is dictionary of times T[n],T[n-1],...,T[2]
while k>1
    randomly choose 2 of the k to coalesce
    mean = 2/(k*(k-1))    #mean coalescent time for k samples
    x    = random floating point number in (0,1)
    T[k] = - mean*log(x)  #sample coalescent time from exp dist
    k    = k-1             #number of samples is decreased
return dictionary T
```

- Note that

- time for k th coalescence is $2N \cdot T[k]$ for diploid population size N
- time for k th coalescence is $N \cdot T[k]$ for haploid population size N

In the literature, this is expressed by stating that the coalescent tree is “scaled” by $2N$ resp. N

Mathematical models for drift in constant-size haploid populations with no selection or mutation

- Case 1: only 2 alleles (red and black) in haploid monoecious population with constant size n
 - Sum of n independent Poisson random variables, each with mean $\mu = 1$. It can be shown (by taking the convolution) that the sum of independent Poisson random variables is a Poisson random variable whose mean is the sum of the means of the constituent Poisson random variables

$$\text{Prob}[X + \dots + X = k] = e^{-n\mu} \cdot \frac{(n\mu)^k}{k!} = e^{-n} \cdot \frac{n^k}{k!}$$

Avise et al. consider this model in

Demographic influences on mitochondrial DNA lineage survivorship in animal populations.
J. Avise, J. Neigel and J. Arnold.
Journal of Molecular Evolution 20:99-105 (1984)

A classical result from **branching processes** (probability theory): the population will eventually die out with probability 1 if $\mu < 1$, while the population will continue to grow exponentially fast with probability 1 if $\mu > 1$.

- Hypergeometric distribution: describes probability of drawing a sample of $n = r + b$ balls, r of which are red and b of which are black, taken from an urn of $N = R + B$ balls, R of which are red and B of which are black without replacement

$$\text{hypergeom prob} = \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}}$$

- Suppose each of n individuals has exactly 2 offspring, and that n of the $2n$ progeny survive, while the parents and the remaining progeny die. Let X denote the random variable for the number of red offspring in the next generation, given a current number r of individuals in a population of size n

$$\text{Prob}[X = k] = \frac{\binom{2r}{k} \binom{2(n-r)}{n-k}}{\binom{2n}{n}}$$

- Case 1 (continued): di-allelic case (red and black alleles) for a haploid, monoecious population
 - Binomial distribution:** describes probability of drawing a sample of $n = r + b$ balls, r of which are red and b of which are black, taken from an urn of $N = R + B$ balls, R of which are red and B of which are black with replacement

$$\begin{aligned}\text{binom prob} &= \binom{n}{r} \cdot \left(\frac{R}{N}\right)^r \cdot \left(\frac{N-R}{N}\right)^{n-r} \\ &= \binom{n}{r} p^r q^{n-r} \\ &= \text{BINOM.DIST}(r, n, p, \text{FALSE}) \quad (\text{using Excel})\end{aligned}$$

- Wright-Fisher model** corresponds to the binomial distribution – compare with the simulation program
 - In the Wright-Fisher model for haploid, di-allelic population of size n with allele A relative frequency p , the probability that there are k copies of allele A in next generation is

$$P[X = k] = \binom{n}{k} \cdot p^k (1-p)^{n-k}$$

- In the Wright-Fisher model for diploid, di-allelic population of size n ,

$$P[X = k] = \binom{2n}{k} \cdot p^k (1-p)^{2n-k}$$

- Case 2: multiple alleles A_1, \dots, A_k (k colors) in haploid monoecious population with constant size n

- Multinomial distribution:

$$\text{Prob}[\alpha_1 = n_1, \dots, \alpha_k = n_k] = \frac{n!}{n_1! \cdots n_k!} \cdot p_1^{n_1} \cdots p_k^{n_k}$$

where $n_1 + \cdots + n_k = n$ and $p_1 + \cdots + p_k = 1$

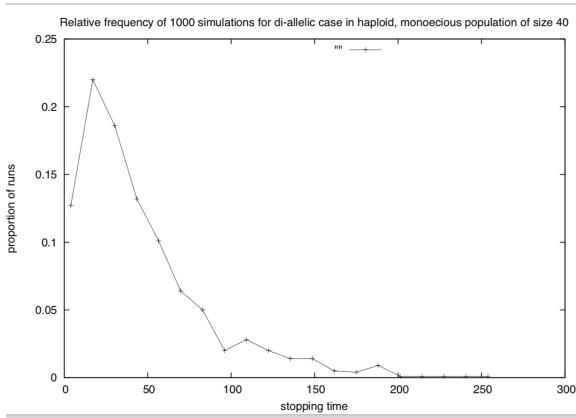
- computer simulations

Effective population size N_e

- Effective population size N_e for a certain model (Wright-Fisher, Moran, Avise, etc.) is defined so that the formulas for expected decrease of heterozygosity $\mathcal{H}_t = \mathcal{H}(1 - \frac{1}{2N})$, equilibrium heterozygosity $\mathcal{H}^* = \frac{4Nu}{1+4Nu}$, etc. hold provided N is replaced by N_e
- **Wright-Fisher model:** given a diploid, monoecious, hermaphroditic species with constant population size N , and proportion $p = \frac{k}{2N}$ of A_1 alleles at fixed bi-allelic site, in the Probability Appendix a proof is given that
 - ▷ mean p
 - ▷ variance $\frac{pq}{2N}$
 - ▷ effective population size $N_e = N$
- **Moran model:** Given a diploid, monoecious, hermaphroditic species with constant population size N , apply $2N$ successive steps to create the next generation, where a single step consists of selecting 2 individuals, the first to be **copied** (birth) and the second to be **removed** (die). Given proportion $p = \frac{k}{2N}$ of A_1 alleles at fixed bi-allelic site, in the Probability Appendix a proof is given that
 - ▷ mean p
 - ▷ variance $\frac{pq}{N}$
 - ▷ effective population size $N_e = \frac{N}{2}$

- **Avise model:** $2N$ independent Poisson random variables X_1, \dots, X_{2N} , each having mean 1 and variance σ^2 , where k of the X_i are A_1 alleles. Pages 50-51 prove that
 - ▷ mean p
 - ▷ variance $\frac{pq\sigma^2}{2N}$
 - ▷ effective population size $N_e = \frac{N}{\sigma^2}$
- This definition of N_e is known as the **variance-based effective population size**, to distinguish it from different notions, such as the **coalescence-based effective population size**

Stopping times for haploid, monoecious population



- histogram of 1000 stopping times for population of constant size $n = 40$ for di-allelic case in a haploid, monoecious population, with estimate for the mean stopping time of $\bar{x} = 52.760$ with sample standard deviation of $s = 40.002$
- it follows that the 95% confidence interval for mean stopping time when $n = 40$ is

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 52.76 \pm 1.96 \cdot \frac{40.002}{\sqrt{1000}} = 52.76 \pm 2.479$$

- this estimate, obtained by 1000 simulation runs, is not quite equal to the **exact** mean stopping time of 53.728, computed to within 3 decimal places ($\epsilon = 0.001$) by my program using the method described in the paper

Solving the Fisher-Wright and coalescence problems with a discrete Markov chain analysis
S.R. Buss and P. Clote
Advances in Applied Probability 36(3):1175-1197 (2004).

- using the coalescent in this **haploid, monoecious** case, we have

$$E[T_n] = \frac{1}{p} = \frac{1}{\binom{n}{2}/N} = \frac{1}{\frac{n(n-1)}{2N}} = \frac{2N}{n(n-1)}$$

and so for the di-allelic case ($n = 2$), the mean stopping time is 40 since

$$E[T_2] = \frac{2N}{2(2-1)} = N$$

- it follows that if one starts with a founding population of N haploid individuals (**mitochondrial Eve problem**), then **half** the expected time to reach **mitochondrial monomorphism** is taken in going from 2 alleles to one allele!

Tajima D-statistic hypothesis testing for neutrality

- **Purpose:** Tajima D is used to identify DNA sequences that do not fit the neutral theory model at equilibrium between mutation and genetic drift
- Idea is to compare two estimates of genetic diversity, which should be approximately equal under the hypothesis of neutral mutation theory:
 - ▶ average number of segregating sites $E[S_n]$
 - ▶ average number of nucleotide differences in DNA
- When we consider selection in a later chapter, we will see that a population carries many (heterozygous) deleterious alleles, each having low frequency – since frequency is low, unless the number of such deleterious mutations is enormous, the contribution to average number of nucleotide differences is small.
- On the other hand, despite having low frequency, such deleterious alleles may contribute greatly to the average number of segregating sites. It follows that if the DNA is under selective pressure, then these two estimates of genetic diversity will differ. This is the intuitive basis of Tajima's D statistic that tests whether DNA satisfies neutral mutation theory.
- Under the neutral mutation hypothesis, we've just derived the estimate $\widehat{\theta}$ with the property that $E[\widehat{\theta}] = \theta = 4Nu$

$$\widehat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}$$

- $\widehat{\theta}$ estimates the average number of segregating sites under neutral theory (balance between mutation and genetic drift, without selection)
- A different estimate $\widehat{\pi}$ is now given for $4Nu$ that is derived from heterozygosity, also with the property that $E[\widehat{\pi}] = \theta = 4Nu$

$$\widehat{\pi} = \frac{n}{n-1} \cdot \sum_{i=1}^{S_n} 2\widehat{p}_i(1 - \widehat{p}_i)$$

where \widehat{p}_i is an estimate for the allele frequency p_i of the A_1 allele at the i th segregating site, where $1 \leq i \leq S_n$. Here we assume (as before) that there are only two alleles A_1, A_2 at a segregating site.

- $\widehat{\pi}$ estimates the average heterozygosity – in Tajima's original paper, and in the literature, notation is \widehat{k} rather than $\widehat{\pi}$
- Under neutral theory, both $\widehat{\theta}$ and $\widehat{\pi}$ are unbiased estimators of $4Nu$, which just means that

$$E[\widehat{\theta}] = 4Nu$$

$$E[\widehat{\pi}] = 4Nu$$

Toy example for Tajima D-statistic

- 4 segregation sites, each with minor allele frequency of $\frac{1}{4}$ suggesting that mutations have proliferated (as in the case of sudden population expansion) – however, since the absolute value of the Tajima D-statistic is not greater than 2, we cannot reject the null hypothesis of neutral mutation in this instance

```
ATTTTTTTTTTTTTTTTT  
TTTTATTTTTTTTTTTT  
TTTTTTTTATTTTTTTT  
TTTTTTTTTTTTATTTT  
  
segregation list: [0, 4, 8, 12]  
harmonic sum: 1.83333333333  
pi: 2.0  
theta: 2.18181818182  
  
normalization constant C: 1.81696929266  
Tajima statistic D: -0.100066733407
```

- 1 segregation site, with minor allele frequency of $\frac{1}{2}$, clearly suggesting that the mutation is advantageous (so not neutral) – however, since the absolute value of the Tajima D-statistic is not greater than 2, we cannot reject the null hypothesis of neutral mutation in this instance

```
TTTTTTTTATTTTTTTTT  
TTTTTTTTATTTTTTTTT  
TTTTTTTTTTTTTTTTT  
TTTTTTTTTTTTTTTTT  
  
segregation list: [7]  
harmonic sum: 1.83333333333  
pi: 0.666666666667  
theta: 0.545454545455  
  
normalization constant C: 0.0742269619025  
Tajima statistic D: 1.63299316186
```

Tajima D-statistic hypothesis parameters and table

- Define normalization constant C , which is the standard deviation for the beta distribution (which provides a better fit than normal distribution) of differences $(\hat{\pi} - \hat{\theta})$

$$C = \sqrt{\frac{c_1}{a_1} S_n + \frac{c_2}{a_1^2 + a_2} \cdot S_n(S_n - 1)}$$

$$a_1 = \sum_{i=1}^{n-1} i^{-1}$$

$$a_2 = \sum_{i=1}^{n-1} i^{-2}$$

$$b_1 = \frac{n+1}{3(n-1)}$$

$$b_2 = 2 \cdot \frac{n^2 + n + 3}{9(n-1)}$$

$$c_1 = b_1 - \frac{1}{a_1}$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

■ Tajima D test statistic

$$D_T = \frac{\widehat{\pi} - \widehat{\theta}}{C}$$

■ hypothesis testing

$H_0 = \pi = \theta$, i.e. sequences satisfy neutral theory (null hypothesis)

$H_1 = \pi \neq \theta$, i.e. sequences do not satisfy neutral theory (alternative hypothesis)

- If $|D_T| > 2$, so $D_T < -2$ or $D_T > 2$, then reject null hypothesis with 99.9% confidence (assuming $n = 10$, with 2-sided test) – use following table:

| <i>n</i> | Confidence limit of <i>D</i> | | | |
|----------|------------------------------|----------------|----------------|----------------|
| | 90% | 95% | 99% | 99.9% |
| 4 | -0.876 ~ 2.081 | -0.876 ~ 2.232 | -0.876 ~ 2.324 | -0.876 ~ 2.336 |
| 5 | -1.255 ~ 1.737 | -1.269 ~ 1.834 | -1.275 ~ 1.901 | -1.276 ~ 1.913 |
| 6 | -1.405 ~ 1.786 | -1.478 ~ 1.999 | -1.540 ~ 2.255 | -1.556 ~ 2.373 |
| 7 | -1.498 ~ 1.728 | -1.608 ~ 1.932 | -1.721 ~ 2.185 | -1.761 ~ 2.311 |
| 8 | -1.522 ~ 1.736 | -1.663 ~ 1.975 | -1.830 ~ 2.313 | -1.909 ~ 2.524 |
| 9 | -1.553 ~ 1.715 | -1.713 ~ 1.954 | -1.916 ~ 2.296 | -2.023 ~ 2.519 |
| 10 | -1.559 ~ 1.719 | -1.733 ~ 1.975 | -1.967 ~ 2.362 | -2.105 ~ 2.640 |
| 11 | -1.572 ~ 1.710 | -1.757 ~ 1.966 | -2.014 ~ 2.359 | -2.174 ~ 2.649 |
| 12 | -1.573 ~ 1.713 | -1.765 ~ 1.979 | -2.041 ~ 2.401 | -2.223 ~ 2.729 |
| 13 | -1.580 ~ 1.708 | -1.779 ~ 1.976 | -2.069 ~ 2.403 | -2.267 ~ 2.741 |
| 14 | -1.580 ~ 1.710 | -1.783 ~ 1.985 | -2.085 ~ 2.432 | -2.299 ~ 2.798 |
| 15 | -1.584 ~ 1.708 | -1.791 ~ 1.984 | -2.103 ~ 2.436 | -2.329 ~ 2.811 |
| 16 | -1.583 ~ 1.709 | -1.793 ~ 1.990 | -2.113 ~ 2.457 | -2.350 ~ 2.854 |
| 17 | -1.585 ~ 1.708 | -1.798 ~ 1.990 | -2.126 ~ 2.461 | -2.372 ~ 2.866 |
| 18 | -1.584 ~ 1.709 | -1.799 ~ 1.996 | -2.132 ~ 2.478 | -2.387 ~ 2.900 |
| 19 | -1.585 ~ 1.708 | -1.802 ~ 1.996 | -2.141 ~ 2.483 | -2.403 ~ 2.911 |
| 20 | -1.584 ~ 1.710 | -1.803 ~ 2.001 | -2.146 ~ 2.496 | -2.414 ~ 2.939 |
| 21 | -1.585 ~ 1.709 | -1.805 ~ 2.001 | -2.152 ~ 2.501 | -2.426 ~ 2.950 |
| 22 | -1.584 ~ 1.711 | -1.804 ~ 2.005 | -2.153 ~ 2.512 | -2.434 ~ 2.973 |
| 23 | -1.584 ~ 1.710 | -1.806 ~ 2.006 | -2.160 ~ 2.516 | -2.443 ~ 2.983 |
| 24 | -1.583 ~ 1.712 | -1.806 ~ 2.009 | -2.162 ~ 2.526 | -2.449 ~ 3.002 |
| 25 | -1.583 ~ 1.712 | -1.807 ~ 2.010 | -2.165 ~ 2.530 | -2.457 ~ 3.011 |
| 26 | -1.582 ~ 1.712 | -1.807 ~ 2.013 | -2.167 ~ 2.538 | -2.461 ~ 3.029 |
| 27 | -1.582 ~ 1.712 | -1.807 ~ 2.014 | -2.170 ~ 2.542 | -2.467 ~ 3.037 |
| 28 | -1.581 ~ 1.713 | -1.807 ~ 2.017 | -2.171 ~ 2.549 | -2.471 ~ 3.052 |
| 29 | -1.581 ~ 1.714 | -1.807 ~ 2.018 | -2.173 ~ 2.553 | -2.475 ~ 3.060 |
| 30 | -1.580 ~ 1.714 | -1.807 ~ 2.020 | -2.173 ~ 2.559 | -2.478 ~ 3.073 |

Math details for Tajima D-statistic

- Since p_i is actually a stochastic quantity (allele frequencies can change over time or geographic location, as seen in later chapters), the expected heterozygosity at a fixed site i is a function of $E[p_i]$, rather than of p_i . Letting \mathcal{H}_i denote the heterozygosity at the i th locus

$$E[\mathcal{H}_i] = E[2p_i(1 - p_i)]$$

since

$$p_i(1 - p_i) = \text{Prob}[A_1 \text{ on paternal DNA strand, } A_2 \text{ on maternal DNA strand}]$$

$$(1 - p_i)p_i = \text{Prob}[A_2 \text{ on paternal DNA strand, } A_1 \text{ on maternal DNA strand}]$$

$$2p_i(1 - p_i) = \text{Prob}[\text{alleles differ on paternal and maternal chromosome}] = \text{Prob}[\text{site } i \text{ is heterozygous}]$$

- Key fact

$$\pi = E\left[\sum_{i=1}^{\infty} 2p_i(1 - p_i)\right] = 4Nu$$

Intuitive justification of key fact:

$$\pi = E\left[\sum_{i=1}^{\infty} 2p_i(1 - p_i)\right] \quad \text{expected heterozygosity, i.e. average number of nucleotide differences}$$

$$E[T_2] = 2N \quad \text{for neutral loci, expected time to MRCA of 2 alleles}$$

$$2uE[T_2] \quad \text{expected number of nucleotide differences between 2 alleles}$$

$$\pi = 2uE[T_2] = 4Nu$$

- Rigorous derivation of key fact given a few slides later, when we discuss the stationary distribution of allele frequency of A_1

- Coalescent provides simple, intuitive justification that heterozygosity of allele A_1 at equilibrium, when $\Delta\mathcal{H} = 0$ is given by

$$\mathcal{H}^* = \frac{4Nu}{1 + 4Nu}$$

- probability of coalescence between two given alleles in any given generation is $\frac{1}{2N}$
- probability that mutation does not occur in 2 given alleles in one generation is $(1 - u)^2$
- probability that at least one mutation occurs among 2 given alleles is

$$1 - (1 - u)^2 = 1 - (1 - 2u + u^2) \approx 2u$$

- probability that a mutation or coalescence occurs between 2 given alleles in one generation is $\approx 2u + \frac{1}{2N}$, so probability that a mutation occurs before a coalescence occurs is

$$\frac{2u}{2u + 1/2N} = \frac{1}{1 + 1/4Nu} = \frac{4Nu}{4Nu + 1}$$

which is the expected heterozygosity at equilibrium

- In practice, $\hat{\pi}$ is computed as follows, given n DNA sequences (from protein-coding gene, or possibly non-coding region suspected of playing regulatory role, such as transcription factor binding site or non-coding RNA)

$$\hat{\pi} = \hat{k} = \frac{\sum_{1 \leq i < j \leq n} k_{i,j}}{\binom{k}{2}}$$

where $\binom{k}{2} = \frac{k(k-1)}{2}$, and $k_{i,j}$ is the number of nucleotide differences between the i th and j th DNA sequence

Interpretation of Tajima's D

- If null hypothesis $\pi = \theta = 4Nu$ cannot be rejected, then the site is deemed to satisfy the neutral mutation hypothesis – there is no evidence of selection
- If the average heterozygosity $\widehat{k} = \frac{\sum_{1 \leq i < j \leq n} k_{i,j}}{\binom{n}{2}}$ is less than the number $\widehat{\theta} = \frac{s_n}{1 + \frac{1}{2} + \dots + \frac{1}{n-1}}$ of segregating sites (with statistical significance), then there is an excess of rare alleles. In other words, if there are statistically significantly fewer haplotypes than number of segregating sites, then this suggests a recent selective sweep, where the population has expanded after a recent bottleneck.
- If the average heterozygosity $\widehat{k} = \frac{\sum_{1 \leq i < j \leq n} k_{i,j}}{\binom{n}{2}}$ is greater than the **Watterson mutation parameter** $\widehat{\theta} = \frac{s_n}{1 + \frac{1}{2} + \dots + \frac{1}{n-1}}$ (with **statistical significance**), then there are few rare alleles, which suggests balancing selection after a sudden population contraction.

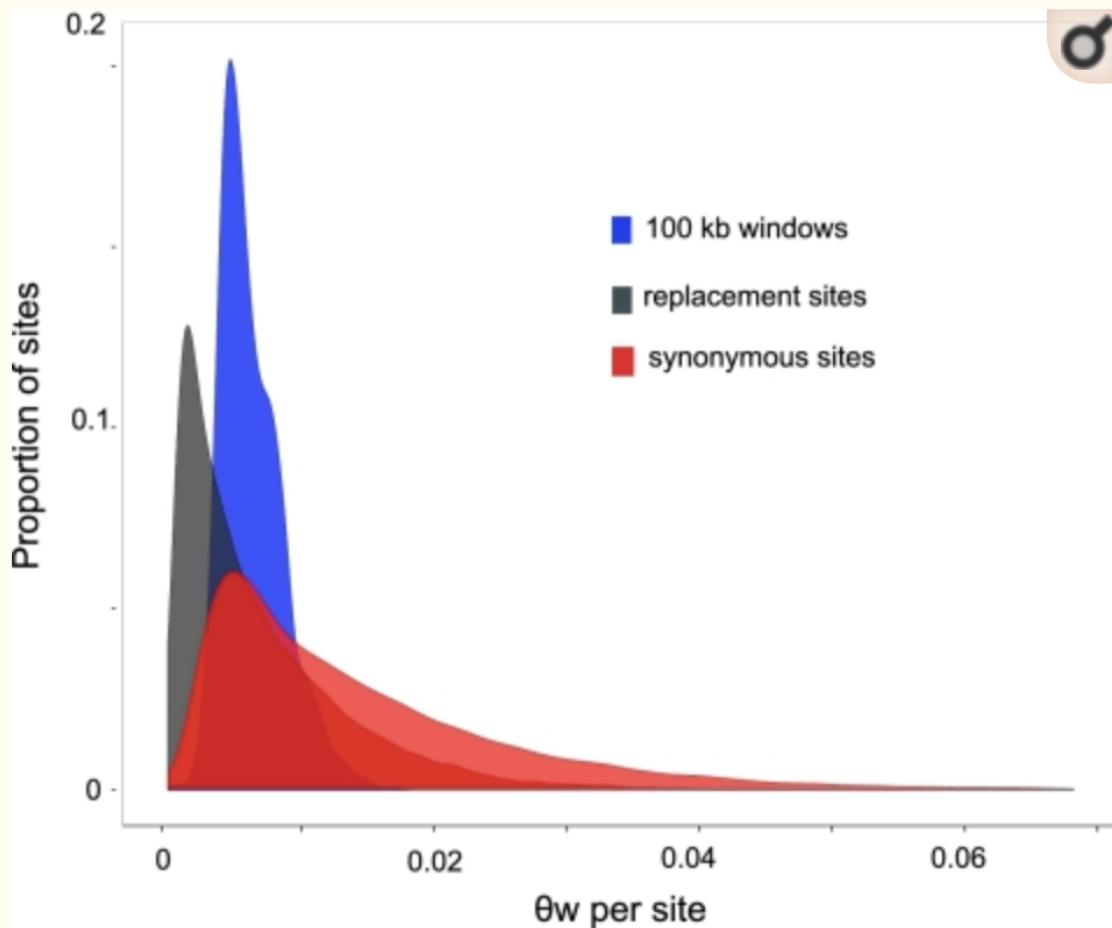


Fig. 1.

Distributions of nucleotide diversity (θ_W) found in 100-kb sliding windows and replacement (gray) and synonymous (red) sites among 30,768 gene models.

- Image from Branca et al. "Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*", Proc Natl Acad Sci U S A. 2011 Oct 18; 108(42): E864-E870.

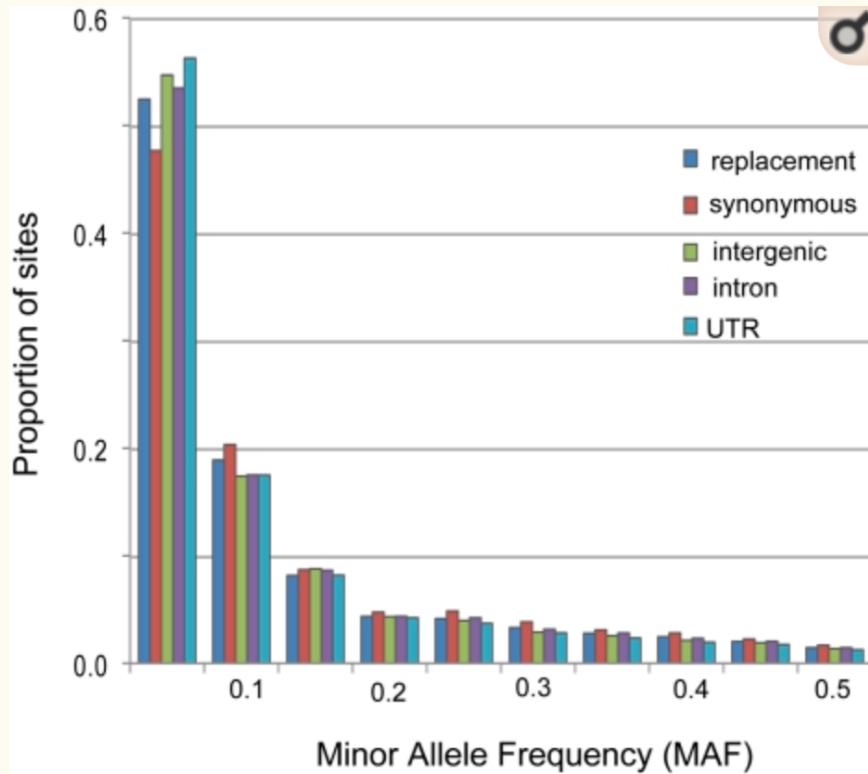


Fig. 2.

MAF at replacement, synonymous, intergenic, intron, and UTR sites.

- Image from Branca et al. "Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*", Proc Natl Acad Sci U S A. 2011 Oct 18; 108(42): E864-E870.

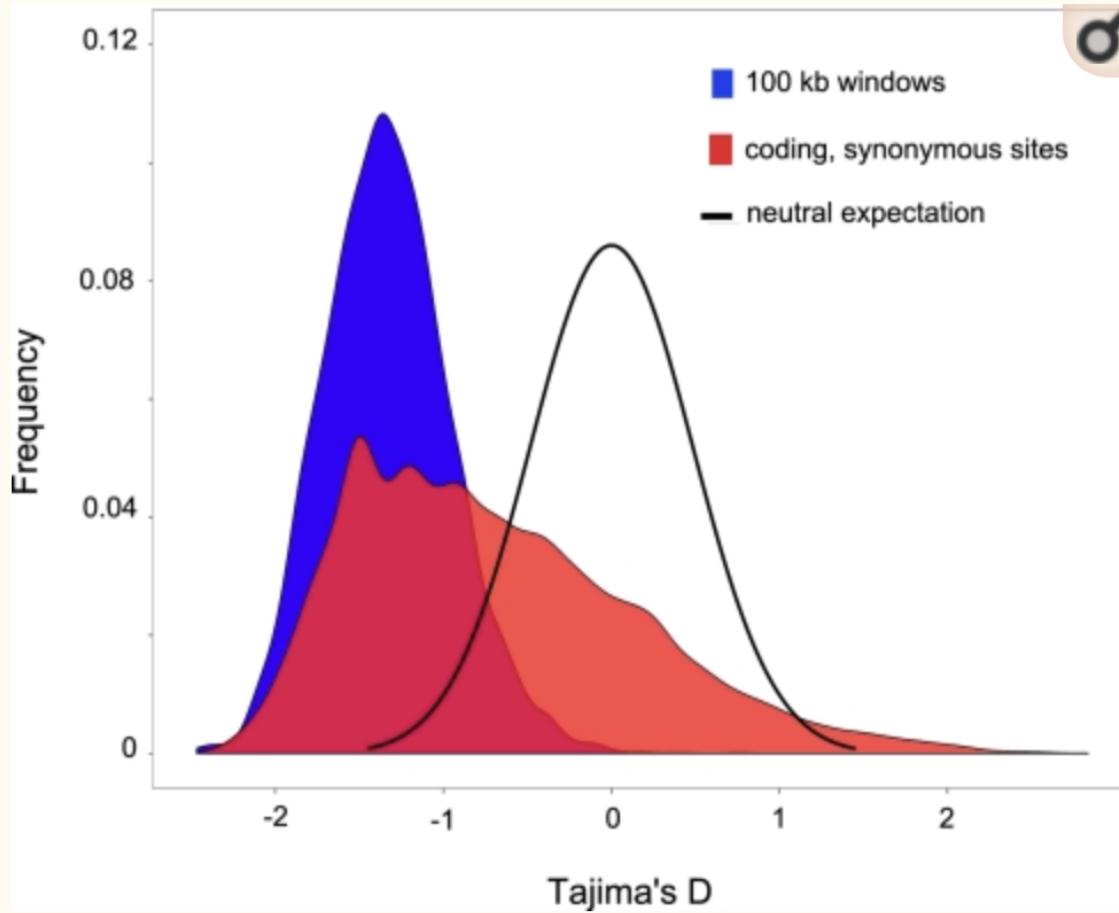
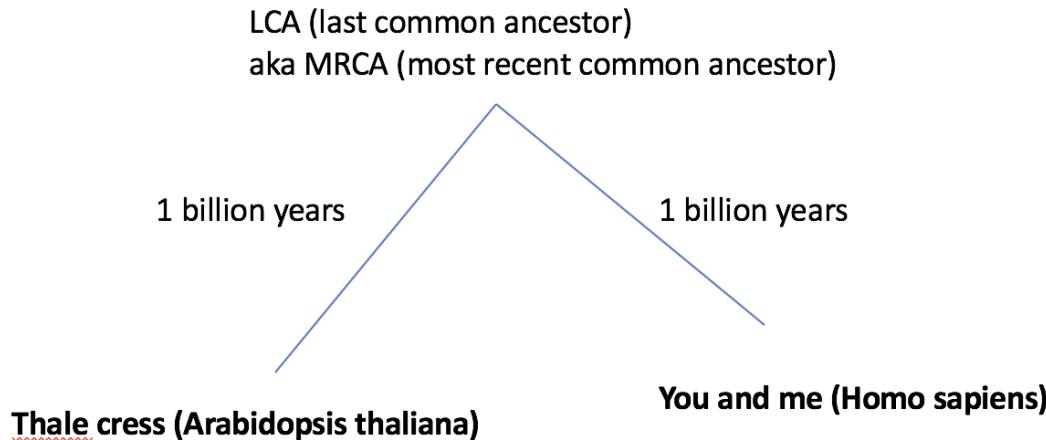


Fig. 3.

Distributions of Tajima's D statistic for 100-kb sliding windows (blue) and synonymous sites found in the 23,468 gene models with more than or equal to two segregating sites. The black line shows the expected distribution of D with no selection in a panmictic population of constant size.

- Image from Branca et al. "Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*", Proc Natl Acad Sci U S A. 2011 Oct 18; 108(42): E864-E870.

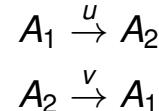
Molecular evolution



- There are 2 differences in 83 amino acids between the aligned portions of histone H4 from *A. thaliana* and *H. sapiens* – this discounts the portions that are poorly aligned.
- It is believed that the MRCA (LCA) of plants and animals lived one billion (10^9) years ago, so each branch has length 10^9 years.
- Thus mutation rate $u = \frac{2}{2 \times 10^9} = 10^{-9}$, and the mutation rate per amino acid is thus $u = \frac{10^{-9}}{83} = 0.012 \times 10^{-9} = 1.2 \times 10^{-11}$. NOTE: add branch lengths between the species, not simply length to MRCA
- Histone H4 is one of the proteins found in the histone barrel about which the DNA in eukaryotic nucleus is wrapped – this explains the extreme conservation.
- Usually, nucleotide mutation rates are given per nucleotide, rather than per amino acid (as in book)

Wright's formula for A_1 allele probability distribution at equilibrium

- Evolutionary model: 2 alleles with reversible mutation rates u, v , where $0 \leq u + v \ll 1$



- Next generation A_1 allele frequency p' given by

$$\begin{aligned} p' &= (1 - u)p + v(1 - p) = p(1 - u - v) + v \\ \Delta p &= p' - p = -up - vp + v = -up + v(1 - p) \end{aligned}$$

The *genetic drift* of A_1 allele frequency is defined to be Δp , where in general Δp depends both on p, t , however in this chapter Δp depends only on p . At equilibrium, $\Delta p = 0$, in which case, the expected value of the stationary distribution of A_1 allele frequency is

$$\Delta p = -up + v(1 - p) = 0 \Rightarrow p = \frac{v}{u + v}$$

- By using the Fokker-Planck equation (also called the forward Kolmogorov equation of 1-dimensional diffusion), Wright showed that the probability density for the A_1 allele frequency at equilibrium is

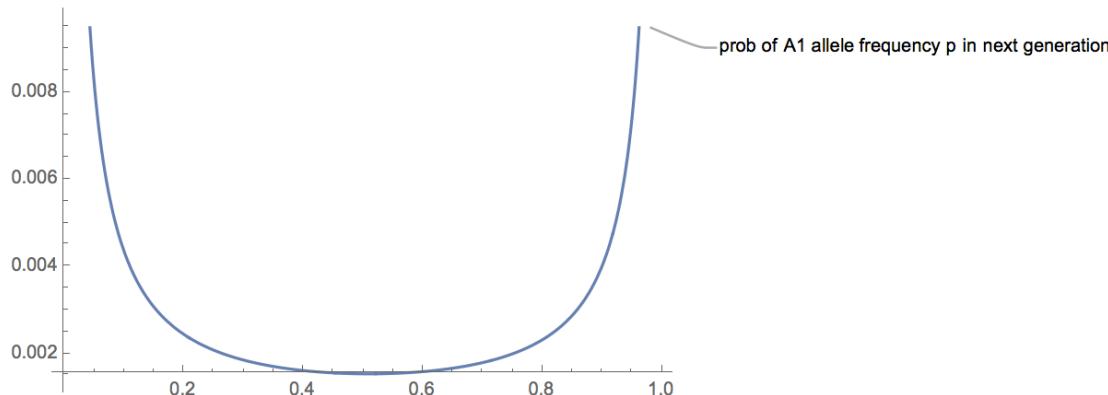
$$\Phi(p) = c \cdot \frac{2 \exp\left(\int^p \frac{m(x)}{v(x)} dx\right)}{v(p)}$$

- c is a normalization constant that ensures $\int_0^1 \Phi(p) dp = 1$, hence

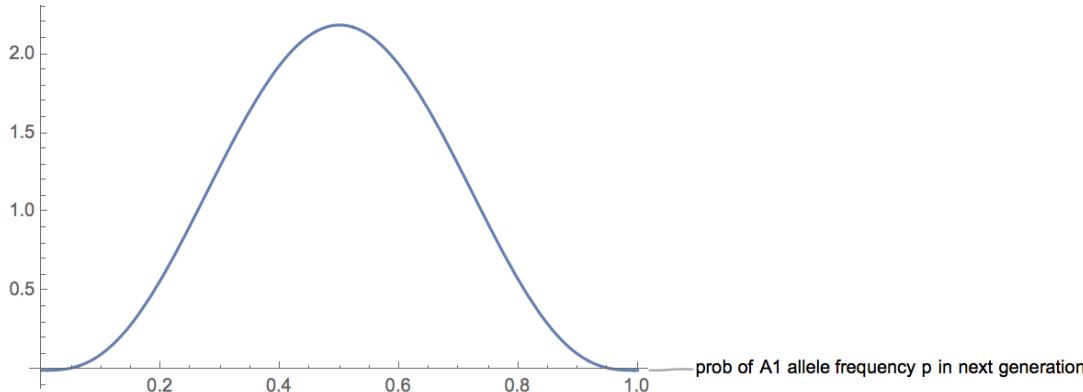
$$c = \left(\int_0^1 \frac{2 \exp\left(\int_0^1 \frac{m(x)}{v(x)} dx\right)}{v(p)} dp \right)^{-1}$$

- Here $x \in [0, p]$, for $p \in [0, 1]$. No lower bound given for the definite integral - intent is that lower bound $\epsilon \approx 0$, yet is strictly positive to ensure $\ln(\epsilon)$ defined.

Mathematica program to compute A_1 allele probability distribution at equilibrium



```
Ne = 10000;
u = 10-6;
v = 10-8;
c = Gamma[4 Ne u + 4 Ne v] / (Gamma[4 Ne u] Gamma[4 Ne v]);
Phi[p_] := cp4Ne v - 1(1 - p)4Ne u - 1;
Plot[Phi[p], p, 0, 1,
 PlotLabels -> "prob of A1 allele frequency p in next generation"]
meanOfP = N[Integrate[p *Phi[p], p, 0, 1]];
varianceOfP = N[Integrate[p2 *Phi[p], p, 0, 1] - meanOfP2];
Print["m(p)=", meanOfP] (* m(p) = 0.99099 *)
Print["v (p) = ", varianceOfP] (* v(p) = 0.009422 *)
```



- recall that at equilibrium

$$\Delta p = -up + v(1-p) = 0 \Rightarrow p = \frac{v}{u+v}$$

- in last slide, we showed an image of the stationary distribution for $N_e = 10000$, $u = 10^{-6}$, $v = 10^{-8}$. For these values of u, v , we have

$$p = \frac{v}{u+v} = \frac{10^{-8}}{10^{-6} + 10^{-8}} = 0.009900990099009903$$

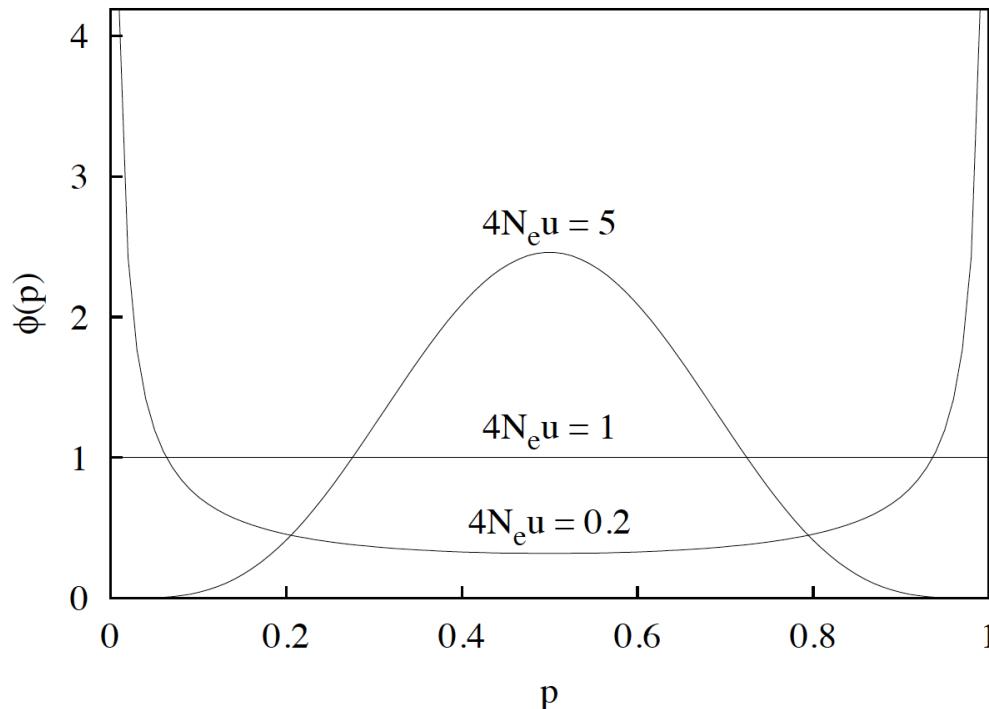
$$q = (1-p) = \frac{u}{u+v} = \frac{10^{-6}}{10^{-6} + 10^{-8}} = (1 - 0.009900990099009903) = 0.9900990099$$

- image above is for values $N_e = 10000$, $u = 10^{-4} = v$, hence at equilibrium

$$p = \frac{v}{u+v} = \frac{1}{2}$$

This value is the expected value of the distribution, but clearly, since molecular evolution is stochastic, it is possible that p at any time is quite different than $\frac{1}{2}$

Equilibrium distribution for A_1 allele frequency p with symmetric reversible mutation



- Figure shows distribution $\Phi(p)$ at equilibrium, when genetic drift balances mutation (balancing mutation), when forward mutation rate u equals backward mutation rate v

$$\begin{aligned}\Phi(p) &= c \cdot \frac{2 \exp \left(\int^p \frac{m(x)}{v(x)} dx \right)}{v(p)} \\ &= \left(\frac{\Gamma(4N_e u + 4N_e v)}{\Gamma(4N_e u) \Gamma(4N_e v)} \right) \cdot p^{4N_e v - 1} \cdot (1 - p)^{4N_e u - 1}\end{aligned}$$

Derivation of Wright's formula in the case of reversible mutation

- mean value of genetic drift

$$m(p) = \Delta p = -up + v(1 - p)$$

- variance of genetic drift

$$v(p) = \frac{pq}{2N_e}$$

- normalization constant c in case of reversible mutation

$$\begin{aligned} c &= \left(\int_0^1 \frac{2 \exp \left(\int_0^p \frac{m(x)}{v(x)} dx \right)}{v(p)} dp \right)^{-1} \\ &= \frac{\Gamma(4N_e u + 4N_e v)}{\Gamma(4N_e u)\Gamma(4N_e v)} \end{aligned}$$

where for complex $z = a + bi$, such that $a > 0$,

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

The Gamma function is an extension of the factorial function to complex numbers, where we recall that $n! = n(n - 1) \cdots 1$ is defined only for non-negative integers. Indeed, for positive integer n ,

$$\Gamma(n) = (n - 1)!$$



$$\begin{aligned} \Phi(p) &= c \cdot \frac{2 \exp \left(\int_0^p \frac{m(x)}{v(x)} dx \right)}{v(p)} \\ &= \left(\frac{\Gamma(4N_e u + 4N_e v)}{\Gamma(4N_e u)\Gamma(4N_e v)} \right) \cdot p^{4N_e v - 1} \cdot (1 - p)^{4N_e u - 1} \end{aligned}$$

Derivation of $\pi = E[\sum_{i=1}^{\infty} 2p_i(1 - p_i)] = 4N_e u$



$$\Phi(p) = \left(\frac{\Gamma(4N_e u + 4N_e v)}{\Gamma(4N_e u)\Gamma(4N_e v)} \right) \cdot p^{4N_e v - 1} \cdot (1 - p)^{4N_e u - 1}$$

is a beta distribution with shape parameters $\alpha = 4N_e v$ and $\beta = 4N_e u$, hence under the hypothesis of symmetric reversible mutation, we have $u = v$, so $\alpha = \beta = 4N_e u = \theta$ and

$$\mu = \frac{\alpha}{\alpha + \beta} = \frac{\theta}{\theta + \theta} = \frac{1}{2}$$

$$\sigma^2 = \frac{1}{4(2\beta + 1)} = \frac{1}{4(2\theta + 1)}$$

$$V[p] = E[p^2] - E[p]^2$$

$$E[p^2] = V[p] + E[p]^2$$

$$\begin{aligned} E[2p(1 - p)] &= 2(E[p] - E[p^2]) = 2(\mu - (\sigma^2 + \mu^2)) \\ &= 2\left(\frac{1}{2} - \left(\frac{1}{4(2\theta + 1)} + \frac{1}{4}\right)\right) 2\left(\frac{1}{4} - \frac{1}{4(2\theta + 1)}\right) \\ &= 2\left(\frac{2\theta + 1 - 1}{4(2\theta + 1)}\right) = \frac{\theta}{2\theta + 1} \end{aligned}$$

- We've just shown that for a specific site s , the expected value $E[2p(1 - p)]$ at that site equals $\frac{\theta_s}{1+2\theta_s}$, where θ_s denotes $4N_e u_s$ and u_s is probability of mutation at site s . However, since u is the overall DNA mutation rate for the DNA sequences under study, if there are m mutation sites, then $\theta_i = \frac{\theta}{m}$, and

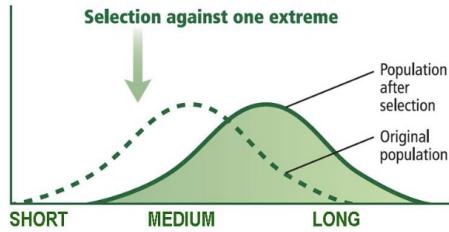
$$E\left[\sum_{i=1}^m 2p_i(1 - p_i)\right] = \sum_{i=1}^m \frac{\theta/m}{2\theta/m + 1} = \frac{\theta}{2\theta/m + 1}$$

$$\pi = \lim_{m \rightarrow \infty} E\left[\sum_{i=1}^m 2p_i(1 - p_i)\right] = \theta$$

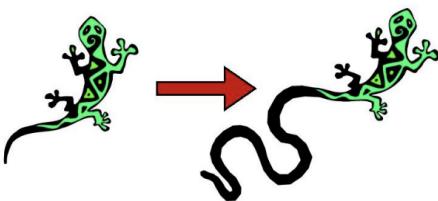
Chapter 3: Natural selection

HOW does the trait change?

Directional Selection

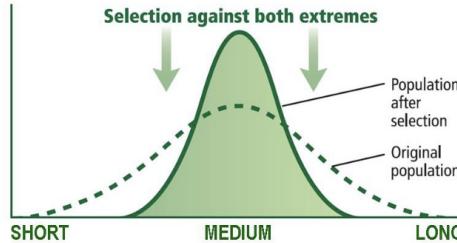


FOR: one extreme trait
AGAINST: the other extreme



EX. Long wiggly tails look like a snake and scare predators. The longer the tail, the more it looks like a snake.

Stabilizing Selection

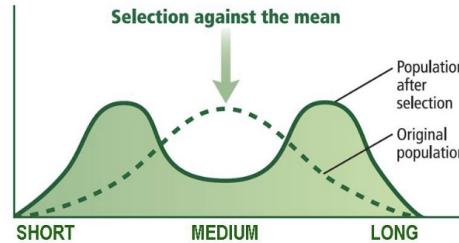


FOR: moderate traits
AGAINST: both extremes

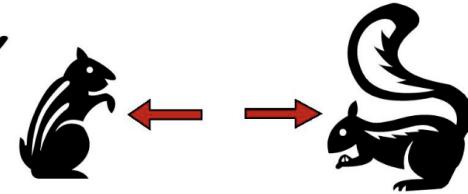


EX. Short tails mess up the cat's balance. Long tails drag on the ground. Medium tails are best.

Disruptive Selection



FOR: both extremes
AGAINST: moderate traits



EX. Short tails help keep predators from catching you on the ground. Long tails are good for balance in the trees. Medium tails don't help.

(Image from Socrative)

Mathematical model of fitness and selection

Current frequencies for alleles A_1, A_2 at A locus

$$p = \text{Prob}(A_1)$$

$$q = 1 - p = \text{Prob}(A_2)$$

$$p + q = 1$$

- **Fitness, or viability of genotype:** probability of survival of genotype in a population; alternatively, fraction of individuals that survive

| genotype | viability (fitness) | probability |
|----------|---------------------|-------------------|
| A_1A_1 | $w_{1,1}$ | p^2 |
| A_1A_2 | $w_{1,2}$ | $2pq = 2p(1 - p)$ |
| A_2A_2 | $w_{2,2}$ | $q^2 = (1 - p)^2$ |

- **expected fitness of population:** \bar{w} :

$$\bar{w} = p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2}$$

● **A_1 allele frequency of next generation p'**

▷ In **absence** of selection and mutation,

$$p' = p^2 + pq = p(p + q) = p$$

▷ In **presence** of selection but **without** mutation,

$$p' = \frac{p^2 w_{1,1} + pq w_{1,2}}{\bar{w}} = \frac{p^2 w_{1,1} + pq w_{1,2}}{p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2}}$$

● Hence in presence of selection without mutation the change Δp in A_1 allele frequency in one generation, caused by selection satisfies

$$\begin{aligned}\Delta p &= p' - p = \frac{p^2 w_{1,1} + pq w_{1,2}}{\bar{w}} - \frac{p \bar{w}}{\bar{w}} = \frac{p^2 w_{1,1} + pq w_{1,2} - p(p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2})}{\bar{w}} \\ &= \frac{p}{\bar{w}} (p(1-p)w_{1,1} + qw_{1,2}(1-2p) - q^2 w_{2,2}) = \frac{p}{\bar{w}} (pq w_{1,1} + qw_{1,2}(1-p) - pq w_{1,2} - q^2 w_{2,2}) \\ \Delta p &= \frac{p}{\bar{w}} (pq w_{1,1} + q^2 w_{1,2} - pq w_{1,2} - q^2 w_{2,2}) = \frac{pq}{\bar{w}} (p(w_{1,1} - w_{1,2}) - q(w_{2,2} - w_{1,2}))\end{aligned}$$

● In words, Δp is $\frac{pq}{\bar{w}}$ times the difference of the **incremental fitness** of the homozygote $A_1 A_1$ over the heterozygote $A_1 A_2$ (weighted by current allele frequency of A_1) and the **incremental fitness** of the homozygote $A_2 A_2$ over the heterozygote $A_1 A_2$ (weighted by current allele frequency of A_2).

Absolute and relative fitness

- Recall that **absolute fitness** $w \in [0, 1]$ is the proportion of individuals of a given genotype that survive
- Note that equation for Δp is **unchanged** if
 - $w_{1,1}$ is replaced by $\frac{w_{1,1}}{\max(w_{1,1}, w_{1,2}, w_{2,2})}$ (or instead if $w_{1,1}$ is replaced by $\frac{w_{1,1}}{w_{1,1}}$, etc.)
 - $w_{1,2}$ is replaced by $\frac{w_{1,2}}{\max(w_{1,1}, w_{1,2}, w_{2,2})}$ (or instead if $w_{1,2}$ is replaced by $\frac{w_{1,2}}{w_{1,1}}$, etc.)
 - $w_{2,2}$ is replaced by $\frac{w_{2,2}}{\max(w_{1,1}, w_{1,2}, w_{2,2})}$ (or instead if $w_{2,2}$ is replaced by $\frac{w_{2,2}}{w_{1,1}}$, etc.)

since

$$\bar{w} = p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2}$$

then becomes

$$\frac{\bar{w}}{\max(w_{1,1}, w_{1,2}, w_{2,2})} = p^2 \frac{w_{1,1}}{\max(w_{1,1}, w_{1,2}, w_{2,2})} + 2pq \frac{w_{1,2}}{\max(w_{1,1}, w_{1,2}, w_{2,2})} + q^2 \frac{w_{2,2}}{\max(w_{1,1}, w_{1,2}, w_{2,2})}$$

and term $\max(w_{1,1}, w_{1,2}, w_{2,2})$ cancels, since it appears both in the numerator and denominator

- Same argument shows that one can replace $w_{i,j}$ by $\frac{w_{i,j}}{c}$ for any positive constant
- Relative fitnesses:** absolute fitnesses divided by the term $w_{1,1}$ (usual case), or by $\max(w_{1,1}, w_{1,2}, w_{2,2})$, etc.

Fitness model with parameters s, h

- One widely used fitness model uses parameters for **selection coefficient** $s \in [0, 1]$: this model requires that allele A_1 is fitter than allele A_2 – otherwise alleles are relabeled to ensure $\text{fitness}(A_1) \geq \text{fitness}(A_2)$
- heterozygosity effect** h : positive or negative values not necessarily bounded by 1 in absolute value, i.e. $h \in \mathbb{R}$

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|------------------|-----------|-----------|-----------|
| absolute fitness | $w_{1,1}$ | $w_{1,2}$ | $w_{2,2}$ |
| relative fitness | 1 | $1 - hs$ | $1 - s$ |

- Selection factors s are generally small, such as $s = 0.001$, but for illustrative purposes, we consider sometimes $s = 0.1$
- Example:** If the number of offspring of individuals of genotype A_1A_1 , A_1A_2 , and A_2A_2 is respectively 4, 2.5 and 0.5, then what are the values of the selection coefficient s and the heterozygous effect h ?
- Answer:** Relative fitnesses of genotypes A_1A_1 , A_1A_2 , and A_2A_2 are respectively $\frac{4}{4} = 1$, $\frac{2.5}{4} = 0.625$, $\frac{0.5}{4} = 0.125$. Homozygote A_2A_2 relative fitness is $1 - s = 0.125$, so $s = 0.875$, and heterozygote A_1A_2 relative fitness is $1 - hs = 1 - 0.875h = 0.625$, so $h = 0.42857142857142855 \approx 0.4286$.
- Notes:
 - in diploid organisms, selection applies to **genotypes** not **alleles** – Dawkin's "selfish gene"?
 - We generally adopt the convention that $w_{1,1} \geq w_{2,2}$, and so relative fitness is usually given by dividing absolute fitness by $w_{1,1}$
 - selection coefficient $s \in [0, 1]$ and is usually small, e.g. $s = 0.1$, $s = 0.001$, etc.
 - when $h = \frac{1}{2}$, alleles are said to be **additive** – i.e. the fitness of the heterozygote is halfway between that of the homozygotes

- For the selection model with parameters h, s , we write $\Delta_s p$ to denote the change $p' - p$ of A_1 allele frequency in one generation due to selection.
- Recall the equation for the change $\Delta_s p$ in frequency of allele A_1

$$\begin{aligned}\Delta_s p &= \frac{pq}{\bar{w}} (p(w_{1,1} - w_{1,2}) - q(w_{2,2} - w_{1,2})) \\ &= \frac{pq}{\bar{w}} (p(1 - (1 - hs)) - q((1 - s) - (1 - hs))) \\ \Delta_s p &= \frac{pq s}{\bar{w}} (ph - q(1 - h))\end{aligned}$$

- Since $p, q, s, \bar{w} \in [0, 1]$ are all non-negative, the sign of $\Delta_s p$ uniquely determined by $ph - q(1 - h)$, i.e. whether allele p increases or decreases under evolution depends on heterozygosity effect h , not on selection factor s (here, recall our assumption that allele A_1 is fitter than A_2 , and if not, we relabel alleles to ensure this)
- For the relative fitness model with parameters s, h , the expected fitness of the population is given by the following.

$$\begin{aligned}\bar{w} &= p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2} \\ &= p^2 + 2pq(1 - hs) + q^2(1 - s) = p^2 + 2pq - 2pqhs + q^2 - q^2 s \\ &= (p^2 + 2pq + q^2) - 2pqhs - sq^2 = 1 - 2pqhs - q^2 s\end{aligned}$$

Note that if selection $s = 0$ then expected fitness is the optimal value 1, i.e. homozygotes and heterozygotes equally fit.

- equilibrium value of allele p : value of p when $\Delta_s p = 0$

$$\Delta_s p = \frac{pq s}{w} (ph + q(1-h)) = 0$$

If $s = 0$, there is no fitness difference and so allele frequencies p, q never change from initial frequencies (case of no mutation or selection). Assuming $s > 0$, there are 3 solutions to the 3rd degree polynomial equation $\Delta_s p = 0$. Following the simple manual solution, please look at the Mathematica code that illustrates how you factor and (symbolically) solve an equation.

- ▷ solution 1: $p = 0, q = 1$
- ▷ solution 2: $p = 1, q = 0$
- ▷ solution 3: Assuming p, q, s all are non-zero

$$\begin{aligned} pq s(ph + q(1-h)) = 0 &\Leftrightarrow ph + q(1-h) = 0 \Leftrightarrow ph + (1-p)(1-h) = 0 \Leftrightarrow 2ph + 1 - h - p = 0 \\ &\Leftrightarrow p(2h - 1) = h - 1 \Leftrightarrow p = \frac{h - 1}{2h - 1} = \frac{1 - h}{1 - 2h} \end{aligned}$$

```

: DeltaP = p q s (p h + q (1 - h));
: q = 1 - p;
Factor[DeltaP]
Solve[DeltaP == 0, p]

: -0.2 p (-0.5 + 0.5 h + 1. p - 1.5 h p - 0.5 p2 + 1. h p2)
: { {p → 0}, {p → 1}, {p →  $\frac{-1+h}{-1+2h}$ } }

```

- Summarizing, the three solutions are

$$\begin{aligned}
 p &= 0 & q &= 1 \\
 p &= 1 & q &= 0 \\
 p &= \frac{h-1}{2h-1} & q &= \frac{h}{2h-1}
 \end{aligned}$$

where the third case obtains only for the cases of overdominance ($h < 0$) or underdominance ($h > 1$). In the case of incomplete dominance $0 < h < 1$, the sign of $\Delta_s p$ is positive, since

$$\Delta_s p = \frac{pq}{W} (ph + q(1-h)) > 0$$

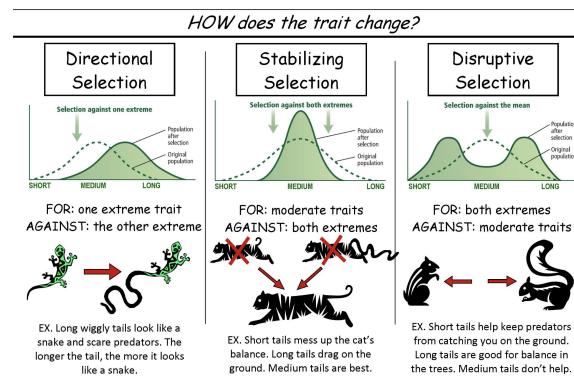
and so A_1 allele frequency p gradually increases until it reaches the stable equilibrium $p = 1$.

Heterozygosity effect

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|------------------|-----------|-------------------|-------------------|
| absolute fitness | $w_{1,1}$ | $w_{1,2}$ | $w_{2,2}$ |
| relative fitness | 1 | $w_{1,2}/w_{1,1}$ | $w_{2,2}/w_{1,1}$ |
| relative fitness | 1 | $1 - hs$ | $1 - s$ |

where $1 - hs = \frac{w_{1,2}}{w_{1,1}}$ and $1 - s = \frac{w_{2,2}}{w_{1,1}}$. Type of dominance and type of natural selection depends on h

| heterozygosity effect | type of dominance | type of selection |
|-----------------------|--|--|
| $h = 0$ | A_1 dominant, A_2 recessive | directional evolution, converging to $p = 1$ |
| $h = 1$ | A_2 dominant, A_1 recessive | directional evolution, converging to $p = 1$ |
| $0 < h < 1$ | incomplete dominance | directional evolution, converging to $p = 1$ |
| $h < 0$ | overdominance (heterozygote is fittest) | balancing selection (e.g. sickle cell trait) |
| $h > 1$ | underdominance (homozygotes are fittest) | disruptive selection |



A_1 allele equilibrium frequency

- Recall that selection coefficient satisfies $0 \leq s \leq 1$. Since fitness is always **non-negative**, it follows that **expected fitness** satisfies $\bar{w} \geq 0$ and that **heterozygote fitness** satisfies $1 - hs \geq 0$, and so $h \leq \frac{1}{s}$.
- Recall that change in A_1 allele frequency $\Delta p = p' - p$ satisfies

$$\Delta p = \frac{pq s(ph + q(1 - h))}{\bar{w}} = \frac{pq s(ph + q(1 - h))}{1 - 2pqhs - q^2s}$$

hence **equilibrium** A_1 allele frequency is reached when $\Delta p = 0$, which occurs when $pq s(ph + q(1 - h)) = 0$

- Case 1:** $s = 0$ If selection coefficient is 0, there is no selection, and by Hardy-Wright theorem, there is no change of A_1 allele frequency – the A_1 allele frequency p never changes from its initial value. This is easily seen, since

$$w_{1,1} = 1$$

$$w_{1,2} = 1 - hs = 1$$

$$w_{2,2} = 1 - s = 1$$

$$\bar{w} = p^2 \cdot w_{1,1} + 2pq \cdot w_{1,2} + q^2 \cdot w_{2,2} = p^2 + 2pq + q^2 = (p + q)^2 = 1$$

$$p' = p^2 \cdot \frac{w_{1,1}}{\bar{w}} + pq \cdot \frac{w_{1,2}}{\bar{w}} = p^2 + pq = p(p + q) = p$$

- Case 2:** $p = 0$ and $q = 1$

$$\bar{w} = p^2 \cdot w_{1,1} + 2pq \cdot w_{1,2} + q^2 \cdot w_{2,2} = w_{2,2}$$

$$p' = p^2 \cdot \frac{w_{1,1}}{\bar{w}} + pq \cdot \frac{w_{1,2}}{\bar{w}} = 0$$

- It follows that if $p = 0$, then in all future generations, $p = 0$ and $q = 1$. As we'll soon see, $p = 0$ is not a **stable equilibrium**, since if $p > 0$ (even $p = 10^{-6}$ or less), then A_1 allele frequency will gradually increase until it reaches the stable equilibrium $p = 1$.

■ **Case 3: $p = 1$ and $q = 0$**

$$\bar{w} = p^2 \cdot w_{1,1} + 2pq \cdot w_{1,2} + q^2 \cdot w_{2,2} = w_{1,1}$$

$$p' = p^2 \cdot \frac{w_{1,1}}{\bar{w}} + pq \cdot \frac{w_{1,2}}{\bar{w}} = p^2 = 1^2 = 1$$

It follows that if $p = 1$, then in all future generations, $p = 1$. We'll soon see that $p = 1$ is a **stable equilibrium**, since if $p < 1$ is sufficiently close to 1, then A_1 allele frequency will increase until $p = 1$.

■ **Case 4: $0 < p, q < 1$ and $ph + q(1 - h) = 0$, whereby $p = \frac{1-h}{1-2h}$**

$$\Delta p = \frac{pq s}{\bar{w}} \cdot [ph + q(1 - h)]$$

$$\Delta p > 0 \Leftrightarrow ph + q(1 - h) > 0$$

It follows that A_1 allele frequency increases in the case of **incomplete dominance**, since if $0 < h < 1$, then $\Delta p > 0$.

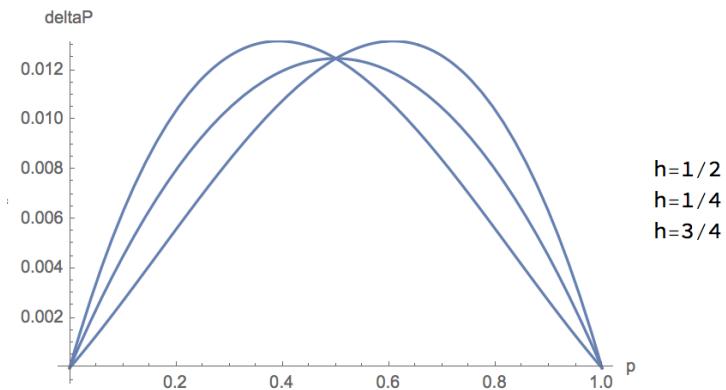
■ Moreover, in this case, we can **not** have an **equilibrium** frequency $p = 1/2 = q$, since $ph + q(1 - h)$ would equal $h/2 - h/2 + q$, and under the current hypothesis we would not have $ph + q(1 - h) = 0$ as assumed. Hence

$$ph + q(1 - h) = 0 \Leftrightarrow h(p - q) = -q \Leftrightarrow h = \frac{q}{q - p}$$

■ For example, if $p = \frac{1}{4}$ and $q = \frac{3}{4}$, then for heterozygosity effect $h = \frac{3/4}{3/4 - 1/4} = 1.5$ causing **underdominance**, and we have $\Delta p = 0$, hence an equilibrium. Is this equilibrium stable?

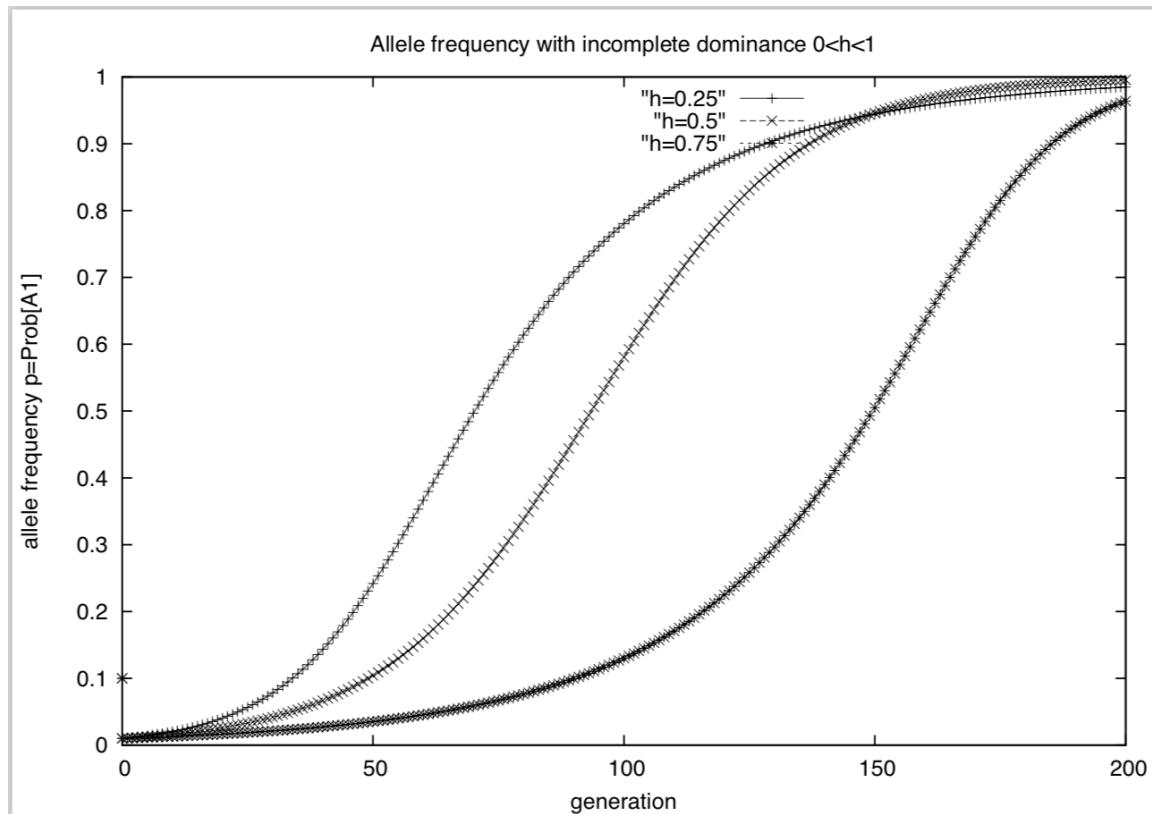
■ Similarly, if $p = \frac{3}{4}$ and $q = \frac{1}{4}$, then for heterozygosity effect $h = \frac{1/4}{1/4 - 3/4} = -1.5$ causing **overdominance** (sickle cell trait), and we have $\Delta p = 0$, hence an equilibrium. Is this equilibrium stable?

Δp plotted as function of p (Incomplete dominance $0 < h < 1$)



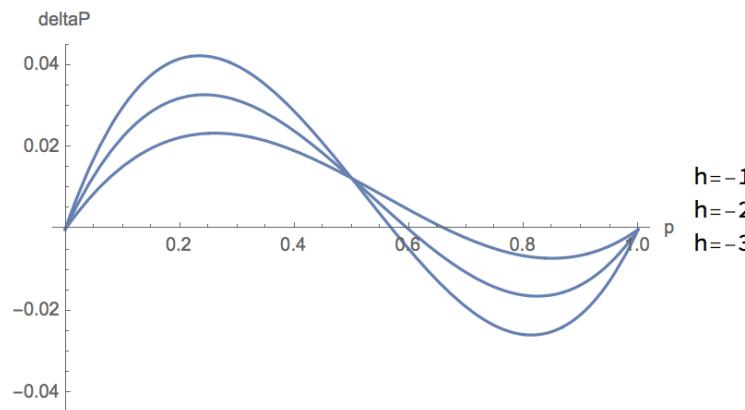
- In the case of incomplete dominance, note that for smaller h (closer to 0), there is less of a penalty for heterozygotes, so heterozygotes provide a larger contribution to A_1 allele frequency p . Similarly, for larger h (closer to 1), there is a greater penalty for heterozygotes, so heterozygotes provide a smaller contribution to A_1 allele frequency p .
- in the case of **incomplete dominance**, where $0 < h < 1$, there is **balancing selection**, ensuring stable frequencies $0 < p, q < 1$ of both alleles (hence the name “balancing” selection)

A_1 allele frequency as function of generation (Incomplete dominance $0 < h < 1$)



In the case of incomplete dominance, note that for smaller h (closer to 0), there is less of a penalty for heterozygotes, so heterozygotes provide a larger contribution to A_1 allele frequency p . Similarly, for larger h (closer to 1), there is a greater penalty for heterozygotes, so heterozygotes provide a smaller contribution to A_1 allele frequency p .

Δp plotted as function of p (overdominance $h < 0$)



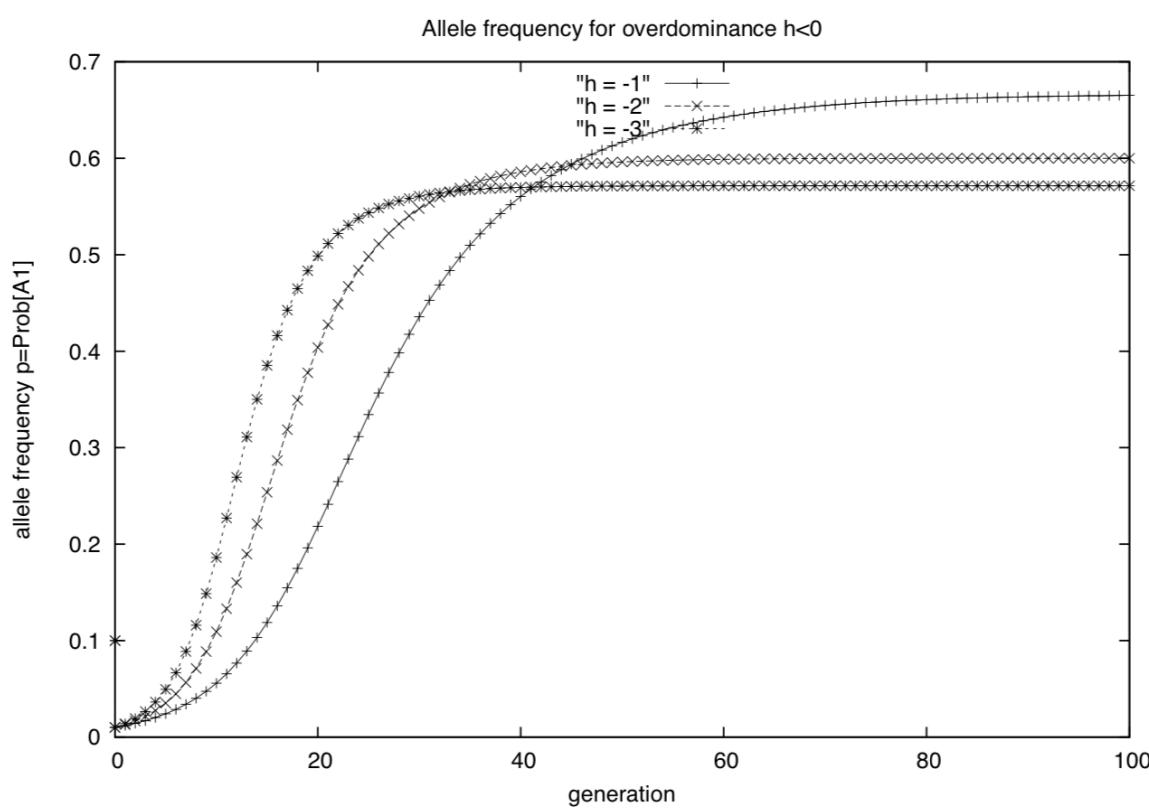
- In the case of overdominance, where $h < 0$, note that for smaller h (negative, but larger in absolute value), the x -intercept lies to the left, since the third equilibrium value of p , where $\Delta p = 0$, is given by $p = \frac{h-1}{2h-1}$. Hence for $h_1, h_2 < 0$, we have

$$\begin{aligned} \frac{h_2 - 1}{2h_2 - 1} &< \frac{h_1 - 1}{2h_1 - 1} \Leftrightarrow (h_2 - 1)(2h_1 - 1) < (h_1 - 1)(2h_2 - 1) \quad (\text{inequality flips when multiply by neg number}) \\ &\Leftrightarrow 2h_1h_2 - 2h_1 - h_2 + 1 < 2h_1h_2 - 2h_2 - h_1 + 1 \\ &\Leftrightarrow h_2 < h_1 \end{aligned}$$

Thus x -intercept for $h = -3$ curve lies to the left of that for $h = -2$ curve, which lies to the left of that for $h = -1$.

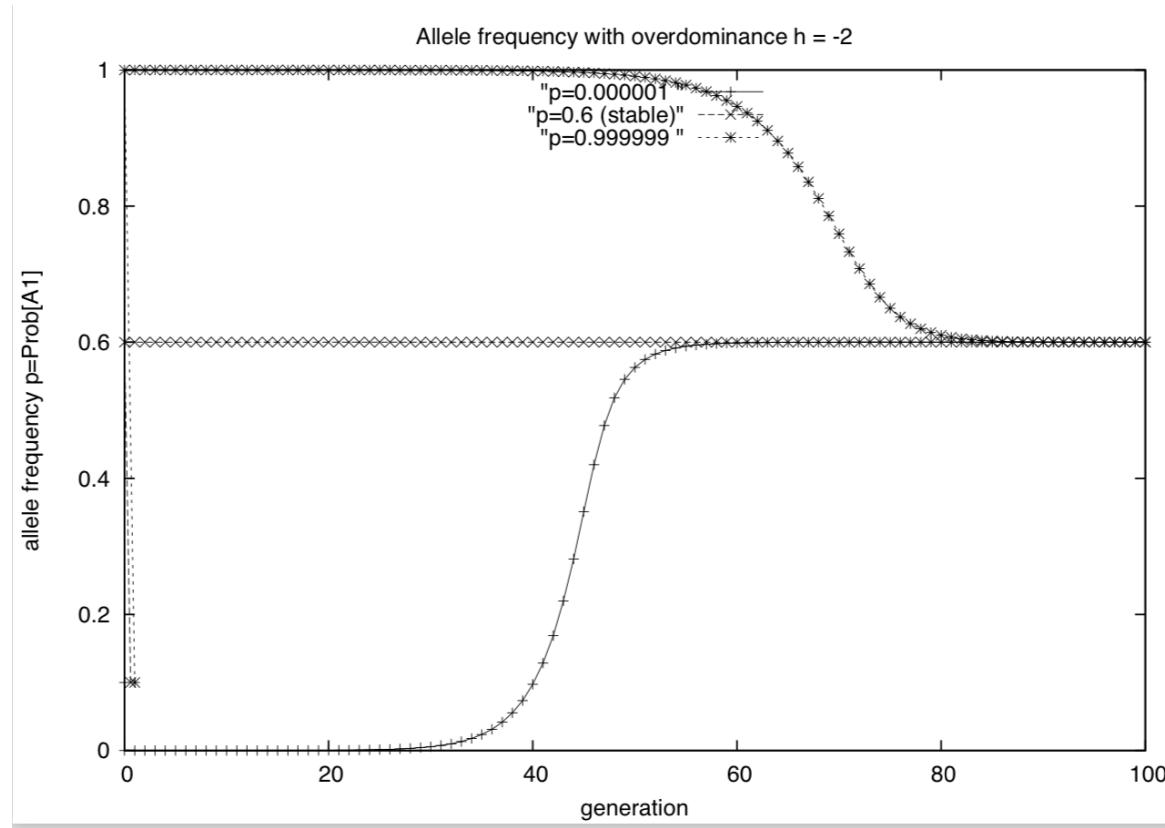
- Consider the overdominance case where heterozygosity effect $h = -2$. There are 3 equilibrium values of p , i.e. $p = 0$ (homozygous for A_2), $p = 1$ (homozygous for A_1), $p = \frac{h-1}{2h-1} = \frac{-2-1}{-4-1} = 3/5 = 0.6$.
- By graphical analysis, Δp is positive for values of $p \in [0, 0.6]$, so any initial value of $0 < p < 0.6$ will ultimately converge to 0.6. Similarly Δp is negative for values of $p \in (0.6, 1]$, so any initial value of $0.6 < p < 1$ will ultimately converge to 0.6. It follows that $p = 0.6$ is a stable equilibrium, while $p = 0$ and $p = 1$ are unstable equilibria.

A_1 allele frequency as function of generation (overdominance $h < 0$)



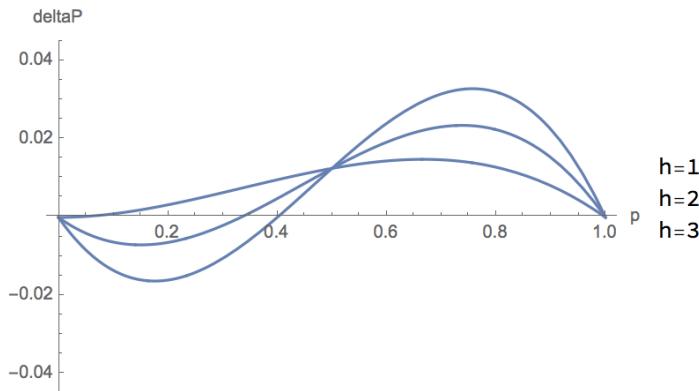
- In the case of overdominance, where $h < 0$, note that for smaller h values (negative, but larger in absolute value), the equilibrium A_1 frequency p is smaller. This is for exactly the same reason explained in the previous slide.
- Note that the equilibrium frequency of p lies strictly between 0 and 1, so that $0 < p, q < 1$. Hence this is **balancing selection**.

A_1 allele frequency for $h = -2$ and $p = \epsilon$, $p = 0.6$, and $p = 1 - \epsilon$ (overdominance $h < 0$)



In the case of overdominance, where in this case $h = -2$, if the initial value of $p \in (0, 1)$, the asymptotic value of p is 0.6 (of course, if initial value of $p = 0.6$, then subsequent values of p never change). In contrast, if the initial value of p is either 0 or 1, then subsequent values of p never change.

Δp plotted as function of p (underdominance $h > 1$)



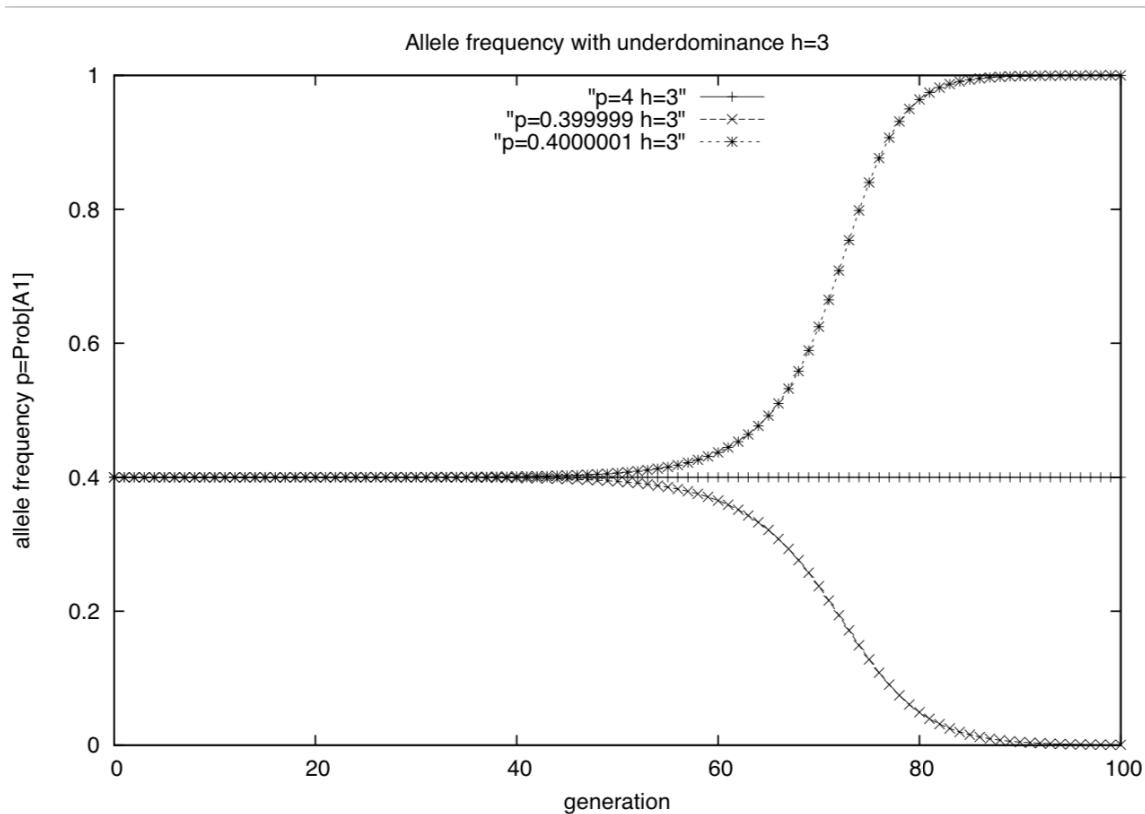
In the case of underdominance, where $h > 1$, note that for larger h values, the corresponding x -intercept lies further to the right, since equilibrium value of p , where $\Delta p = 0$, is given by $p = \frac{h-1}{2h-1}$. Hence

$$\begin{aligned} \frac{h_2 - 1}{2h_2 - 1} &> \frac{h_1 - 1}{2h_1 - 1} \Leftrightarrow (h_2 - 1)(2h_1 - 1) > (h_1 - 1)(2h_2 - 1) \\ &\Leftrightarrow 2h_1h_2 - 2h_1 - h_2 + 1 > 2h_1h_2 - 2h_2 - h_1 + 1 \Leftrightarrow h_2 > h_1 \end{aligned}$$

Thus x -intercept for $h = 3$ curve lies to the right of that for $h = 2$, which lies to the right of that for $h = 1$. Here, it should be stressed that $h = 1$ is not a case of underdominance, but if the initial value of $p = 1$, then all future generations have $p = 1$.

Consider the underdominance case where heterozygosity effect $h = 3$. There are 3 equilibrium values of p , i.e. $p = 0$ (homozygous for A_2 , $p = 1$ (homozygous for A_1 , $p = \frac{h-1}{2h-1} = \frac{3-1}{6-1} = 2/5 = 0.4$). By graphical analysis, Δp is negative for values of $p \in [0, 0.4)$, so any initial value of $p < 0.4$ will ultimately converge to 0. Similarly Δp is positive for values of $p \in (0.4, 1]$, so any initial value of $p > 0.4$ will ultimately converge to 1. It follows that $p = 0.4$ is an unstable equilibrium, while $p = 0$ and $p = 1$ are stable equilibria.

A₁ allele frequency as function of generation (underdominance $h > 0$)



| initial value $p = \text{Prob}[A_1]$ | value of p after 100 generations | asymptotic value of p |
|--------------------------------------|------------------------------------|-------------------------|
| 0.399999 | 0.000373202390326 | 0 |
| $p = 0.4$ | 0.4 | 0.4 |
| 0.400001 | 0.999969226803 | 1 |

As explained on the previous slide, graphical analysis indicates that the equilibria $p = 0$ and $p = 1$ are stable, while for $h = 3$, the remaining equilibrium value $p = \frac{h-1}{2h-1} = 0.4$ is unstable. This is shown clearly by the current figure and table.

A₁ allele change proportional to mean fitness of population

$$\bar{w} = p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2} \quad (\text{absolute fitness})$$

$$\frac{d\bar{w}}{dp} = 2pw_{1,1} + 2w_{1,2}(q - p) - 2qw_{2,2} = 2[p(w_{1,1} - w_{1,2}) + q(w_{1,2} - w_{2,2})]$$

$$\Delta p = \frac{pq}{\bar{w}} [p(w_{1,1} - w_{1,2}) + q(w_{1,2} - w_{2,2})]$$

$$\Delta p = \frac{pq}{2\bar{w}} \frac{d\bar{w}}{dp}$$

Fisher Fundamental Theorem of Natural Selection: change in mean fitness proportional to additive genetic variation in fitness.

According to Gillespie, the Fundamental Theorem is only valid for a theoretical population with simple selection at a single locus, but is not true in general (time-dependent fitness values reflecting perhaps a changing environment). Nevertheless, due to its importance in population genetics, we now give a precise formulation and proof of Fisher's result.

Fisher's Fundamental Theorem

- The goal of this section is to prove a fundamental result of population genetics discovered by R.M. Fisher (1930) that relates the genetic variation of a population with the rate of increase in average fitness of a population. To that end, we will do the following:
 - ▷ define **additive genetic variance**, also called **genic variance**
 - ▷ introduce a **continuous time model** for which we compute the (instantaneous) change of average population fitness $\frac{d\bar{w}}{dt}$
 - ▷ compute the additive genetic variance V_g and prove Fisher's theorem

$$\frac{d\bar{w}}{dt} = V_g$$

- Fisher formulated his **Fundamental Theorem (1930)** as follows:

“The rate of increase of fitness of any organism at any time is equal to its genetic variance at that time.”
- This section follows the approach of Crow and Kimura in Chapters 4 and 5 of **An Introduction to Population Genetics Theory**
- Before defining additive genetic variance, note that the **variance** found in any trait in the population (fitness, intelligence, height, weight, etc.) is determined partly by genetic factors and partly by environmental factors. If genetic and environmental factors are not independent, then their **covariance** must also be considered.

- The **additive genetic variance**, also called **genic variance**, is the variance due solely to genetic factors – i.e. without any contribution due to environment, mutation, etc.
- Although Fisher proved a general form of this fundamental theorem, we here consider only the simple model with random mating hypothesis, but no mutation – since we work with fitness parameters, this model certainly does consider selection.
- We originally defined fitness in terms of genotypes, and not alleles. However, in the table below, α [resp. β] is defined to be the **deviation** from **mean genic fitness** of allele A_1 [resp. A_2] – Fisher called α [resp. β] the **average effect** of allele A_1 [resp. A_2].

| genotype frequency fitness (genotype) fitness (genic value) | A_1A_1 p^2 $w_{1,1} = \frac{\bar{w}}{w + 2\alpha}$ | A_1A_2 $2pq$ $w_{1,2} = \frac{\bar{w}}{w + \alpha + \beta}$ | A_2A_2 q^2 $w_{2,2} = \frac{\bar{w}}{w + 2\beta}$ |
|--|--|---|---|
|--|--|---|---|

- Clearly $a_{1,1} = 2\alpha$, $a_{1,2} = \alpha + \beta$, $a_{2,2} = 2\beta$, but it can also be shown that

$$\alpha = pa_{1,1} + qa_{1,2}$$

$$\beta = qa_{2,2} + pa_{1,2}$$

by computing the values α, β which **minimize** the expected squared distance (a technique known as the **least squares method**)

$$Q(\alpha, \beta) = p^2 (a_{1,1} - 2\alpha)^2 + 2pq (a_{1,2} - \alpha - \beta)^2 + q^2 (a_{2,2} - 2\beta)^2$$

This is done by setting the **partial derivatives** $\frac{\partial Q}{\partial \alpha} = 0$, $\frac{\partial Q}{\partial \beta} = 0$ and then solving this system of 2 equations in 2 unknowns – see the homework for details.

- Here you should notice that the values of α, β which minimize the expected distance are necessarily the same values that minimize the expected squared distance – similar to but different from our observation that the likelihood is maximized at the same location as the log likelihood.

- Note that the average of the deviations from mean fitness is required to be 0, hence

$$p^2 a_{1,1} + 2pqa_{1,2} + q^2 a_{2,2} = 0$$

- Since $a_{1,1} = 2\alpha$, $a_{1,2} = \alpha + \beta$ and $a_{2,2} = 2\beta$, we have

$$\begin{aligned} 0 &= p^2 \cdot (2\alpha) + 2pq \cdot (\alpha + \beta) + q^2 \cdot (2\beta) = 2p^2\alpha + 2pq(\alpha + \beta) + 2q^2\beta \\ &= 2[p^2\alpha + pq\alpha] + 2[q^2\beta + pq\beta] = 2p\alpha(p + q) + 2q\beta(q + p) \\ &= 2(p\alpha + q\beta) \end{aligned}$$

- Thus it follows that

$$p\alpha + q\beta = 0$$

This is an important fact that we will use later – please make note of it. In particular, this fact will be used in the derivation of the formula for additive genetic fitness V_g :

$$p\alpha + q\beta = 0$$

Formula for additive genetic variance

- Let X temporarily denote deviation from mean genic fitness.

$$\begin{aligned}V[X] &= E[X^2] - E[X]^2 \\&= E[X^2] \quad \text{since average deviation from mean fitness is 0}\end{aligned}$$

hence the additive genetic variance (genic variance) $V_g = V[X]$ satisfies

$$\begin{aligned}V_g &= p^2 \cdot (2\alpha)^2 + 2pq \cdot (\alpha + \beta)^2 + q^2 \cdot (2\beta)^2 = 4p^2\alpha^2 + 2pq(\alpha^2 + 2\alpha\beta + \beta^2) + 4q^2\beta^2 \\&= [2p^2\alpha^2 + 2pq\alpha^2] + [2p^2\alpha^2 + 4pq\alpha\beta + 2q^2\beta^2] + [2q^2\beta^2 + 2pq\beta^2] \\&= 2p\alpha[p\alpha + q\alpha] + [2(p\alpha + q\beta)]^2 + 2q\beta[q\beta + p\beta] = 2p\alpha[\alpha(p + q)] + [2(p\alpha + q\beta)]^2 + 2q\beta[\beta(p + q)] \\&= 2p\alpha^2 + 4(p\alpha + q\beta)^2 + 2q\beta^2 = 2p\alpha^2 + 2q\beta^2 = 2(p\alpha^2 + q\beta^2)\end{aligned}$$

- Summarizing, in the bi-allelic case, we have proved that additive genetic variance equals twice the second moment of the deviation of fitness of the alleles:

$$V_g = 2(p\alpha^2 + q\beta^2)$$

- This formula generalizes in the multi-allelic case to the following:

$$V_g = 2 \sum_{i=1}^n p_i \alpha_i^2$$

Expected fitness of alleles A_1, A_2

- Define the expected fitness w_1 [resp. w_2] of allele A_1 [resp. A_2] to be

$$w_1 = \frac{p^2 w_{1,1} + pq w_{1,2}}{p} = pw_{1,1} + qw_{1,2}$$
$$w_2 = \frac{q^2 w_{2,2} + pq w_{1,2}}{q} = qw_{2,2} + pw_{1,2}$$

- Note that expected fitness of genotype A_1A_1 plus one half the expected fitness of genotype A_1A_2 equals $p^2 w_{1,1} + pq w_{1,2}$, and that this represents the contribution of allele A_1 to the entire population fitness. By dividing by A_1 allele frequency p , we obtain the **conditional expected fitness** w_1 of allele A_1 . Similarly for allele A_2 .
- Clearly we have

$$\bar{w} = p \cdot \left[\frac{p^2 w_{1,1} + pq w_{1,2}}{p} \right] + q \cdot \left[\frac{q^2 w_{2,2} + pq w_{1,2}}{q} \right] = p \cdot [pw_{1,1} + qw_{1,2}] + q \cdot [qw_{2,2} + pw_{1,2}]$$

hence

$$\bar{w} = pw_1 + qw_2$$

Continuous time model and computation of change of expected fitness

- Since Fisher's fundamental theorem states that the **change of average population fitness is equal to additive genetic variation** V_g , we proceed to compute the derivative of expected fitness \bar{w} with respect to time in a **continuous time model** – until now, we have only considered a **discrete time model**, where each new generation corresponds to $\Delta t = 1$.
- When discussing Fisher's theorem (and only in this instance), allele frequencies p, q , relative fitnesses $w_{1,1}, w_{1,2}, w_{2,2}$, expected population fitness \bar{w} , etc. are all functions of **continuous** time, so that one should write $p(t), q(t), \bar{w}(t)$. Population size N is also a function of time, so $N(t)$ should be written, where on average the population size remains constant. Such is the case for the number $k(t)$ of A_1 alleles at time t , etc. However, to avoid cumbersome notation, the dependence on continuous time variable t will not be written.
- In the case of **discrete time**, if fitness $w_1(t)$ is a measure of the amount to which A_1 alleles contribute to the population of the next generation $N(t + 1)$, then

$$k(t + 1) = k(t) \cdot w_1$$

and similarly

$$N(t + 1) = N(t) \cdot \bar{w}$$

- The previous statements are clear if fitness w_1 of allele A_1 and expected fitness of the population are understood to be the average number of offspring produced – thus w_1 is understood to be the average number of next generation A_1 alleles produced by a single A_1 allele, and \bar{w} is understood to be the average number of offspring individuals produced per current individual (e.g. according to 2016 statistics, the average birth rate is 1.8 in the US and 1.96 in France).

- By analogy, it follows that in the case of **continuous time**

$$\frac{dk(t)}{dt} = kw_1$$
$$\frac{dN(t)}{dt} = N\bar{w}$$

- Our goal is to compute $\frac{d\bar{w}}{dt}$. Since $\bar{w} = p^2w_{1,1} + 2pqw_{1,2} + q^2w_{2,2}$, we will need to compute the derivative of allele frequencies $\frac{dp}{dt}$ and $\frac{dq}{dt}$, which is why we computed $\frac{dk(t)}{dt}$ and $\frac{dN(t)}{dt}$ – recall that $p = k/N$.

- Compute the derivative of A_1 and A_2 allele frequencies as follows.

$$\begin{aligned}\frac{dp}{dt} &= \frac{d}{dt} \left(\frac{k}{N} \right) = \frac{1}{N} \frac{dk}{dt} - \frac{k}{N^2} \cdot \frac{dN}{dt} \quad (\text{since } p = k/N) \\ &= w_1 \cdot \frac{k}{N} - \frac{k}{N^2} \cdot \bar{w}N = w_1 \cdot \frac{k}{N} - \frac{k}{N} \cdot \bar{w} = pw_1 - p\bar{w} = p(w_1 - \bar{w})\end{aligned}$$

$$\frac{dq}{dt} = q(w_2 - \bar{w}) \quad (\text{by an analogous proof})$$

- Recall that

$$\bar{w} = p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2}$$

hence

$$\begin{aligned}\frac{d\bar{w}}{dt} &= (2pw_{1,1} + 2qw_{1,2}) \frac{dp}{dt} + (2qw_{2,2} + 2pw_{1,2}) \frac{dq}{dt} \quad (\text{product rule of differentiation}) \\ &= 2w_1 \cdot \frac{dp}{dt} + 2w_2 \cdot \frac{dq}{dt} \\ &= 2w_1 \cdot p(w_1 - \bar{w}) + 2w_2 \cdot q(w_2 - \bar{w})\end{aligned}$$

We need to prove the following non-obvious claim.

■ **Claim:**

$$p(w_1 - \bar{w})^2 + q(w_2 - \bar{w})^2 = pw_1(w_1 - \bar{w}) + qw_2(w_2 - \bar{w})$$

Proof of Claim:

$$\begin{aligned} p(w_1 - \bar{w})^2 + q(w_2 - \bar{w})^2 &= p(w_1 - \bar{w})(w_1 - \bar{w}) + q(w_2 - \bar{w})(w_2 - \bar{w}) \\ &= pw_1(w_1 - \bar{w}) - p\bar{w}(w_1 - \bar{w}) + qw_2(w_2 - \bar{w}) - q\bar{w}(w_2 - \bar{w}) \\ &= [pw_1(w_1 - \bar{w}) + qw_2(w_2 - \bar{w})] - \bar{w} [p(w_1 - \bar{w}) + q(w_2 - \bar{w})] \\ &= [pw_1(w_1 - \bar{w}) + qw_2(w_2 - \bar{w})] - \bar{w} [pw_1 + qw_2] + \bar{w}^2 [p + q] \\ &= [pw_1(w_1 - \bar{w}) + qw_2(w_2 - \bar{w})] - \bar{w} [\bar{w}] + \bar{w}^2 \\ &= pw_1(w_1 - \bar{w}) + qw_2(w_2 - \bar{w}) \end{aligned}$$

- We're almost done, but now need to note that from the table on page 21 (second page of section on Fisher's theorem),

$$w_{1,1} = \bar{w} + a_{1,1}$$

$$w_{1,2} = \bar{w} + a_{1,2}$$

$$pw_{1,1} = p\bar{w} + pa_{1,1}$$

$$qw_{1,2} = q\bar{w} + qa_{1,2}$$

- Add the last 2 equations together to get

$$\begin{aligned} pw_{1,1} + qw_{1,2} &= (p+q)\bar{w} + pa_{1,1} + qa_{1,2} \\ pw_{1,1} + qw_{1,2} - \bar{w} &= pa_{1,1} + qa_{1,2} \end{aligned}$$

- Similarly

$$\begin{aligned} w_{1,2} &= \bar{w} + a_{1,2} \\ w_{2,2} &= \bar{w} + a_{2,2} \\ pw_{1,2} &= p\bar{w} + pa_{1,2} \\ qw_{2,2} &= q\bar{w} + qa_{2,2} \end{aligned}$$

and by adding the last 2 equations together

$$\begin{aligned} pw_{1,2} + qw_{2,2} &= (p+q)\bar{w} + pa_{1,2} + qa_{2,2} \\ pw_{1,2} + qw_{2,2} - \bar{w} &= pa_{1,2} + qa_{2,2} \end{aligned}$$

Now

$$\begin{aligned} w_1 - \bar{w} &= pw_{1,1} + qw_{1,2} - \bar{w} \\ &= pa_{1,1} + qa_{1,2} \\ &= \alpha \end{aligned}$$

Similarly,

$$\begin{aligned} w_2 - \bar{w} &= qw_{2,2} + pw_{1,2} - \bar{w} \\ &= qa_{2,2} + pa_{1,2} \\ &= \beta \end{aligned}$$

End of proof of Fisher's Theorem

- It now follows that

$$\begin{aligned}\frac{d\bar{w}}{dt} &= 2p(w_1 - \bar{w})^2 + 2q(w_2 - \bar{w})^2 \\ &= 2p\alpha^2 + 2q\beta^2 \\ &= 2(p\alpha^2 + q\beta^2)\end{aligned}$$

- In the multi-allelic case

$$\frac{d\bar{w}}{dt} = 2 \sum_{i=1}^n p_i \alpha_i^2$$

- Since we've shown that

$$V_g = 2 \sum_{i=1}^n p_i \alpha_i^2$$

we have now established Fisher's Fundamental Theorem of population genetics:

$$\frac{d\bar{w}}{dt} = V_g$$

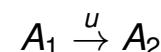
- Whew!

Mutation-selection balance

- Our goal in this section is to prove the mutation-selection balance equation

$$q^* \approx \frac{u}{hs}$$

where q^* denotes the A_2 equilibrium frequency, in which mutation and selection balance each other. Here, u denotes the rate of mutation of allele A_1 to A_2



where A_2 is assumed not to back-mutate to A_1 or to mutate to yet another allele A_3 . This equation is only (approximately) valid when $q = \text{Prob}[A_2]$ is small but non-zero, hence under the assumption that s is small, but non-zero (i.e. selection against A_1 is small).

- Consider the change in A_1 allele frequency due solely to mutation, where mutation rate $u \approx 10^{-9}$ in the case of nucleotide mutation rate (given DNA error correction) and $u \approx 10^{-6}$ in the case of mutation to a locus in Drosophila leading to a visual phenotype change.

$$\begin{aligned} p' &= p(1 - u) \\ \Delta_u p &= p(1 - u) - p = -up = -u(1 - q) = -u + qu \approx -u \end{aligned}$$

when q is small (since product of q and u is tiny).

- Now consider the change in A_1 allele frequency due solely to selection, given by

$$p' = \frac{p^2 + pq(1 - hs)}{\bar{w}} = \frac{p^2 + pq(1 - hs)}{1 - 2pqhs - q^2s}$$

$$\Delta_s p = p' - p = \frac{pq s}{\bar{w}} (ph + q(1 - h)) = \frac{pq s(ph + q(1 - h))}{1 - 2pqhs - q^2s} \approx qhs$$

under the assumption that q is small but non-zero, and assuming

$$\begin{aligned} q^2 &\approx 0 \\ qu &\approx 0 \\ p &\approx 1 \\ p^2 &\approx 1 \\ 2pqhs &\approx 0 \end{aligned}$$

The last assumption is justified, since the proportion of heterozygotes is $2pq \approx 2q$, hence small, provided that q is very small. Moreover, the selection factor s must be **non-negligible** (relatively large, such as $s \geq 0.1$). Indeed, if $s \approx u$, then there is insufficient selection of A_1 allele over A_2 , and so more and more A_1 alleles will be mutated to A_2 .

- Since

$$\Delta p = \Delta_u p + \Delta_s p \approx -u + qhs$$

at equilibrium, $\Delta p = 0$ implies that equilibrium A_2 frequency q^* satisfies the **mutation-selection balance equation**

$$\begin{aligned} u &\approx q^* hs \\ q^* &\approx \frac{u}{hs} \end{aligned}$$

Genetic load due to detrimental and lethal alleles

- Genetic load is the relative difference between maximal fitness and average fitness of a population

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}}$$

Under the hypothesis of mutation-selection balance, which assumes that $q = \text{Prob}[A_2]$ small but non-zero (see other assumptions on previous slide), then equilibrium allele frequencies p^* , q^* will satisfy the **mutation-selection balance** equation

$$\begin{aligned} q^* &\approx \frac{u}{hs} \\ p^* = 1 - q &\approx \frac{hs - u}{hs} \approx 1 \end{aligned}$$

so that equilibrium expected fitness is

$$\bar{w} = 1 - 2p^*q^*hs - q^*s \approx 1 - 2p^*q^*hs = 1 - 2p^* \cdot \frac{u}{hs} \cdot hs \approx 1 - 2u$$

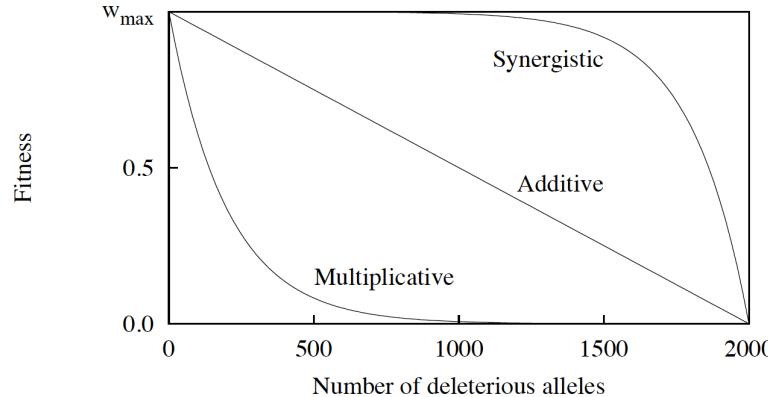
In the case of incomplete dominance ($0 < h < 1$) and underdominance ($h > 1$), maximal fitness $w_{\max} = 1$, it follows that the genetic load

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}} \approx \frac{1 - (1 - 2u)}{1} = 2u$$

under mutation-selection balance.

- Why doesn't the genetic load L depend on how detrimental a mutation is? Explanation: load is approximately the same if the population has many, only slightly deleterious mutations as when there are only a few lethal mutations. (Average population fitness is the proportion that survive, so individuals with a large number of somewhat deleterious alleles will be somewhat impaired in producing offspring, having the same net effect as if only a few had no offspring).

Total genetic load due to detrimental alleles at n distinct loci



- Under multiplicative epistasis, which models the deleterious contribution to fitness of n distinct loci as the product of the deleterious contribution of each, we have

$$\begin{aligned} \bar{W} &= \prod_{i=1}^n \bar{w}_i = \exp\left(\ln \prod_{i=1}^n \bar{w}_i\right) = \exp\left(\sum_{i=1}^n \ln(\bar{w}_i)\right) \\ \ln(\bar{w}_i) &\approx \ln(1 - 2u_i) = -2u_i + \frac{(-2u_i)^2}{2} + \frac{(-2u_i)^3}{3} + \frac{(-2u_i)^4}{4} + \dots \approx -2u_i \\ U &= 2\left(\sum_{i=1}^n u_i\right) \\ \bar{W} &\approx e^{-U} \end{aligned}$$

Note that the Taylor expansion and subsequent approximation of $\ln(1 - 2u_i)$ is explained on page 34 in the section "Some useful approximations" of the Probability Appendix. Total genetic load is the relative difference between maximal total fitness and average total fitness of a population

$$L_{\text{tot}} = \frac{W_{\max} - \bar{W}}{W_{\max}} = \frac{1 - \bar{W}}{1} \approx 1 - e^{-U} \quad (\text{see exponentially decreasing curve in figure})$$

- Gillespie argues that genetic **load** L only has relevance when the selection coefficient s is **not small**, i.e. $s > 0.001$. Why?
- The answer lies in the following argument. For simplicity, assume heterozygosity effect $h = \frac{1}{2}$ and that the fitness of the i th locus is

$$1 - X_i \cdot hs = 1 - \frac{X_i s}{2}$$

where $X_i \in \{0, 1, 2\}$ is the number of deleterious alleles at the i th locus. Assume that q is the probability of a deleterious allele at the i th locus, where q does not depend on i (probability of a deleterious mutation does not depend on location in genome).

- Under **multiplicative epistasis**, the fitness of the individual is Y , where

$$\begin{aligned} Y &= \prod_{i=1}^n (1 - X_i \cdot hs) = \exp \left(\ln \prod_{i=1}^n (1 - X_i hs) \right) = \exp \left(\sum_{i=1}^n \ln(1 - X_i hs) \right) \\ \ln(1 - X_i hs) &\approx -X_i hs + \frac{(-X_i hs)^2}{2} + \frac{(-X_i hs)^3}{3} + \dots \approx -X_i hs \\ Y &\approx \exp \left(-hs \sum_{i=1}^n X_i \right) \end{aligned}$$

Warning: Here Y is individual fitness under multiplicative epistasis (in contrast, Gillespie uses Y to represent $\sum_i X_i$ on page 73).

From the figure on the previous slide, it is clear that there is a precipitous drop in fitness with only a few deleterious mutations – since natural populations show robustness and vitality, multiplicative epistasis is a mathematically convenient but unreasonable model when discussing load.

- Assuming instead **additive epistasis**, the fitness of the individual is Y , where

$$Y = \sum_{i=1}^n (1 - X_i \cdot hs)$$

- Since the random variable $X_i \in \{0, 1, 2\}$ represents the number of A_1 alleles present at the i th locus for a pair of chromosomes, the number n of loci could be the number of nucleotides in protein coding regions

$$n \approx 20,000 \text{ genes} \times 300 \text{ nt} \approx 6 \cdot 10^5$$

which is too conservative an estimate, since we now know that noncoding RNAs (non protein-coding regions of the genome that are transcribed) play important regulatory roles. For that reason, a more reasonable estimate of n is the number of nucleotides in the haploid genome (haploid to avoid overcounting, since $X_i = 2$ means that both chromosomes contain the A_2 allele) – thus $n \approx 10^9$.

- Clearly, we have

$$\begin{aligned} \text{Prob}[X_i = 0] &= p^2 \\ \text{Prob}[X_i = 1] &= 2pq \\ \text{Prob}[X_i = 2] &= q^2 \end{aligned}$$

so we compute

$$\begin{aligned} E[X_i] &= p^2 \cdot 0 + 2pq \cdot 1 + q^2 \cdot 2 = 2q(p + q) = 2q \\ E[X_i^2] &= p^2 \cdot 0^2 + 2pq \cdot 1^2 + q^2 \cdot 2^2 = 2pq + 4q^2 = 2q(p + 2q) = 2q(p + q + q) = 2q(1 + q) \\ V[X_i] &= E[X_i^2] - E[X_i]^2 = 2q(1 + q) - (2q)^2 = 2q + 2q^2 - 4q^2 = 2q - 2q^2 = 2q(1 - q) = 2pq \\ V[1 - \frac{X_i s}{2}] &= V[1] + (-1)^2 V[\frac{X_i s}{2}] = 0 + \frac{s^2}{4} V[X_i] = \frac{2pq s^2}{4} = \frac{pq s^2}{2} \\ V[Y] &= V\left[\sum_{i=1}^n \left(1 - \frac{X_i s}{2}\right)\right] = \sum_{i=1}^n V\left[\left(1 - \frac{X_i s}{2}\right)\right] = \frac{npq s^2}{2} \end{aligned}$$

- For $n = 10^9$ and $s = 10^{-5}$, it follows that

$$V[Y] = \frac{npqs^2}{2} \leq 0.1 \frac{pq}{2} \leq 0.05q$$

Since minor allele frequency $q < \frac{1}{2}$ (in the literature, minor allele frequencies are not counted unless $q \geq 0.05$), it follows that $V[Y] \leq 0.025$, so the standard deviation $\sqrt{V[Y]} < 0.1582$. In the standard normal distribution, 68.26% (approximately $\frac{2}{3}$) of the area under the curve lies within 1 standard deviation of the mean, it follows that $68.26\% + \frac{100-68.26\%}{2} = 84.13\%$ of the population has an overall fitness of at least $E[Y] - 0.1582$. According to Gillespie, such numbers are acceptable.

- For this reason, genetic load is of little consequence when the selection factor s is tiny. However, when $s > 0.001$, then genetic load is an important factor in deleterious mutations.

Goals of Greenberg-Crow experiment

1. create inbred homozygotic $+_n/+_n$ flies
2. create outbred heterozygotic $+_n/+_{n+1}$ flies
3. compare fitness (viability) of inbred versus outbred flies

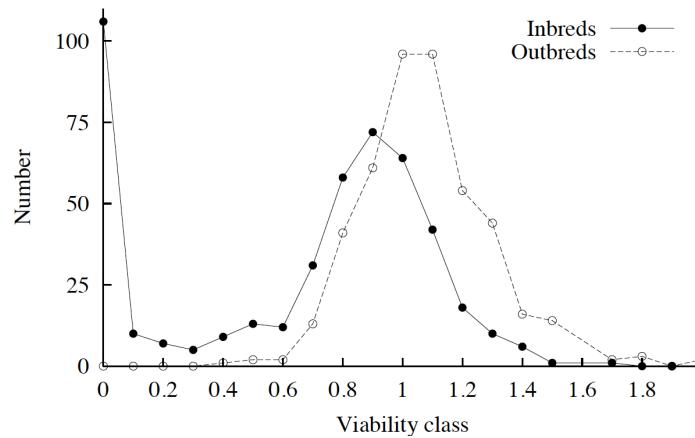


Figure 3.7: The numbers of lines in each viability class among the inbred and outbred flies in the Greenberg and Crow experiment.

Note that inbred fly viability histogram is bimodal, with large population of lethals.

- $A = 1.008$: average relative viability of outbred flies
- $B = 0.632$: average relative viability of inbred flies
- $C = 0.842$: average relative viability of inbred flies, without lethal mutations on chromosome 2

Question: Is it clear how to average viabilities above (relative viability will be explained shortly)?

Explanation of D. melanogaster crosses

Overview of chromosome 2 allele types

- $n = 1, 2, \dots, 465$ wild type males, whose chromosome 2 is $+_n/+'_n$
- Cy: curly wing dominant allele
- cn: cinnabar eye recessive allele
- bw: brown eye recessive allele
- bw^D : brown eye dominant allele
- males: no recombination
- females: pericentric inversion preventing recombination

Computation of viability and relative viability

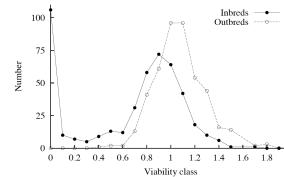


Figure 3.7: The numbers of lines in each viability class among the inbred and outbred flies in the Greenberg and Crow experiment.

1. compute relative frequency histograms from the absolute frequency histograms, which results in probability distributions for the viability of inbred and outbred flies
2. compute expected viability histograms for (a) inbred, (b) inbred non-lethal, (c) outbred flies, where expected viability is

$$E[\text{fitness}] = \sum_{i=0}^{20} \text{Prob}[\text{class}(i/10)] \cdot \text{fitness}[\text{class}(i/10)]$$

3. compute relative viability of inbred flies by

$$\frac{2N(+_n/+_n)}{N(Cy/+_n) + N(bw^D/+_n)}$$

using the number of controls $Cy/+_n$ and $bw^D/+_n$)

4. compute relative viability of outbred flies by

$$\frac{2N(+_n/+_{n+1})}{N(Cy/+_n) + N(bw^D/+_{n+1})}$$

using the number of controls $Cy/+_n$ and $bw^D/+_{n+1}$) – note presence of $+_{n+1}$ and not $+'_n$

Crosses used in Greenberg-Crow experiment

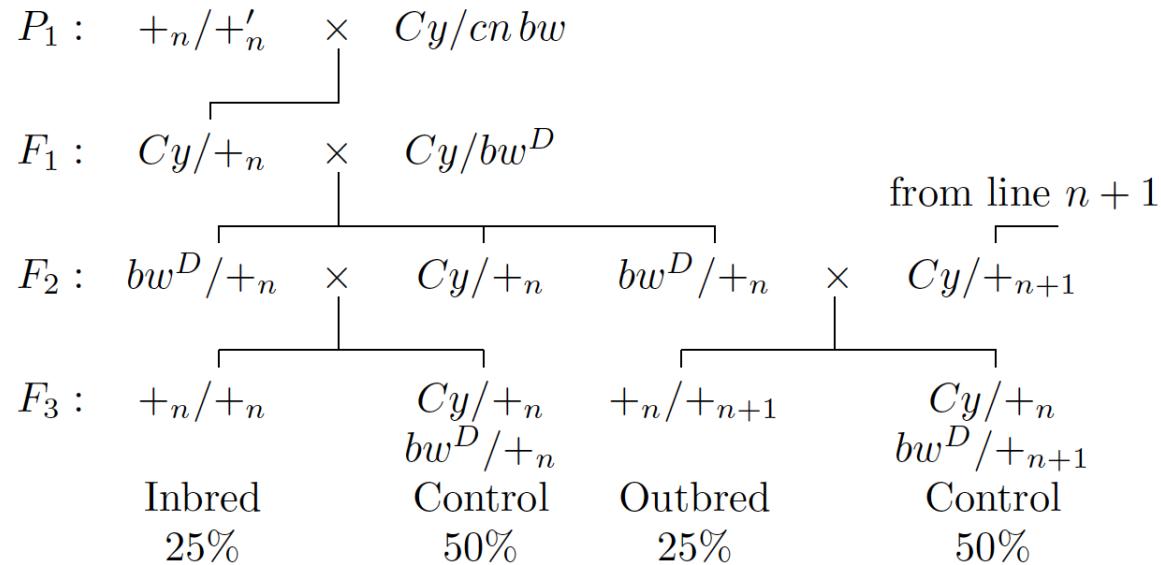


Figure 3.6: The *Drosophila melanogaster* crosses used to uncover hidden variation. In each cross, the male is on the left.

Drosophila melanogaster crosses used to uncover hidden variation. (Image from Gillespie Figure 3.6)

Experimental protocol

- P: parent generation

$$+_n/+'_n \quad (m) \times Cy/cn, bw \quad (f)$$

yielding offspring:

$$+_n/Cy \quad +_n/cn, bw \quad +'_n/Cy \quad +_n/cn, bw$$

From offspring, select $Cy/+_n(m)$. In the next round, $Cy/+_n(m)$ will be mated with outside $Cy/bw^D(f)$ – here we confound $+_n$ and $+'_n$.

- F_1 : generation

$$Cy/+_n \quad (m) \times Cy/bw^D \quad (f)$$

yielding offspring:

$$Cy/Cy \quad Cy/bw^D \quad +_n/Cy \quad +_n/bw^D$$

Homozygous curly wing is lethal. From offspring, select $+_n/bw^D \quad (m)$ and $Cy/+_n \quad (f)$ for brother-sister cross next round.

- F_2 : brother-sister breeding of F_2 generation

$$+_n/bw^D \quad (m) \times Cy/+_n \quad (f)$$

yielding offspring:

$$+_n/Cy \quad +_n/+_n \quad bw^D/Cy \quad bw^D/+_n$$

By Mendelian segregation, we have 25% inbred $+_n/+_n$ flies, and 25% control $Cy/+_n$ flies and 25% control $bw^D/+_n$ flies (disregard bw^D/Cy flies)

- F_2 : interbreed between n th and $(n + 1)$ st line

$$+_n/bw^D \quad (m) \times Cy/+_{n+1} \quad (f)$$

yielding offspring:

$$+_n/Cy \quad +_n/+_{n+1} \quad bw^D/Cy \quad bw^D/+_{n+1}$$

By Mendelian segregation, we have 25% outbred $+_n/+_{n+1}$ flies, and 25% control $Cy/+_n$ flies and 25% control $bw^D/+_{n+1}$ flies (disregard bw^D/Cy flies)

This produces inbred and outbred flies that allow a quantification of the loss of viability due to homozygous deleterious alleles in nature.

- recessive lethal genes found on $\frac{106}{465}$ of chromosomes examined, yielding that 23% of 2nd chromosome carries at least one lethal mutation when in homozygous state

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|---------------------|----------|----------|----------|
| relative fitness | 1 | $1 - hs$ | $1 - s$ |
| inbred frequencies | p | 0 | q |
| outbred frequencies | p^2 | $2pq$ | q^2 |

- inbred expected fitness (mean inbred viability): $p(1) + q(1 - s) = p + q - qs = 1 - qs$
- outbred expected fitness (mean outbred viability): $p^2(1) + 2pq(1 - hs) + q^2(1 - s) = 1 - 2pqhs - q^2s$
- outbred fitness > inbred fitness > if and only if

$$\begin{aligned} 1 - 2pqhs - q^2s &> 1 - qs \Leftrightarrow qs > 2pqhs + q^2s \\ &\Leftrightarrow 1 > 2ph + q \Leftrightarrow 1 - q > 2ph \\ &\Leftrightarrow p > 2ph \Leftrightarrow 1 > 2h \\ &\Leftrightarrow h < \frac{1}{2} \end{aligned}$$

- This makes sense, since when $h < \frac{1}{2}$, heterozygote relative fitness $1 - hs$ is closer to A_1A_1 homozygote relative fitness of 1 than to A_2A_2 homozygote fitness of $1 - s$.

Conclusions from Greenberg-Crow experiment

- inbreds less viable than outbreds
- Inbreeding depression is the reduced fitness in a given population as a result of inbreeding, which means that average fitness of a homozygous population is smaller than that of a heterozygous population:
inbreeding depression $\Leftrightarrow h < \frac{1}{2}$
- $\approx 20\%$ of genome contains mutations that are lethal in homozygous state
- inverse heterozygous-homozygous effect, described in detail on next slide: hs approximately constant, with an estimate of

$$hs = \frac{u}{q^*} \approx \frac{3.86 \cdot 10^{-6}}{0.0025} = 0.001544 \approx 1.5 \cdot 10^{-3}$$

This estimate was obtained by using the mutation-selection balance equation

$$q^* \approx \frac{u}{hs}$$

together with the experimentally determined (deleterious) **null allele** frequency $q = 0.0025$ in Drosophila (null allele means no measurable activity of enzyme), and with mutation rate of $u = 3.86 \cdot 10^{-6}$ to null alleles.

The inverse heterozygous-homozygous effect implies:

- deleterious mutations with **large (severe) effect** ($0 \ll s \leq 1$) are almost **recessive** ($0 \approx h$)
- deleterious mutations with **small (mild) effect** ($0 \approx s \leq 1$) are **widely found** in heterozygotes ($0 \ll h \approx \frac{1}{2}$)

Argument to justify inverse heterozygous-homozygous effect

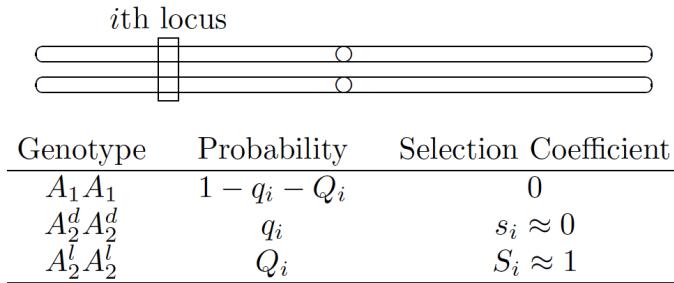


Figure 3.8: The possible states of a typical locus in an inbred fly. A superscript *d* on an allele indicates that it is a deleterious mutant. A superscript *l* identifies a lethal mutant.

Figure 3.8 of Gillespie, shows locus in inbred fly, where ‘d’ means deleterious and ‘l’ means lethal.

- detimental load $D = \sum q_i s_i$, where q_i is probability that inbred fly is homozygous for deleterious allele at *i*th locus (detimental load is distinct from genetic load)
- lethal load $L = \sum Q_i S_i$, where Q_i is probability that inbred fly is homozygous for lethal allele at *i*th locus (lethal load is distinct from genetic load)

- Recall that
 - ▷ $A = 1.008$: probability of survival of outbred fly to adulthood
 - ▷ $B = 0.632$: probability of survival of inbred fly to adulthood
 - ▷ $C = 0.842$: probability of survival of inbred fly to adulthood, discounting flies that suffered death from lethal mutation in homozygous state
- assume loci act independently (multiplicative epistasis):

$$\begin{aligned}
 B &= A \prod_{i=1}^n (1 - q_i s_i)(1 - Q_i S_i) \\
 &= A \exp(\ln(\prod_{i=1}^n (1 - q_i s_i)(1 - Q_i S_i))) \\
 &= A \exp\left(\sum_{i=1}^n \ln(1 - q_i s_i) + \sum_{i=1}^n \ln(1 - Q_i S_i)\right) \\
 &\approx A \exp\left(-\sum_{i=1}^n q_i s_i - \sum_{i=1}^n Q_i S_i\right) \quad (\text{using first-order approximation of } \ln(1 + x) \text{ for small } x) \\
 &= A \exp(-D - L)
 \end{aligned}$$



$$D + L = \ln(A) - \ln(B) = \ln(1.008) - \ln(0.632) = 0.4668$$

$$\begin{aligned} C &= A \exp(\ln(\prod_{i=1}^n (1 - q_i s_i))) \\ &= A \exp\left(\sum_{i=1}^n \ln(1 - q_i s_i)\right) = A \exp\left(\sum_{i=1}^n -q_i s_i\right) = Ae^{-D} \end{aligned}$$

$$D = \ln(A) - \ln(C) = \ln(1.008) - \ln(0.842) = 0.1799$$

$$L = 0.4668 - 0.1799 = 0.2868$$

$$\frac{D}{L} = \frac{\sum q_i s_i}{\sum Q_i S_i} = 0.627267 \approx 0.627$$

- However, by mutation-selection balance equation

$$q^* \approx \frac{u}{hs}$$

this implies that

$$\begin{aligned} \frac{D}{L} &= \frac{\sum u_i/h_i}{\sum U_i/H_i} \\ &\approx \frac{\sum u_i}{\sum U_i} \quad (\text{assuming constant heterozygous effects}) \end{aligned}$$

However, the ratio of deleterious to lethal mutation rate (independently experimentally determined) is $\approx 10 \gg 0.617$ according to current research, contradicting the theoretical derivation which assumed multiplicative epistasis.

- Greenberg and Crow argue that the unreasonably low theoretical ratio $\frac{D}{L} = 0.617$ compared with the experimental ratio of $\frac{D}{L} > 10$ suggests that one can not assume that the heterozygous effect is constant – i.e. one can not assume $h_i \approx h_j \approx H_i \approx H_j$ all i, j .
- Using the mutation-selection balance equation $q^* \approx \frac{u}{hs}$, we have

$$\begin{aligned}\frac{D}{L} &= \frac{\sum q_i s_i}{\sum Q_i S_i} = 0.627 \\ &= \frac{\sum (u_i/h_i s_i) h_i}{\sum (U_i/H_i S_i) H_i}\end{aligned}$$

If we **assume** that the product of selection and heterozygosity factors is approximately constant, i.e. if $h_i s_i \approx h_j s_j \approx H_i S_i \approx H_j S_j$, then

$$\frac{D}{L} = \frac{\sum (u_i/h_i s_i) h_i}{\sum (U_i/H_i S_i) H_i} \approx \frac{\sum u_i h_i}{\sum U_i H_i}$$

and using values determined from a different experiment, the last term is approximately 0.711 – a good fit.

Wright's enzyme model of synergistic epistasis

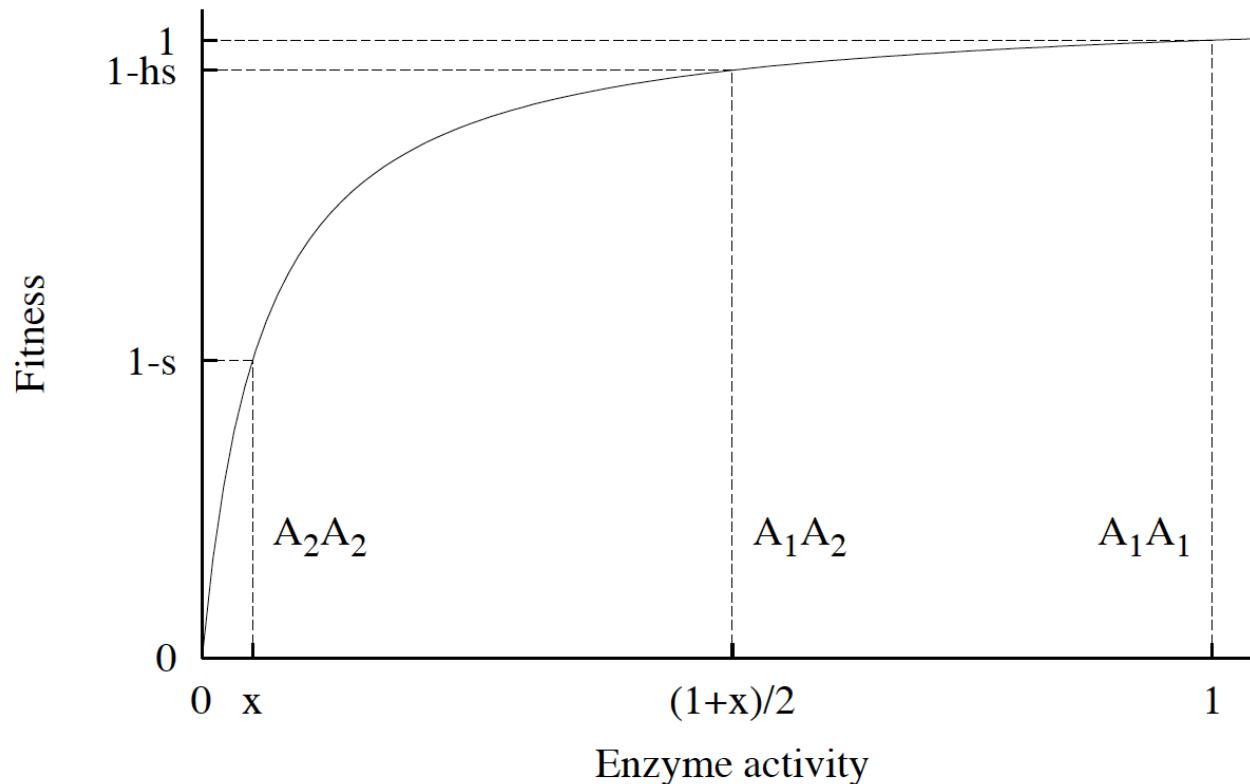
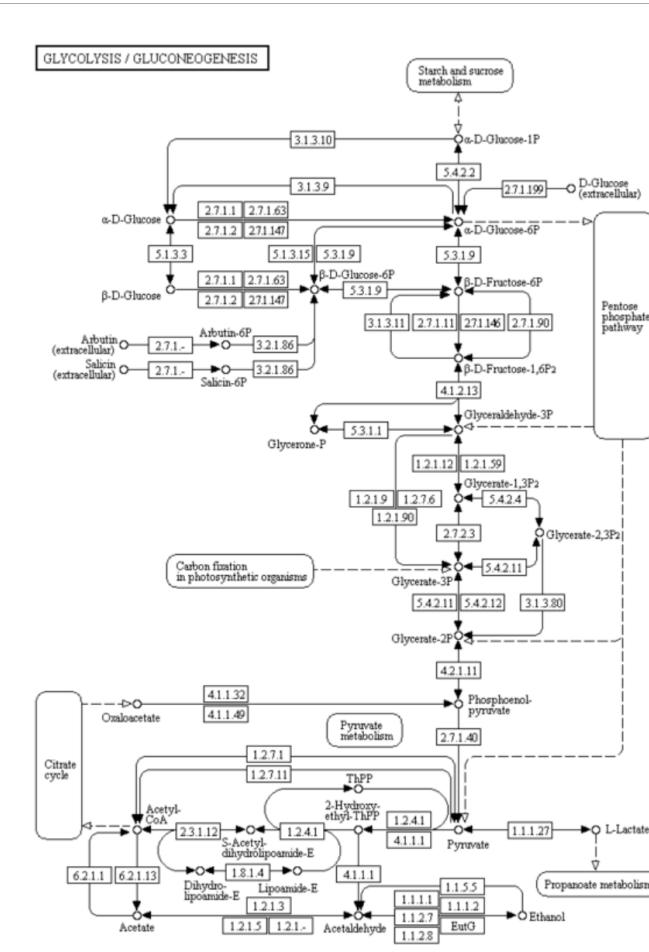


Figure 3.9: Sewall Wright's model of dominance applied to viability.

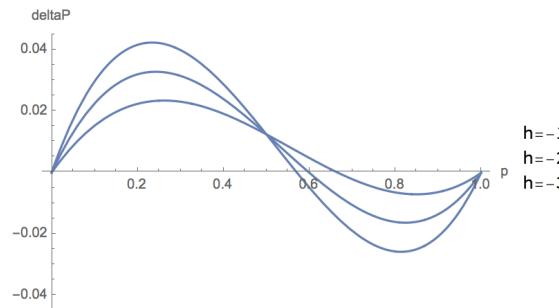
Robustness of biological pathways due to redundancy



Glycolysis pathway taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) web site <https://www.genome.jp/kegg/>

Changing environments

- Both mutation-selection balance and balancing selection (caused by overdominance) are the only mechanisms we've learned so far, that are capable of ensuring stable major and minor allele frequencies $0 < q < p < 1$
- It is believed that overdominance is rare, arising only in specific conditions, such as the presence of sickle cell trait in malaria-infested African regions
- Question: Why are major and minor allele frequencies $0 < q < p < 1$ apparently stable and not cases of incomplete dominance, with an ultimate fixation of the major allele?
- Answer: Environments change over both space and time, and humans don't mate randomly. Consider geographic restriction of potential mates.
- Balancing selection requires allele A_1 to increase if $A_1 \approx 0$, as well as for allele A_1 to **decrease** if $A_1 \approx 1$. Indeed, in overdominance, depicted in the figure below, A_1 frequency increases if $0 < p < \frac{h-1}{2h-1}$, while A_1 frequency decreases if $\frac{h-1}{2h-1} < p < 1$.



| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|---------------------|-----------------|-----------------|-----------------|
| fitness at time t | $w_{1,1}^{(t)}$ | $w_{1,2}^{(t)}$ | $w_{2,2}^{(t)}$ |

- Slowly changing, **fluctuating** environments can be modeled by **fluctuating** fitness values at time t : $w_{1,1}^{(t)}$, $w_{1,2}^{(t)}$, $w_{2,2}^{(t)}$ as shown in the table above.
- Assuming p is very close to 0, and q is very close to 1, the A_1 allele frequency p' in the next generation satisfies:

$$\begin{aligned} p' &= \frac{p^2 w_{1,1} + pq w_{1,2}}{p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2}} \approx \frac{pw_{1,2}}{2pw_{1,2} + w_{2,2}} \quad (\text{assuming } p \approx 0, q \approx 1) \\ &\approx \frac{pw_{1,2}}{w_{2,2}} \quad (\text{assuming } p \approx 0, q \approx 1) \end{aligned}$$

$$p'' \approx p' \frac{w'_{1,2}}{w'_{2,2}} \approx p \frac{w_{1,2}}{w_{2,2}} \cdot \frac{w'_{1,2}}{w'_{2,2}}$$

$$p^{(t)} \approx p \frac{\prod_{k=0}^{t-1} w_{1,2}^{(k)}}{\prod_{k=0}^{t-1} w_{2,2}^{(k)}}$$

$$w_{i,j}^g = \left(\prod_{k=0}^{t-1} w_{i,j}^{(k)} \right)^{1/t} \quad (\text{geometric mean, which is } \leq \text{ arithmetic mean - see Prob Appendix})$$

$$p^{(t)} \approx p \left(\frac{w_{1,2}^g}{w_{2,2}^g} \right)^t$$

- It follows that if p is close to 0, and q is close to 1, the A_1 allele frequency $p^{(t)}$ in the t th generation **increases** if and only if the **geometric mean of heterozygote A_1A_2 fitness is greater than the geometric mean of homozygote A_2A_2 fitness**:

$$p^{(t)} > p \Leftrightarrow (w_{1,2}^g)^t > (w_{2,2}^g)^t \Leftrightarrow w_{1,2}^g > w_{2,2}^g \Leftrightarrow \left(\prod_{k=0}^{t-1} w_{1,2}^{(k)} \right)^{1/t} > \left(\prod_{k=0}^{t-1} w_{2,2}^{(k)} \right)^{1/t}$$

- Similarly, if p is very close to 1, and q is very close to 0, the A_2 allele frequency $q^{(t)}$ in the t th generation satisfies:

$$q^{(t)} > q \Leftrightarrow (w_{1,2}^g)^t > (w_{1,1}^g)^t \Leftrightarrow w_{1,2}^g > w_{1,1}^g \Leftrightarrow \left(\prod_{k=0}^{t-1} w_{1,2}^{(k)} \right)^{1/t} > \left(\prod_{k=0}^{t-1} w_{1,1}^{(k)} \right)^{1/t}$$

Since

$$q^{(t)} > q \Leftrightarrow -q^{(t)} < -q \Leftrightarrow 1 - q^{(t)} < 1 - q \Leftrightarrow p^{(t)} < p$$

it follows that if p is close to 1, and q is close to 0, then the A_1 allele frequency $p^{(t)}$ in the t th generation **decreases** if and only if the **geometric mean of heterozygote A_1A_2 fitness is greater than the geometric mean of homozygote A_1A_1 fitness**.

- In summary, **balancing selection is operative** if geometric fitness of heterozygotes is greater than geometric fitness of either homozygote.
- However, there is **no stable equilibrium** allele frequency p^* , since fitness parameters $w_{1,1}^{(t)}, w_{1,2}^{(t)}, w_{2,2}^{(t)}$ are not static, but rather **change over time** due to changing environments. This is different than the case of overdominance, where there is a stable equilibrium $p^* = \frac{h-1}{2h-1}$, which we derived using the hypothesis of unchanging fitness parameters.
- The proof, that balancing selection can occur when the geometric mean of heterozygote fitness exceeds the geometric mean of homozygote fitness, depends **heavily** on approximations that are only valid when p is close to 0 or close to 1.

Variance of time-dependent fitness

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|----------|-----------------|---------------------------------------|-----------------|
| fitness | $1 + Y_1^{(t)}$ | $1 + \frac{Y_1^{(t)} + Y_2^{(t)}}{2}$ | $1 + Y_2^{(t)}$ |

- The goal of this section is to show that **balancing selection** can occur in the case of **incomplete dominance** when the **variance** of the (stochastic) time-dependent fitness is **greater than the difference of mean fitness** of the A_1 and A_2 alleles.
- Comparing the notation used in Fisher's Fundamental Theorem (pages 20–30) with the notation below, we see that Fisher's deviations from expected fitness values $a_{1,1}$, as well as 2α , correspond to $Y_1^{(t)}$ below, that $a_{1,2}$, as well as $\alpha + \beta$ correspond to $\frac{Y_1^{(t)} + Y_2^{(t)}}{2}$, that $a_{2,2}$ and 2β correspond to $Y_2^{(t)}$ below. Bear in mind that $Y_1/2$ denotes the **deviation** from mean fitness of allele A_1 and $Y_2/2$ denotes the **deviation** from mean fitness of allele A_2 , where we have **time-dependent, stochastic** values $Y_1^{(t)}$ and $Y_2^{(t)}$ at time t . Comparing the table above with the table on page 21 in the proof of Fisher's Fundamental Theorem, note that **expected fitness** $\bar{w} = 1$. We could have handled arbitrary \bar{w} with small additional technical work. The reason for taking $\bar{w} = 1$ is to use a truncated Taylor expansion to approximate expressions such as $\ln(1 + Y_1^{(t)})$.
- Assume that fitnesses are additive (i.e. $h = 1/2$), with stochastic fluctuations over time, and that the expectation and variance of the **deviation** from expected fitness are given by

$$\mu_1 = E[Y_1^{(t)}]$$

$$\mu_2 = E[Y_2^{(t)}]$$

$$\sigma^2 = V[Y_1^{(t)}] = V[Y_2^{(t)}] \quad (\text{variance } \sigma^2 \text{ is assumed to be } \underline{\text{independent of } t})$$

■ Then

$$E\left[\frac{Y_1^{(t)} + Y_2^{(t)}}{2}\right] = \frac{E[Y_1^{(t)}] + E[Y_2^{(t)}]}{2} = \frac{\mu_1 + \mu_2}{2}$$

$$V[1 + Y_1^{(t)}] = V[Y_1^{(t)}] = \sigma^2$$

$$V[1 + Y_2^{(t)}] = V[Y_2^{(t)}] = \sigma^2$$

$$V\left[1 + \frac{Y_1^{(t)} + Y_2^{(t)}}{2}\right] = V\left[\frac{Y_1^{(t)} + Y_2^{(t)}}{2}\right]$$

$$V\left[\frac{Y_1^{(t)} + Y_2^{(t)}}{2}\right] = \frac{1}{4} \cdot (V[Y_1^{(t)}] + V[Y_2^{(t)}]) = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2}$$

- Compute an approximation of the homozygotic **geometric mean fitness** $w_{1,1}^g$:

$$\begin{aligned} w_{1,1}^g &= \left(\prod_{k=0}^{t-1} 1 + Y_1^{(k)} \right)^{1/t} = \exp \ln \left(\prod_{k=0}^{t-1} 1 + Y_1^{(k)} \right)^{1/t} \\ &= \exp \left(\frac{1}{t} \ln \left(\prod_{k=0}^{t-1} 1 + Y_1^{(k)} \right) \right) = \exp \left(\frac{1}{t} \sum_{k=0}^{t-1} \ln(1 + Y_1^{(k)}) \right) \end{aligned}$$

On page 34 of the Probability Appendix, you find a proof that

$$\ln(1 + x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

hence by taking the second-order approximation (dropping terms of order 3 and higher), we approximate $\ln(1 + Y_1^{(k)})$ by $Y_1^{(k)} - \frac{(Y_1^{(k)})^2}{2}$. **Question:** Why doesn't it suffice to take only a first-order approximation?

- This yields

$$w_{1,1}^g \approx \exp \left(\frac{1}{t} \cdot \sum_{k=0}^{t-1} \left(Y_1^{(k)} - \frac{\{Y_1^{(k)}\}^2}{2} \right) \right) = \exp \left(\frac{\sum_{k=0}^{t-1} Y_1^{(k)}}{t} - \frac{\sum_{k=0}^{t-1} \{Y_1^{(k)}\}^2}{2t} \right)$$

$$w_{1,1}^g \approx e^{\mu_1 - \sigma^2/2} \quad (\text{if } t \text{ large, and both } \mu_1 = E[Y_1^{(t)}] \text{ and } \sigma^2 = V[Y_1^{(t)}] \text{ are close to 0})$$

$$w_{1,1}^g \approx 1 + \mu_1 - \frac{\sigma^2}{2} \quad (\text{recall } e^x = 1 + x + x^2/2! + x^3/3! + \dots \approx 1 + x \text{ for small positive } x)$$

- By replacing Y_1 by Y_2 in the previous proof, we also have

$$w_{2,2}^g \approx 1 + \mu_2 - \frac{\sigma^2}{2}$$

- Now compute an approximation of the heterozygotic geometric mean fitness.

$$\begin{aligned}
 w_{1,2}^g &= \left(\prod_{k=0}^{t-1} 1 + \frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right)^{1/t} = \exp \ln \left(\prod_{k=0}^{t-1} 1 + \frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right)^{1/t} \\
 &= \exp \left(\frac{1}{t} \ln \left(\prod_{k=0}^{t-1} 1 + \frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right) \right) = \exp \left(\frac{1}{t} \sum_{k=0}^{t-1} \ln \left(1 + \frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right) \right) \\
 &\approx \exp \left(\frac{1}{t} \cdot \sum_{k=0}^{t-1} \left(\frac{Y_1^{(k)} + Y_2^{(k)}}{2} - \frac{1}{2} \cdot \left\{ \frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right\}^2 \right) \right) \\
 &= \exp \left(\frac{\sum_{k=0}^{t-1} (Y_1^{(k)} + Y_2^{(k)})}{2t} - \frac{\sum_{k=0}^{t-1} \{Y_1^{(k)} + Y_2^{(k)}\}^2}{8t} \right)
 \end{aligned}$$

$$Z^{(k)} = \frac{Y_1^{(k)} + Y_2^{(k)}}{2} \quad (\text{for clarity, define new random variable})$$

$$E[Z^{(k)}] = E \left[\frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right] = \frac{\mu_1 + \mu_2}{2}$$

$$V[Z^{(t)}] = V \left[\frac{Y_1^{(k)} + Y_2^{(k)}}{2} \right] = \frac{\sigma^2 + \sigma^2}{4} = \frac{\sigma^2}{2}$$

$$w_{1,2}^g \approx \exp \left(\frac{\sum_{k=0}^{t-1} Z^{(k)}}{t} - \frac{\sum_{k=0}^{t-1} \{Z^{(k)}\}^2}{2t} \right)$$

$$\approx \exp \left(\frac{\mu_1 + \mu_2}{2} - \frac{\sigma^2}{4} \right) \quad (\text{if } t \text{ large, and values } \mu_1 = E[2Y_1^{(k)}], \mu_2 = E[2Y_2^{(k)}], \frac{\sigma^2}{2} = V[Z^{(k)}] \text{ close to 0})$$

$$w_{1,2}^g \approx 1 + \frac{\mu_1 + \mu_2}{2} - \frac{\sigma^2}{4} \quad (\text{recall } e^x = 1 + x + x^2/2! + x^3/3! + \dots \approx 1 + x \text{ for small positive } x)$$

- It follows that

$$\begin{aligned}
 w_{1,1}^g < w_{1,2}^g &\Leftrightarrow \left(1 + \mu_1 - \frac{\sigma^2}{2}\right) < \left(1 + \frac{\mu_1 + \mu_2}{2} - \frac{\sigma^2}{4}\right) \\
 &\Leftrightarrow \left(\frac{\mu_1}{2} - \frac{\sigma^2}{4} < \frac{\mu_2}{2}\right) \Leftrightarrow \left(\frac{\mu_1}{2} - \frac{\mu_2}{2} < \frac{\sigma^2}{4}\right) \\
 &\Leftrightarrow \mu_1 - \mu_2 < \frac{\sigma^2}{2} \\
 w_{2,2}^g < w_{1,2}^g &\Leftrightarrow \left(1 + \mu_2 - \frac{\sigma^2}{2} < 1 + \frac{\mu_1 + \mu_2}{2} - \frac{\sigma^2}{4}\right) \\
 &\Leftrightarrow \left(\mu_2/2 - \frac{\sigma^2}{4} < \mu_1/2\right) \Leftrightarrow \left(\mu_2/2 - \mu_1/2 < \frac{\sigma^2}{4}\right) \\
 &\Leftrightarrow \left(\mu_2 - \mu_1 < \frac{\sigma^2}{2}\right)
 \end{aligned}$$

- Important result to retain (**take-home message**):

$$\max(w_{1,1}^g, w_{2,2}^g) < w_{1,2}^g \Leftrightarrow |\mu_1 - \mu_2| < \frac{\sigma^2}{2}$$

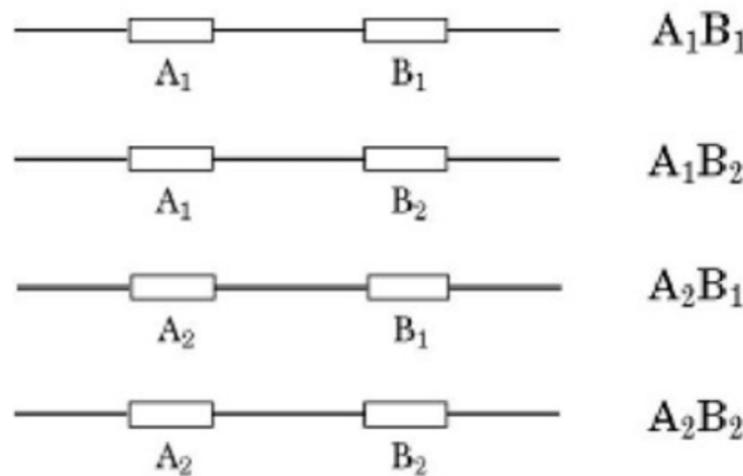
- Conclusion:** heterozygote geometric mean fitness is greater than both homozygote geometric fitnesses if and only if **heterozygotic variance** is greater than the **absolute difference** of homozygotic mean fitnesses over time.
- Notice the importance of **variance** in this result. In a similar fashion, Fisher's Fundamental Theorem indicated the importance of **variance**.



Chapter 4: Two Locus Dynamics

Linkage disequilibrium

Locus A Locus B Haplotype



| haplotype | haplotype frequency | haplotype frequency at equilibrium |
|-------------------------------|-----------------------------|--|
| A ₁ B ₁ | $x_1 = \text{Prob}[A_1B_1]$ | $p_1p_2 = \text{Prob}[A_1] \cdot \text{Prob}[B_1]$ |
| A ₁ B ₂ | $x_2 = \text{Prob}[A_1B_2]$ | $p_1q_2 = \text{Prob}[A_1] \cdot \text{Prob}[B_2]$ |
| A ₂ B ₁ | $x_3 = \text{Prob}[A_2B_1]$ | $q_1p_2 = \text{Prob}[A_2] \cdot \text{Prob}[B_1]$ |
| A ₂ B ₂ | $x_4 = \text{Prob}[A_2B_2]$ | $q_1q_2 = \text{Prob}[A_2] \cdot \text{Prob}[B_2]$ |

- If D denotes linkage disequilibrium, then we will show

$$D = x_1 - p_1 p_2 \quad (1)$$

$$-D = x_2 - p_1 q_2 \quad (2)$$

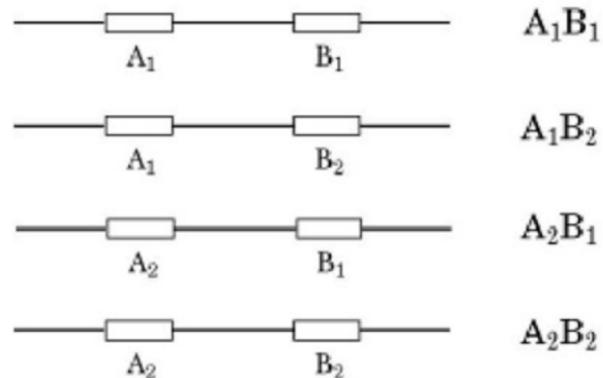
$$-D = x_3 - q_1 p_2 \quad (3)$$

$$D = x_4 - q_1 q_2 \quad (4)$$

- Moreover, we will show that

$$D = x_1 x_4 - x_2 x_3$$

Locus A Locus B Haplotype



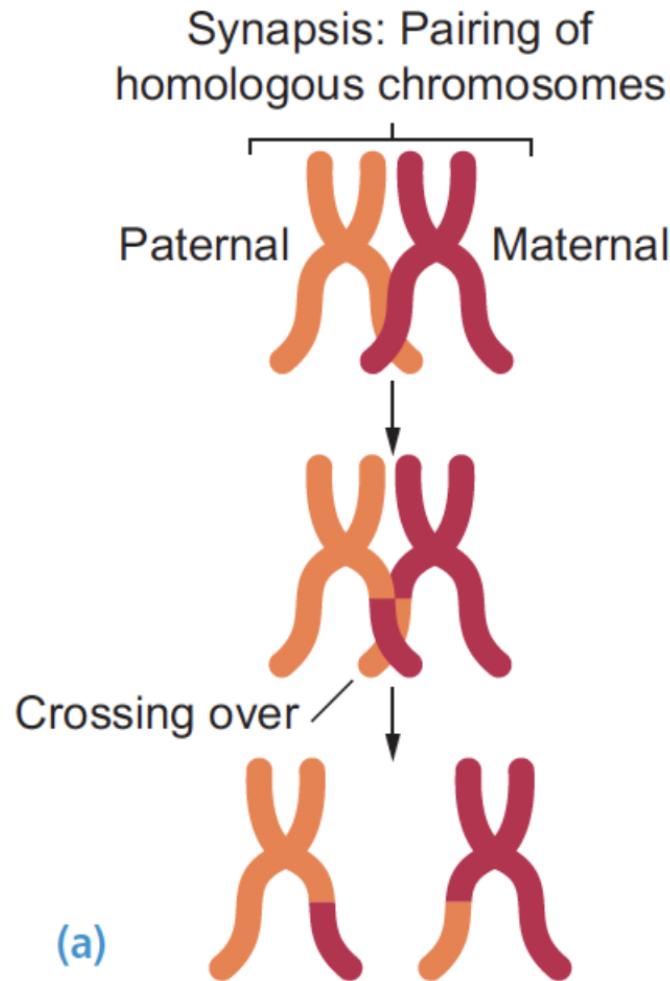
| | |
|----|----|
| x1 | x2 |
| x3 | x4 |

determinant is $(x_1 x_4 - x_3 x_4)$

Crossover in meiosis

One centimorgan (abbreviated cM) or genetic mapping unit is a measure of genetic linkage, defined as the distance between chromosome positions (loci or markers) for which the expected number crossover breakpoints in a single generation is 0.01. Not a true physical distance.

One centimorgan $\approx 10^6$ bps in humans: specifically 1.09×10^6 bps in men and 0.64×10^6 bps in women. In the malarial parasite *Plasmodium falciparum*, 1cM ≈ 15 kbps, i.e. within a given 15,000 nt region, one expects to find one parasite in every 100 parasites having a crossover event in this region (in one generation).



Mathematical model of crossover

Imagine a long straight line (chromosome), where crossover breakpoints are uniformly randomly chosen. Divide the line into equal length segments, and perform two computations:

- 1) Count the number of crossover breakpoints per segment
- 2) Compute the distance between successive crossover points

There are known probability distributions that describe both:

- 1) $P[X=k] = \exp(-\mu) \cdot \mu^k/k!$ (Poisson distribution, discrete)
- 2) $P[X \leq n] = 1 - \exp(-n/\mu)$ (exponential distribution, continuous)

Poisson probability of k breakpoints is (1), and the exponential probability that the distance is bounded by n is (2), where μ is the average respectively for each.

Relation between cM distance and probability of crossover event

Genetic recombination can only be detected if there are an odd number of chromosomal crossovers between two markers. Assume the distance between two markers is d cM.

$$\begin{aligned} Pr[\text{recombination detected}] &= \sum_{k=0}^{\infty} \exp\left(-\frac{d}{100}\right) \cdot \frac{\left(\frac{d}{100}\right)^{2k+1}}{(2k+1)!} \\ &= \exp\left(-\frac{d}{100}\right) \cdot \sum_{k=0}^{\infty} \frac{\left(\frac{d}{100}\right)^{2k+1}}{(2k+1)!} \\ &= \exp\left(-\frac{d}{100}\right) \cdot \sinh\left(\frac{d}{100}\right) \\ &= \exp\left(-\frac{d}{100}\right) \cdot \left[\frac{\exp\left(\frac{d}{100}\right) - \exp\left(-\frac{d}{100}\right)}{2} \right] \\ &= \frac{1 - \exp\left(-\frac{2d}{100}\right)}{2} \end{aligned}$$

Some math identities used

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \dots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - x^3/3! + x^5/5! + \dots$$

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - x^2/2! + x^4/4! + \dots$$

$$\exp(ix) = \cos(x) - i \sin(x)$$

$$\exp(i\pi) = -1$$

$$\sinh(x) = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!}$$

$$\cosh(x) = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}$$

$$\sinh(x) = \frac{\exp(x) - \exp(-x)}{2}$$

Recombination rate

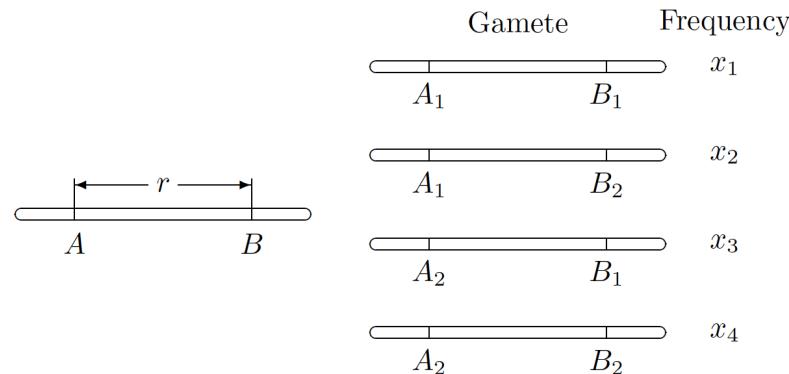


Figure 4.1: The chromosome on the left shows the position of the A and B loci. The right side illustrates the four possible gametes with their frequencies.

- Recombination rate

$$r = \text{Prob}[\text{recombinant gamete produced}]$$

- 4 possible gametes, with following probabilities

$$x_1 = \text{Prob}[A_1 B_1]$$

$$x_2 = \text{Prob}[A_1 B_2]$$

$$x_3 = \text{Prob}[A_2 B_1]$$

$$x_4 = \text{Prob}[A_2 B_2]$$

$$x_1 + x_2 + x_3 + x_4 = 1$$

$$p_1 = \text{Prob}[A_1] = x_1 + x_2$$

$$q_1 = \text{Prob}[A_2] = 1 - p_1 = x_3 + x_4$$

$$p_2 = \text{Prob}[B_1] = x_1 + x_3$$

$$q_2 = \text{Prob}[B_2] = 1 - p_2 = x_2 + x_4$$

- Given recombination rate r , and frequency x_1 of haplotype A_1B_1 in current generation, let x'_1 be frequency of haplotype A_1B_1 in next generation, where NO mutation or selection is allowed.

- Then

$$x'_1 = (1 - r)x_1 + rp_1p_2$$

where term $(1 - r)x_1$ represents the proportion of A_1B_1 haplotypes from the current generation which are not subject to recombination, and where rp_1p_2 is the proportion of current haplotypes (not necessarily only A_1B_1 haplotypes) that after undergoing recombination produce haplotype A_1B_1 in the next generation.

- The change in A_1B_1 haplotype frequency between two successive generations is thus

$$\Delta x_1 = x'_1 - x_1 = -rx_1 + rp_1p_2 = -r(x_1 - p_1p_2)$$

- Define **linkage disequilibrium** D to be $(x_1 - p_1p_2)$. Since this expression depends on the frequency x_1 of the A_1B_1 haplotype, the subscript 1 is added, so

$$D_1 = x_1 - p_1p_2$$

$$\Delta x_1 = -r(x_1 - p_1p_2) = rD_1$$

- In a similar fashion, we have

$$x'_2 = (1 - r)x_2 + rp_1q_2$$

$$\Delta x_2 = x'_2 - x_2 = -r(x_2 - p_1q_2) = -rD_2$$

$$D_2 = x_2 - p_1q_2$$

$$x'_3 = (1 - r)x_3 + rq_1p_2$$

$$\Delta x_3 = x'_3 - x_3 = -r(x_3 - q_1p_2) = -rD_3$$

$$D_3 = x_3 - q_1p_2$$

$$x'_4 = (1 - r)x_4 + rq_1q_2$$

$$\Delta x_4 = x'_4 - x_4 = -r(x_4 - q_1q_2) = -rD_4$$

$$D_4 = x_4 - q_1q_2$$

- **Claim:** $D_2 = -D_1 = -D$ and $D_3 = -D_1 = -D$ and $D_4 = D_1 = D$.

$$\begin{aligned} D_1 + D_2 &= (x_1 - p_1 p_2) + (x_2 - p_1 q_2) = (x_1 + x_2) - (p_1 p_2 + p_1 q_2) \\ &= (x_1 + x_2) - p_1(p_2 + q_2) = (x_1 + x_2) - p_1 = (x_1 + x_2) - (x_1 + x_2) = 0 \\ D_2 &= -D_1 = -D \end{aligned}$$

$$\begin{aligned} D_1 + D_3 &= (x_1 - p_1 p_2) + (x_3 - q_1 p_2) = (x_1 + x_3) - (p_1 p_2 + q_1 p_2) \\ &= (x_1 + x_3) - p_2(p_1 + q_1) = (x_1 + x_3) - p_2 = (x_1 + x_3) - (x_1 + x_3) = 0 \\ D_3 &= -D_1 = -D \end{aligned}$$

$$\begin{aligned} D_1 - D_4 &= (x_1 - p_1 p_2) - (x_4 - q_1 q_2) = (x_1 - x_4) - (p_1 p_2 - q_1 q_2) \\ q_1 q_2 &= (1 - p_1)(1 - p_2) = 1 - p_1 - p_2 + p_1 p_2 \\ p_1 p_2 - q_1 q_2 &= p_1 + p_2 - 1 = (x_1 + x_2) + (x_1 + x_3) - 1 \\ &= x_1 + [x_1 + x_2 + x_3 - 1] = x_1 - x_4 \\ D_1 + D_4 &= (x_1 - x_4) - (x_1 - x_4) = 0 \\ D_4 &= D_1 = D \end{aligned}$$

- It follows that

$$\begin{aligned} D &= x_1 - p_1 p_2 \\ -D &= x_2 - p_1 q_2 \\ -D &= x_3 - q_1 p_2 \\ D &= x_4 - q_1 q_2 \end{aligned}$$

so

$$\begin{aligned} x_1 &= p_1 p_2 + D \\ x_2 &= p_1 q_2 - D \\ x_3 &= q_1 p_2 - D \\ x_4 &= q_1 q_2 + D \end{aligned}$$

- The following table summarizes the results from the previous slide.

| haplotype frequency | A_1B_1 x_1 $p_1p_2 + D$ | A_1B_2 x_2 $p_1q_2 - D$ | A_2B_1 x_3 $q_1p_2 - D$ | A_2B_2 x_4 $q_1q_2 + D$ |
|---------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
|---------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|

- Claim:**

$$D = x_1x_4 - x_2x_3 \quad (5)$$

$$x_1x_4 = (p_1p_2 + D)(q_1q_2 + D) = p_1p_2q_1q_2 + D^2 + D(p_1p_2 + q_1q_2)$$

$$x_2x_3 = (p_1q_2 - D)(q_1p_2 - D) = p_1p_2q_1q_2 + D^2 - D(p_1q_2 + q_1p_2)$$

$$\begin{aligned} x_1x_4 - x_2x_3 &= D(p_1p_2 + q_1q_2) + D(p_1q_2 + q_1p_2) \\ &= D(p_1p_2 + q_1q_2 + p_1q_2 + q_1p_2) = D \end{aligned}$$

- Coupling gametes: A_1B_1 and A_2B_2 (corresponding to x_1 and x_4)

- Repulsion gametes: A_1B_2 and A_2B_1 (corresponding to x_2 and x_3)

- The sign $\pm D$ in the table is easy to remember: + for coupling gametes, – for repulsion gametes. Moreover, the result $D = x_1x_4 - x_2x_3$ states that **linkage disequilibrium** D equals the product of frequencies of the **coupling** gametes (haplotypes) minus the product of frequencies of the **repulsion** gametes (haplotypes)

- After one round of random mating in an infinite population, but in absence of mutation, selection (but allowing recombination)

$$x_1 = p_1p_2 + D$$

$$x'_1 = p'_1p'_2 + D' \quad \text{in next generation}$$

$x'_1 = p_1p_2 + D'$ since allele frequencies constant in absence of drift, mutation, selection

$$\Delta x_1 = x'_1 - x_1 = D' - D$$

$$x'_1 = (1 - r)x_1 + rp_1p_2$$

$$\Delta x_1 = x'_1 - x_1 = -rD$$

$$D' = D + \Delta x_1 = D - rD = D(1 - r)$$

- Iterating this, we have

$$\begin{aligned} D'' &= D'(1 - r) = D(1 - r)^2 \\ D''' &= D''(1 - r) = D'(1 - r)^2 = D(1 - r)^3 \end{aligned}$$

- If we consider D_0 to be the initial linkage disequilibrium (generation 0 with a crossover event that caused A_1B_1 haplotype), and let $D^{(t)}$ denote the linkage disequilibrium in the t th generation, then by induction

$$D^{(t)} = D^{(t-1)}(1 - r) = D_0(1 - r)^t = D(1 - r)^t \quad (6)$$

which is reminiscent of the decay of heterozygosity due to drift $\mathcal{H}^{(t)} = \mathcal{H}_0 \cdot (1 - \frac{1}{2N})^t$.

- Three pages before, we showed that

$$x_1 = p_1 p_2 + D \quad x_2 = p_1 q_2 - D \quad x_3 = q_1 p_2 - D \quad x_4 = q_1 q_2 + D$$

Since x_1, x_2, x_3, x_4 are probabilities, $x_1 \geq 0$, etc., we have

$$\begin{aligned} p_1 p_2 + D \geq 0 &\Rightarrow D \geq -p_1 p_2 \Rightarrow p_1 p_2 \geq -D \\ p_1 q_2 - D \geq 0 &\Rightarrow p_1 q_2 \geq D \\ q_1 p_2 - D \geq 0 &\Rightarrow q_1 p_2 \geq D \\ q_1 q_2 + D \geq 0 &\Rightarrow D \geq -q_1 q_2 \Rightarrow q_1 q_2 \geq -D \\ D \geq 0 &\Rightarrow D \leq \min(p_1 q_2, q_1 p_2) \\ D \leq 0 &\Rightarrow -D \leq \min(p_1 p_2, q_1 q_2) \end{aligned}$$

Since linkage disequilibrium values $D = x_1 x_4 - x_2 x_3$ depend on genotype frequencies, $|D|$ can be quite low. The normalized value D' , defined below, is more useful, since it satisfies $-1 \leq D' \leq 1$. (Warning: D' here has nothing to do with the temporarily used symbol D' on the previous page.)

$$D' = \begin{cases} \frac{D}{\min(p_1 q_2, q_1 p_2)} & \text{if } D \geq 0 \\ \frac{-D}{\min(p_1 p_2, q_1 q_2)} & \text{if } D < 0 \end{cases}$$

Linkage disequilibrium $D = \text{Cov}[X, Y]$

- Here we show that linkage disequilibrium D is equal to the **covariation** between genes A and B – technically, $D = \text{Cov}[X, Y]$, the **covariance** between X and Y , the **indicator functions** respectively for genes A and B .
- Suppose that X, Y are random variables that represent allele probabilities at genes A, B , so that

$$X = \begin{cases} 1 & \text{if allele is } A_1 \\ 0 & \text{if allele is } A_2 \end{cases}$$

and

$$Y = \begin{cases} 1 & \text{if allele is } B_1 \\ 0 & \text{if allele is } B_2 \end{cases}$$

- Random variables, which are defined to equal 1 if a condition applies and 0 if the condition does not apply, are called **indicator functions**.
- Let \bar{X} denote the expected value $E[X]$ and \bar{Y} denote the expected value $E[Y]$, and recall the definitions of haplotype (gamete) relative frequencies x_1, x_2, x_3, x_4 respectively for haplotypes $A_1B_1, A_1B_2, A_2B_1, A_2B_2$.

$$E[X] = 1 \cdot p_1 + 0 \cdot q_1 = p_1$$

$$E[Y] = 1 \cdot p_2 + 0 \cdot q_2 = p_2$$

$$E[XY] = 1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 = x_1$$

$$V[X] = E[X^2] - E[X]^2 = p_1 - p_1^2 = p_1(1 - p_1) = p_1 q_1$$

$$\sigma_X = \sqrt{V[X]} = \sqrt{p_1 q_1}$$

$$V[Y] = E[Y^2] - E[Y]^2 = p_2 - p_2^2 = p_2(1 - p_2) = p_2 q_2$$

$$\sigma_Y = \sqrt{V[Y]} = \sqrt{p_2 q_2}$$

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - \bar{X})(Y - \bar{Y})] = E[XY - p_1 Y - p_2 X + p_1 p_2] \\ &= E[XY] - p_1 p_2 - p_2 p_1 + p_1 p_2 = E[XY] - p_1 p_2 \quad (\text{by additivity of expectation}) \\ &= x_1 - p_1 p_2 = D \end{aligned}$$

- It follows that **linkage disequilibrium** D is simply the **covariance** between the **indicator functions** for genes A, B .
- From statistics, recall that the **correlation coefficient** of random variables X and Y is defined by

$$\rho = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y respectively denote the standard deviation of X respectively Y . Usually ρ denotes the correlation coefficient of a **population**, while r denotes the correlation coefficient of a **sample**. Gillespie uses r for the former, so we'll confound ρ and r without cause of confusion. Hence

$$r = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y} = \frac{D}{\sqrt{p_1 q_1 p_2 q_2}}$$

$$r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$$

Statistical test if linkage disequilibrium is non-zero

- Our goal is to perform a χ^2 **goodness-of-fit test** for whether the **haplotype frequencies** x_1, x_2, x_3, x_4 are approximately equal to expected haplotype frequencies if genes A and B are not genetically linked.
- Null hypothesis** H_0 is the assertion that $D = 0$, i.e. that $\text{Cov}[X, Y] = 0$ there is **no genetic linkage** between genes A and B . Given desired significance $\alpha = 0.05$, determine contingency table of observed and expected counts. If p -value for χ^2 statistic with $df = 1$ is less than α , then **reject** the null hypothesis.
- To compute the χ^2 statistic

$$\chi^2 = \sum_{i=1}^m \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i}$$

where there m classes, and \mathcal{O}_i denotes the **observed number** of instances in the i th class, and \mathcal{E}_i denotes the **expected number** of instances in the i th class.

- The number of **degrees of freedom** $df = m - 1$, unless additional degrees of freedom can be removed.
- In practice, one **estimates** relative frequencies p_1 [resp. p_2] from **observed** data by dividing the first row sum [resp. first column sum] by the number n of observations. From p_1, p_2 , one immediately obtains $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.
- In the χ^2 goodness-of-fit test for $m = 4$ classes, the number of **degrees of freedom** $df = m - 1 = 3$. However, we remove 2 additional degrees of freedom due to having estimated p_1 and q_1 from the data, so $df = 1$.

Observed counts

| $A \setminus B$ | B_1 | B_2 | sum |
|-----------------|--------|--------|--------|
| A_1 | nx_1 | nx_2 | np_1 |
| A_2 | nx_3 | nx_4 | nq_1 |
| sum | np_2 | nq_2 | n |

Expected counts if alleles independent

| $A \setminus B$ | B_1 | B_2 | sum |
|-----------------|------------|------------|--------|
| A_1 | $np_1 p_2$ | $np_1 q_2$ | np_1 |
| A_2 | $nq_1 p_2$ | $nq_1 q_2$ | nq_1 |
| sum | np_2 | nq_2 | n |



$$\begin{aligned}
 \chi^2 &= \frac{(nx_1 - np_1 p_2)^2}{np_1 p_2} + \frac{(nx_2 - np_1 q_2)^2}{np_1 q_2} + \frac{(nx_3 - nq_1 p_2)^2}{nq_1 p_2} + \frac{(nx_4 - nq_1 q_2)^2}{nq_1 q_2} \\
 &= n \cdot \frac{D^2}{p_1 q_1 p_2 q_2} \quad (\text{this inference is justified in a homework problem}) \\
 &= nr^2
 \end{aligned}$$

- It follows that the χ^2 statistic is just

$$\chi^2 = nr^2 = n \cdot \frac{D^2}{p_1 q_1 p_2 q_2} = n \cdot \frac{(x_1 x_4 - x_2 x_3)^2}{p_1 q_1 p_2 q_2}$$

- So to test whether genes A, B are **statistically linked**, perform χ^2 test for test statistic nr^2 with $df = 1$ since the number of degrees of freedom is $(\text{rows} - 1)(\text{cols} - 1)$. Null hypothesis H_0 is the assertion “linkage disequilibrium $D = 0$ ”. Using Excel, for instance, if p -value from χ^2 test with $df = 1$

$$p = CHISQ.DIST.RT(nr^2, 1)$$

is less than (say) $\alpha = 0.05$, then we REJECT the null hypothesis, and have sufficient statistical evidence to assert that $D \neq 0$. The value of D may be positive or negative (genes may be positively or negatively correlated). If $D > 0$ and $p < \alpha = 0.05$, then we can **not** assert that we have sufficient statistical evidence that $D > 0$ – we can though if $p < \frac{\alpha}{2}$.

Computing LD from genotype data

| | | B locus | | | | | | | | | | |
|--|------|---------|-------|-------|-------------------------|--|--|--|--|--|--|--|
| | | "11" | "12" | "22" | Totals | | | | | | | |
| A locus | "11" | a | b | c | a+b+c | | | | | | | |
| | "12" | d | e | f | d+e+f | | | | | | | |
| | "22" | g | h | i | g+h+i | | | | | | | |
| Totals | | a+d+g | b+e+h | c+f+i | $N = a+b+c+d+e+f+g+h+i$ | | | | | | | |
| D = $P(A1,B1) - P(A1)*P(B1)$ | | | | | | | | | | | | |
| $D = (a + b/2 + d/2 + e/4)/N - p(x=1)*p(y=1)$ | | | | | | | | | | | | |
| $\rho = D/\sqrt{P(A1) * P(A2) * P(B1) * P(B2)}$ | | | | | | | | | | | | |
| If $D > 0$, D' is $D/\min(P(A1)*P(B2), P(A2)*P(B1))$ | | | | | | | | | | | | |
| If $D < 0$, D' is $-D/\min(P(A1)*P(B1), P(A2)*P(B2))$ | | | | | | | | | | | | |

| | | B locus | | | |
|---------|------|---------|------|------|--------|
| | | "11" | "12" | "22" | Totals |
| A locus | "11" | 10 | 50 | 20 | 80 |
| | "12" | 15 | 60 | 25 | 100 |
| | "22" | 5 | 0 | 35 | 40 |
| Totals | | 30 | 110 | 80 | 220 |

$$p_A = P(A1) = P(\text{diploid A locus is "11"}) + \frac{1}{2} P(\text{diploid A locus is "12"})$$

$$p_B = P(B1) = P(\text{diploid B locus is "11"}) + \frac{1}{2} P(\text{diploid B locus is "12"})$$

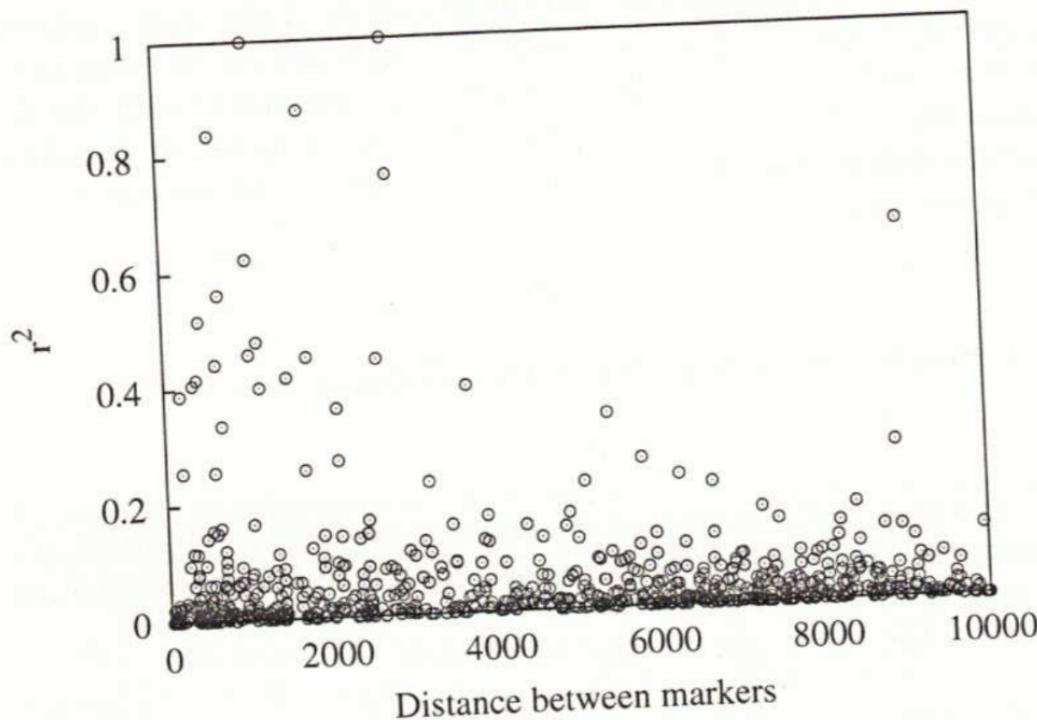
| | | | | | |
|----------|--------|--------|-------|---------|-------|
| pA | 0.5909 | (1-pA) | 0.409 | rho | 0.138 |
| pB | 0.3864 | (1-pB) | 0.614 | n rho^2 | 4.195 |
| P(A1,B1) | 0.261 | | | p-val | 0.041 |
| D | 0.0331 | | | | |

If D > 0, D' is $D/\min(P(A1)*P(B2), P(A2)*P(B1))$

If D < 0, D' is $-D/\min(P(A1)*P(B1), P(A2)*P(B2))$

$$D' \quad 0.2092$$

$$r^2 = \frac{D^2}{p_1 q_1 p_2 q_2} \text{ plotted as function of base pair distance}$$



- In this figure, large values of r^2 are more common for gene pairs A, B within 5 kb (kilo base pairs) than more distant pairs.
- Plot of linkage disequilibrium, as measured by $r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$, between pairs of restriction sites. For 2010 data, see Robbins et al. "Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato", Journal of Experimental Botany, Vol. 62, No. 6, pp. 1831-1845, 2011 (Image from *Population Genetics* by Gillespie.)

Mean, first and third quartile LD values

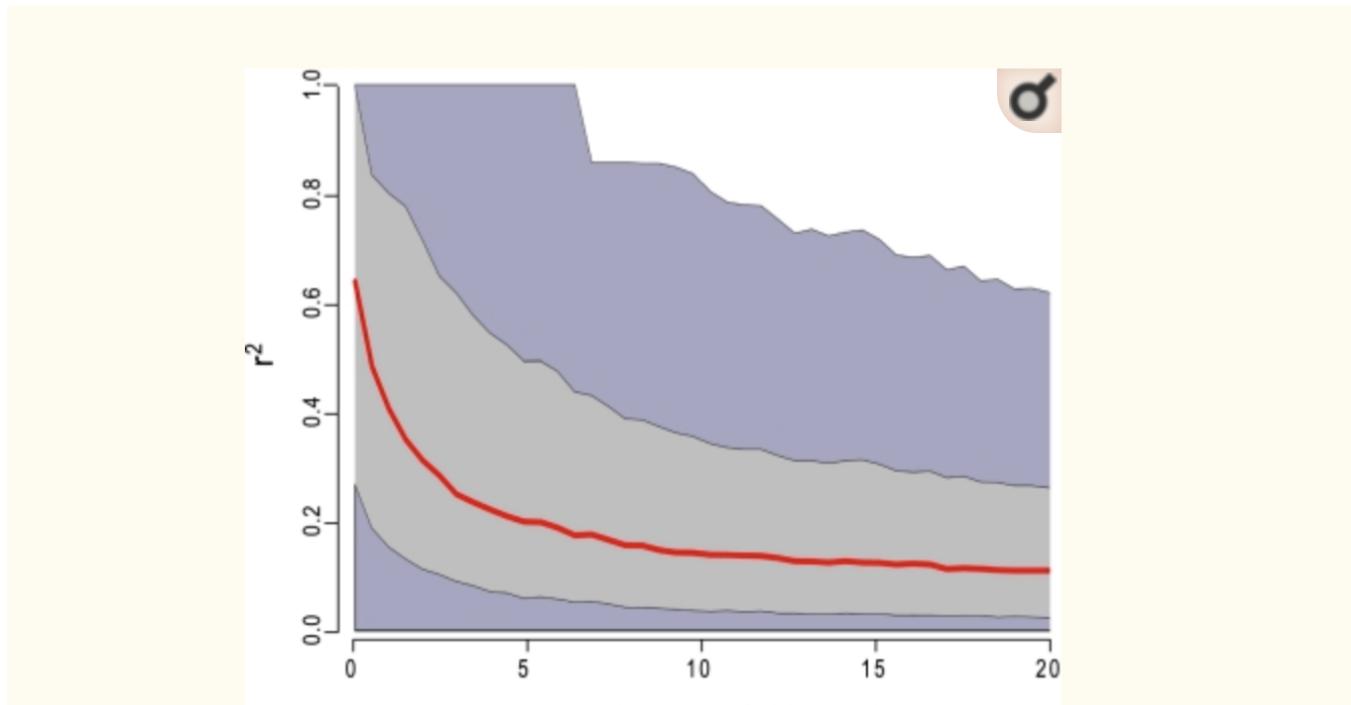


Fig. 6.

Mean LD decay (red line) as measured by pairwise r^2 , with the 50% and 90% ranges of values shown in light and dark gray, respectively.

- Image from Branca et al. "Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*", Proc Natl Acad Sci U S A. 2011 Oct 18; 108(42): E864-E870.

Haplotype blocks and recombination hot spots

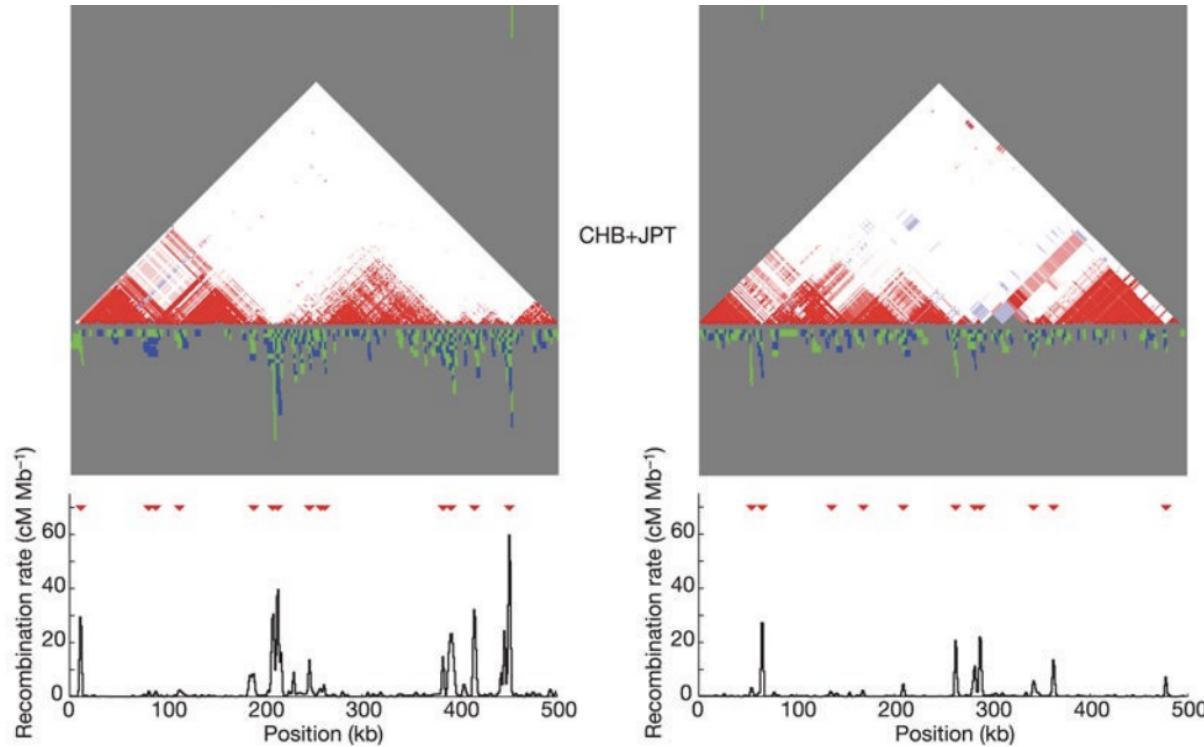


Figure 8. Comparison of linkage disequilibrium and recombination for two ENCODE regions
For each region (ENr131.2q37.1 and ENm014.7q31.33), D' plots for the YRI, CEU and CHB +JPT analysis panels are shown: white, $D' < 1$ and LOD < 2; blue, $D' = 1$ and LOD < 2; pink, $D' < 1$ and LOD ≥ 2; red, $D' = 1$ and LOD ≥ 2. Below each of these plots is shown the intervals where distinct obligate recombination events must have occurred (blue and green indicate adjacent intervals). Stacked intervals represent regions where there are multiple recombination events in the sample history. The bottom plot shows estimated recombination rates, with hotspots shown as red triangles⁴⁶.

- Image from International HapMap Consortium, Nature (2005).

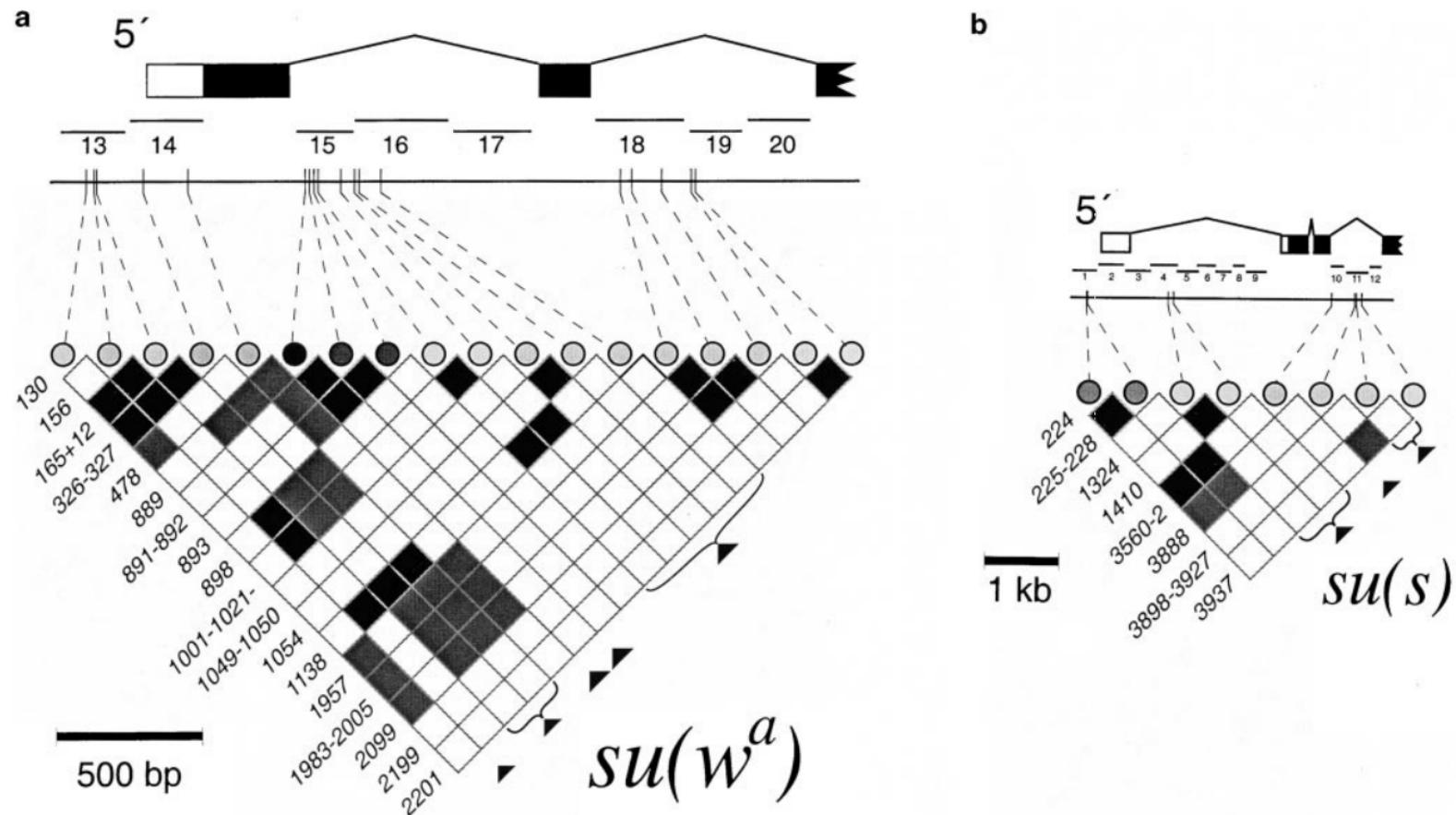


FIGURE 2.—Polymorphism and linkage disequilibria in the European sample. The positions of polymorphic sites detected (ticks on the third level) in the survey of the 5' portion of the *su(w^a)* (a) and of *su(s)* (b) in 51 alleles from a European population are indicated below a depiction of the gene structure (the open box is the 5' untranslated portion, the solid boxes are coding portions of exons, and the thin lines are the introns). Only those polymorphisms at which the least common state occurred twice or more in the sample are shown. Directly below the gene structure are the positions and sizes of the PCR fragments surveyed by SSCP and DNA sequencing. The triangular matrix of squares represents the statistical significance determined by Fisher's exact test (uncorrected for multiple tests); white, $P > 0.05$; light gray, $P < 0.05$; dark gray, $P < 0.01$; and black, $P < 0.005$. The dashed lines connect the polymorphic sites to their corresponding columns in the matrix. The shading of the circles at the top of each column increases with the expected heterozygosity of the polymorphic site. Along the left margin of the matrix are the positions in the published sequence of the corresponding polymorphic sites. Along the right margin are the inferred positions of the “minimum” number of recombination events in the history of the sample (HUDSON and KAPLAN 1985).

Selection and fitness of 2 loci A, B

Fitness for genotype $A_iB_j/A_{i'}B_{j'}$

| | A_1B_1 | A_1B_2 | A_2B_1 | A_2B_2 |
|----------|-----------------|-------------------|-------------------|-------------------|
| A_1B_1 | $x_1^2 w_{1,1}$ | $2x_1x_2 w_{1,2}$ | $2x_1x_3 w_{1,3}$ | $2x_1x_4 w_{1,4}$ |
| A_1B_2 | . | $x_2^2 w_{2,2}$ | $2x_2x_3 w_{2,3}$ | $2x_2x_4 w_{2,4}$ |
| A_2B_1 | . | . | $x_3^2 w_{3,3}$ | $2x_3x_4 w_{3,4}$ |
| A_2B_2 | . | . | . | $x_4^2 w_{4,4}$ |

- Identify 1 with haplotype A_1B_1 , 2 with haplotype A_1B_2 , 3 with haplotype A_2B_1 , and 4 with haplotype A_2B_2 . This identification was tacitly assumed when we defined frequencies x_1, x_2, x_3, x_4 respectively for haplotypes $A_1B_1, A_1B_2, A_2B_1, A_2B_2$.
- In a similar fashion, $w_{i,j}$ is defined to be the fitness, when haplotype i is located on one chromosome and haplotype j on the other chromosome in a **diploid** organism. For instance, fitness of genotype A_2B_1/A_1B_2 is denoted by $w_{3,2}$. Since paternal and maternal chromosomes cannot be identified, $w_{3,2} = w_{2,3}$, and more generally

$$w_{i,j} = w_{j,i}$$

Moreover, we assume that

$$w_{1,4} = w_{2,3}$$

which asserts there are no **fitness cis-trans effects** – i.e. fitness depends only on the presence of both alleles A_1, A_2 at the locus of gene A , and on the presence of both alleles B_1, B_2 at the locus of gene B , so fitness $w_{1,4}$ of A_1B_1/A_2B_2 is equal to the fitness $w_{2,3}$ of A_1B_2/A_2B_1 , because both genotypes contain the same individual alleles A_1, A_2, B_1, B_2 .

- Let \bar{w}_i denote the **marginal fitness** while \bar{w} denotes the **expected fitness** of the population.

$$\bar{w}_i = \sum_{j=1}^4 x_j w_{i,j}$$

$$\bar{w}_1 = x_1 w_{1,1} + x_2 w_{1,2} + x_3 w_{1,3} + x_4 w_{1,4}$$

$$\bar{w}_2 = x_1 w_{2,1} + x_2 w_{2,2} + x_3 w_{2,3} + x_4 w_{2,4}$$

$$\bar{w}_3 = x_1 w_{3,1} + x_2 w_{3,2} + x_3 w_{3,3} + x_4 w_{3,4}$$

$$\bar{w}_4 = x_1 w_{4,1} + x_2 w_{4,2} + x_3 w_{4,3} + x_4 w_{4,4}$$

$$\begin{aligned}\bar{w} &= \sum_{i=1}^4 \sum_{j=1}^4 x_i x_j w_{i,j} = \sum_{i=1}^4 x_i^2 w_{i,i} + 2 \sum_{1 \leq i < j \leq 4} x_i x_j w_{i,j} \\ &= \sum_{i=1}^4 x_i \cdot \left(\sum_{j=1}^4 x_j w_{i,j} \right) = \sum_{i=1}^4 x_i \cdot \bar{w}_i\end{aligned}$$

- Our next step is to compute the gamete frequency x'_i in the next generation, which depends on contributions from all **pairs of gametes** that contain the corresponding genotype of type i . Consider x'_1 first, which is obtained by **fitness-weighted** contributions from all **pairs of gametes** containing $A_1 B_1$.

The product $x_1 x_1$ concerns genotype $A_1 B_1 / A_1 B_1$, which is invariant under recombination. The product $x_1 x_2$ concerns genotype $A_1 B_1 / A_1 B_2$, which is invariant under recombination. The product $x_1 x_3$ concerns genotype $A_1 B_1 / A_2 B_1$, which is invariant under recombination. The product $x_1 x_4$ concerns genotype $A_1 B_1 / A_2 B_2$, which becomes $A_2 B_1 / A_1 B_2$ under recombination, and so only the **non-recombinant** pairs of gametes make a contribution to x'_1 . It follows that product $x_1 x_4$ must be multiplied by the probability $(1 - r)$ of not recombining, thus yielding $(1 - r)x_1 x_4$.

Moreover, the product $x_2 x_3$ concerns genotype $A_1 B_2 / A_2 B_1$ which changes to $A_2 B_2 / A_1 B_1$ under recombination, and so **only the recombinant** portion contributes towards x'_1 . It follows that product $x_2 x_3$ must be multiplied by the probability r of recombining, thus yielding $r x_2 x_3$. No other products $x_i x_j$ can contribute to frequency x'_1 of $A_1 B_1$.

- Recall that linkage disequilibrium D satisfies

$$D = x_1x_4 - x_2x_3$$

- Given our discussion on the previous page, it follows that in the presence of selection, the frequency x'_1 of the A_1B_1 haplotype next generation satisfies

$$x'_1 = \frac{x_1^2 w_{1,1} + x_1 x_2 w_{1,2} + x_1 x_3 w_{1,3} + x_1 x_4 (1 - r) w_{1,4} + r x_2 x_3 w_{2,3}}{\bar{w}}$$

The first 3 terms of the numerator are clear. Factor $(1 - r)$ is applied to the 4th term, since this applies in the case there was **no recombination** occurring between gene A locus and gene B locus – recombination with A_1B_1/A_2B_2 produces A_1B_2/A_2B_1 . The 5th term applies in the case there **is recombination** occurring between gene A locus and gene B locus – here, recombination with A_1B_2/A_2B_1 produces A_1B_1/A_2B_2 .

- More generally, recombination is of consequence only cases of (1) two distinct coupling gametes, or (2) two distinct repulsion gametes.

- Recalling that

$$\bar{w}_1 = x_1 w_{1,1} + x_2 w_{1,2} + x_3 w_{1,3} + x_4 w_{1,4}$$

we have

$$\begin{aligned} x'_1 &= \frac{x_1 \bar{w}_1 - r [x_1 x_4 w_{1,4} - x_2 x_3 w_{2,3}]}{\bar{w}} \\ &= \frac{x_1 \bar{w}_1 - r w_{1,4} [x_1 x_4 - x_2 x_3]}{\bar{w}} \quad (\text{assuming that } w_{1,4} = w_{2,3}) \\ &= \frac{x_1 \bar{w}_1 - r w_{1,4} \cdot D}{\bar{w}} \\ \Delta x_1 = x'_1 - x_1 &= \frac{x_1 \bar{w}_1 - r w_{1,4} \cdot D}{\bar{w}} - \frac{x_1 \bar{w}}{\bar{w}} = \frac{x_1 (\bar{w}_1 - \bar{w}) - r w_{1,4} \cdot D}{\bar{w}} \end{aligned}$$

- The 2-locus genotype fitness table (2 slides ago) is symmetric, so using the second row of the table, we compute the frequency x'_2 from the frequencies of computation with the genotypes A_2B1/A_1B_1 , A_2B1/A_1B_2 , A_2B1/A_2B_1 , A_2B1/A_2B_2 .

$$\begin{aligned} D &= x_1 x_4 - x_2 x_3 \\ x'_2 &= \frac{x_1 x_2 w_{1,2} + x_2^2 w_{2,2} + x_2 x_3 (1 - r) w_{2,3} + x_2 x_4 w_{2,4} + r x_1 x_4 w_{1,4}}{\bar{w}} \\ &= \frac{x_2 \bar{w}_2 - r [x_2 x_3 w_{2,3} - x_1 x_4 w_{1,4}]}{\bar{w}} \\ &= \frac{x_2 \bar{w}_2 - r w_{1,4} [x_2 x_3 - x_1 x_4]}{\bar{w}} \quad (\text{assuming that } w_{1,4} = w_{2,3}) \\ &= \frac{x_2 \bar{w}_2 + r w_{1,4} \cdot D}{\bar{w}} \end{aligned}$$

- In homework, you are asked to similarly show that

$$x'_3 = \frac{x_3 \bar{w}_3 + r w_{1,4} \cdot D}{\bar{w}}$$

$$x'_4 = \frac{x_4 \bar{w}_4 - r w_{1,4} \cdot D}{\bar{w}}$$

- We have

$$\Delta x_2 = x'_2 - x_2 = \frac{x_2 \bar{w}_2 + r w_{1,4} \cdot D}{\bar{w}} - \frac{x_2 \bar{w}}{\bar{w}} = \frac{x_2 (\bar{w}_2 - \bar{w}) + r w_{1,4} \cdot D}{\bar{w}}$$

- Similarly, one shows analogous results for Δx_3 and Δx_4 . Summarizing, we have the following.

$$\Delta x_1 = \frac{x_1 (\bar{w}_1 - \bar{w}) - r w_{1,4} \cdot D}{\bar{w}}$$

$$\Delta x_2 = \frac{x_2 (\bar{w}_2 - \bar{w}) + r w_{1,4} \cdot D}{\bar{w}}$$

$$\Delta x_3 = \frac{x_3 (\bar{w}_3 - \bar{w}) + r w_{1,4} \cdot D}{\bar{w}}$$

$$\Delta x_4 = \frac{x_4 (\bar{w}_4 - \bar{w}) - r w_{1,4} \cdot D}{\bar{w}}$$

- Note well that the **coupling** terms have $-w_{1,4}D$, while the **repulsion** terms have $w_{1,4}D$. Since next generation values x'_1, x'_2, x'_3, x'_4 differ from current values x_1, x_2, x_3, x_4 , linkage disequilibrium $D' = x'_1 x'_4 - x'_2 x'_3$ in the next generation differs from current value D – however, recall we proved $D' = (1 - r)D$ assuming **no selection**, and here we **include selection**.

Fitness values for gene hitchhiking model of J. Maynard Smith and J. Haigh

- **Assumptions:** (1) selection operates on the locus A , while locus B is **neutral**. (2) **homozygote** A_1/A_1 fitness is greater than or equal to **heterozygote** A_1/A_2 fitness, which is greater than or equal to **homozygote** A_2/A_2 fitness.
- Fitness model of **J. Maynard Smith and J. Haigh**

| | A_1B_1 | A_1B_2 | A_2B_1 | A_2B_2 |
|----------|----------|----------|----------|----------|
| A_1B_1 | 1 | 1 | 1-hs | 1-hs |
| A_1B_2 | . | 1 | 1-hs | 1-hs |
| A_2B_1 | . | . | 1-s | 1-s |
| A_2B_2 | . | . | . | 1-s |

- Previous table is **symmetric**, while following table explicitly presents the all fitness values – this makes it clear that fitness only depends on the allele of **gene A**.

| | A_1B_1 | A_1B_2 | A_2B_1 | A_2B_2 |
|----------|----------|----------|----------|----------|
| A_1B_1 | 1 | 1 | 1-hs | 1-hs |
| A_1B_2 | 1 | 1 | 1-hs | 1-hs |
| A_2B_1 | 1-hs | 1-hs | 1-s | 1-s |
| A_2B_2 | 1-hs | 1-hs | 1-s | 1-s |

- In Smith-Haigh model for gene hitchhiking, **fitness** of the **gamete** A_iB_j depends **only** on the fitness of the allele of gene A (since B is considered to be neutral). Thus fitness equals 1 if **both** chromosomes contain A_1 at locus – i.e. locus is A_1 -**homozygous**. Fitness equals $1 - hs$ if one chromosome contains A_1 and the other A_2 at locus – i.e. locus is A_1/A_2 -**heterozygous**. Fitness equals $1 - s$ if **both** chromosomes contain A_2 at locus – i.e. locus is A_2 -**homozygous**.

Change in A_1 and B_1 allele frequency when B locus hitchhikes along with selected A locus

- On page 6 of Chapter 3 notes, we showed that change in A_1 allele frequency under selection satisfies

$$\begin{aligned}\Delta_s p &= \frac{pq}{\bar{w}} (p(w_{1,1} - w_{1,2}) - q(w_{2,2} - w_{1,2})) \\ &= \frac{pq}{\bar{w}} (p(1 - (1 - hs)) - q((1 - s) - (1 - hs))) \\ \Delta_s p &= \frac{pq s}{\bar{w}} (ph - q(1 - h))\end{aligned}$$

where in Chapter 3, $w_{1,1}$ denotes the fitness of the A_1/A_1 homozygote (denoted A_1A_1 homozygote in that chapter), $w_{1,2}$ denotes fitness of A_1/A_2 heterozygote (denoted A_1A_2 in that chapter), and $w_{2,2}$ denotes fitness of A_2/A_2 homozygote (denoted A_2A_2 in that chapter). Note that this usage of $w_{i,j}$ is in **stark contrast** to the **different** usage of $w_{i,j}$ in the current Chapter 4, where for instance $w_{2,3}$ denotes the fitness of A_1B_2/A_2B_1 .

- If gene B is neutral and only hitchhikes along with gene A , which latter is under selection, it can be shown that

$$\Delta x_1 + \Delta x_2 = \frac{p_1 q_1 s [p_1 h + q_1 (1 - h)]}{\bar{w}} = \frac{p_1 q_1 s [p_1 h + q_1 (1 - h)]}{1 - 2p_1 q_1 hs - q_1^2 s}$$

This is because x_1 is the frequency of A_1B_1/A_1B_1 . and x_2 the frequency of A_1B_1/A_1B_2 . so $\Delta x_1 + \Delta x_2 = \Delta p_1 = \Delta_s p$.

- The effect of **hitchhiking** can be numerically quantified by computing the change Δp_2 of B_1 allele frequency, which is a neutral hitchhiker, while gene A is under selection. This is given by the following expression, which will soon be proved:

$$\Delta x_1 + \Delta x_3 = \frac{Ds [p_1 h + q_1 (1 - h)]}{\bar{w}} = \frac{Ds [p_1 h + q_1 (1 - h)]}{1 - 2p_1 q_1 hs - q_1^2 s}$$

- Note that the only difference between this formula and the previous formula is that term $p_1 q_1$ is replaced by linkage disequilibrium D .

Numerical quantification of hitchhiking effect

- Our goal is to compute, specifically for the **Smith-Haigh fitness** model, the change Δp_2 in B_1 allele frequency, which is caused by hitchhiking neutral B locus along with the selected A locus.
- In the Smith-Haigh model, marginal fitnesses satisfy

$$\begin{aligned}
 \bar{w}_1 &= x_1 w_{1,1} + x_2 w_{1,2} + x_3 w_{1,3} + x_4 w_{1,4} = x_1 \cdot 1 + x_2 \cdot 1 + x_3(1 - hs) + x_4(1 - hs) \\
 &= (x_1 + x_2 + x_3 + x_4) - hs(x_3 + x_4) = 1 - q_1 hs \\
 \bar{w}_2 &= x_1 w_{2,1} + x_2 w_{2,2} + x_3 w_{2,3} + x_4 w_{2,4} = x_1 \cdot 1 + x_2 \cdot 1 + x_3(1 - hs) + x_4(1 - hs) \\
 &= (x_1 + x_2 + x_3 + x_4) - hs(x_3 + x_4) = 1 - q_1 hs \\
 \bar{w}_3 &= x_1 w_{3,1} + x_2 w_{3,2} + x_3 w_{3,3} + x_4 w_{4,4} = x_1(1 - hs) + x_2(1 - hs) + x_3(1 - s) + x_4(1 - s) \\
 &= (x_1 + x_2 + x_3 + x_4) - hs(x_1 + x_2) - s(x_3 + x_4) = 1 - p_1 hs - q_1 s \\
 \bar{w}_4 &= x_1 w_{4,1} + x_2 w_{4,2} + x_3 w_{4,3} + x_4 w_{4,4} = x_1(1 - hs) + x_2(1 - hs) + x_3(1 - s) + x_4(1 - s) \\
 &= (x_1 + x_2 + x_3 + x_4) - hs(x_1 + x_2) - s(x_3 + x_4) = 1 - p_1 hs - q_1 s
 \end{aligned}$$

- We now compute

$$\begin{aligned}
 \Delta x_1 &= \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{1,4} \cdot D}{\bar{w}} \\
 \Delta x_3 &= \frac{x_3(\bar{w}_3 - \bar{w}) + rw_{1,4} \cdot D}{\bar{w}} \\
 \Delta x_1 + \Delta x_3 &= \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{1,4} \cdot D}{\bar{w}} + \frac{x_3(\bar{w}_3 - \bar{w}) + rw_{1,4} \cdot D}{\bar{w}} \\
 &= \frac{(x_1 \bar{w}_1 + x_3 \bar{w}_3) - \bar{w}(x_1 + x_3)}{\bar{w}}
 \end{aligned}$$

- Recall that we have shown

$$\begin{aligned}x_1 &= p_1 p_2 + D \\x_3 &= q_1 p_2 - D\end{aligned}$$

- In the **first term** of the **numerator** of $\Delta x_1 + \Delta x_3$, replace x_1, x_3 and \bar{w}_1, \bar{w}_3 by equivalent expressions from above to get

$$\begin{aligned}x_1 \bar{w}_1 + x_3 \bar{w}_3 &= x_1(1 - q_1 hs) + x_3(1 - p_1 hs - q_1 s) \\&= (p_1 p_2 + D)(1 - q_1 hs) + (q_1 p_2 - D)(1 - p_1 hs - q_1 s) \\&= (p_1 p_2 + D - p_1 q_1 p_2 hs - D q_1 hs) + (q_1 p_2 - D - p_1 q_1 p_2 hs + D p_1 hs - q_1^2 p_2 s + D q_1 s) \\&= p_2(p_1 + q_1) - 2p_1 q_1 p_2 hs + Dhs(p_1 - q_1) - q_1^2 p_2 s + Dq_1 s \\&= p_2 - 2p_1 q_1 p_2 hs + Dhs(p_1 - q_1) + Dq_1 s - q_1^2 p_2 s\end{aligned}$$

- Similarly, the **second term** of the **numerator** of $\Delta x_1 + \Delta x_3$, replace x_1, x_3 and \bar{w} by equivalent expressions from above to get

$$\begin{aligned}\bar{w}(x_1 + x_3) &= (1 - 2p_1 q_1 hs - q_1^2 s)(x_1 + x_3) \\&= (1 - 2p_1 q_1 hs - q_1^2 s)[(p_1 p_2 + D) + (q_1 p_2 - D)] \\&= (1 - 2p_1 q_1 hs - q_1^2 s)[p_1 p_2 + q_1 p_2] = (1 - 2p_1 q_1 hs - q_1^2 s)[p_2(p_1 + q_1)] \\&= p_2(1 - 2p_1 q_1 hs - q_1^2 s) \\&= p_2 - 2p_1 q_1 p_2 hs - q_1^2 p_2 s\end{aligned}$$

- Thus the numerator of $\Delta x_1 + \Delta x_3$ is

$$\begin{aligned}\Delta x_1 + \Delta x_3 &= \frac{[p_2 - 2p_1 q_1 p_2 hs + Dhs(p_1 - q_1) + Dq_1 s - q_1^2 p_2 s] - [p_2 - 2p_1 q_1 p_2 hs - q_1^2 p_2 s]}{\bar{w}} \\&= \frac{Dhsp_1 - Dhsq_1 + Dq_1 s}{\bar{w}} = \frac{Ds(hp_1 - hq_1 + q_1)}{\bar{w}} = \frac{Ds[p_1 h + q_1(1 - h)]}{\bar{w}}\end{aligned}$$

Comments on hitchhiking

- The equation for $\Delta x_1 + \Delta x_3$ derived at the bottom of the previous page depends **only** on the **linkage disequilibrium** D and A_1 allele frequency p_1 – it does **not** depend on any allele frequencies at the B locus.
- If the **covariation** (linkage) between genes A and B is large, then the change in B_1 allele frequency, given by $\Delta p_2 = \Delta x_1 + \Delta x_3$ is large, since

$$\Delta x_1 + \Delta x_3 = \frac{Ds[p_1 h + q_1(1 - h)]}{\bar{w}}$$

- A good way to understand the effect of gene hitchhiking is to perform a computer simulation when a new mutation introduces a single **advantageous** A_1 allele on a chromosome near an existent B_1 allele. Thus the initial frequency of the new A_1 allele is $p_1 = \frac{1}{2N}$, where N is population size. Consider a neutral gene B not under selection. In the absence of genetic drift, which could cause the loss of the advantageous A_1 allele, **selection** will cause the A_1 allele to **sweep** through the population to **fixation**, as we've seen in Chapter 3 (see some of the figures generated in Chapter 3 notes by the code `naturalSelectionAlleleFrequencyPerGeneration.py` in the DEMOS directory for our class).
- The following table gives haplotype frequencies after the new advantageous mutation A_1 is introduced.

| haplotype | A_1B_1 | A_1B_2 | A_2B_1 | A_2B_2 |
|-----------|----------------|----------|------------------------------|---------------|
| frequency | $\frac{1}{2N}$ | 0 | $\frac{1}{2} - \frac{1}{2N}$ | $\frac{1}{2}$ |

- Note that linkage disequilibrium $D = x_1x_4 - x_2x_3 = \frac{1}{4N}$ initially.
- From the table above, the **heterozygosity** of the B locus is initially $H = 2p_2q_2 = \frac{1}{2}$.

Scaled heterozygosity of locus B as a function of recombination rate r between A and B

- As A_1 **sweeps** to fixation in the population, what happens to B_1 ? If B_1 hitchhikes, then it will increase in frequency, so the **heterozygosity** of the B locus will **decrease**. However the extent to which it decreases depends on its **distance** from the A locus – more specifically it depends on the **recombination rate** between the A and B loci.
- The following image is taken from Gillespie's text, while the slide after shows the output of the Python program `hitchhiking.py` in the DEMOS directory of our web server.

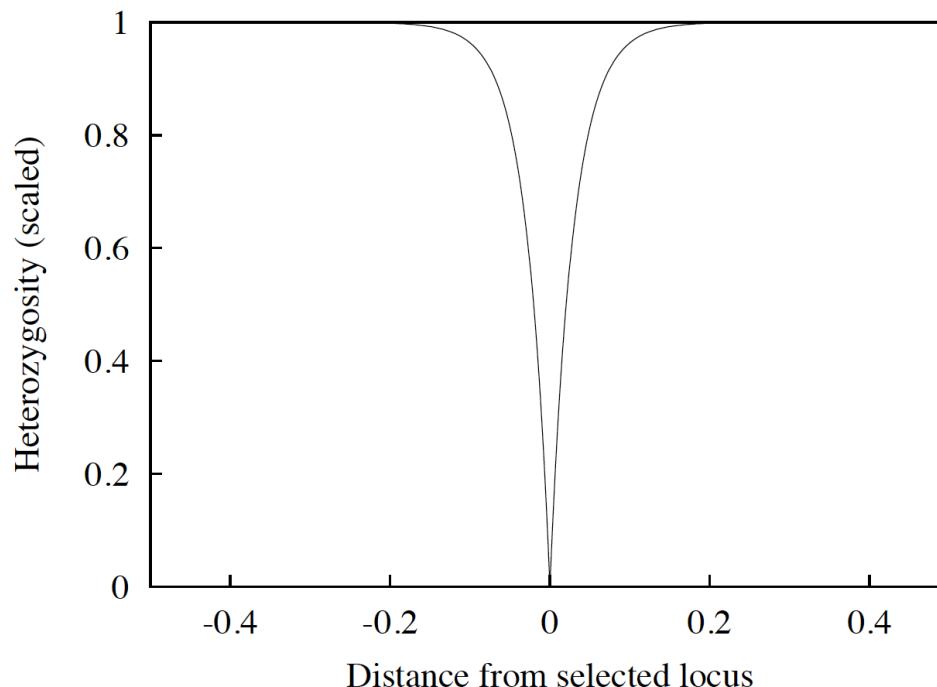
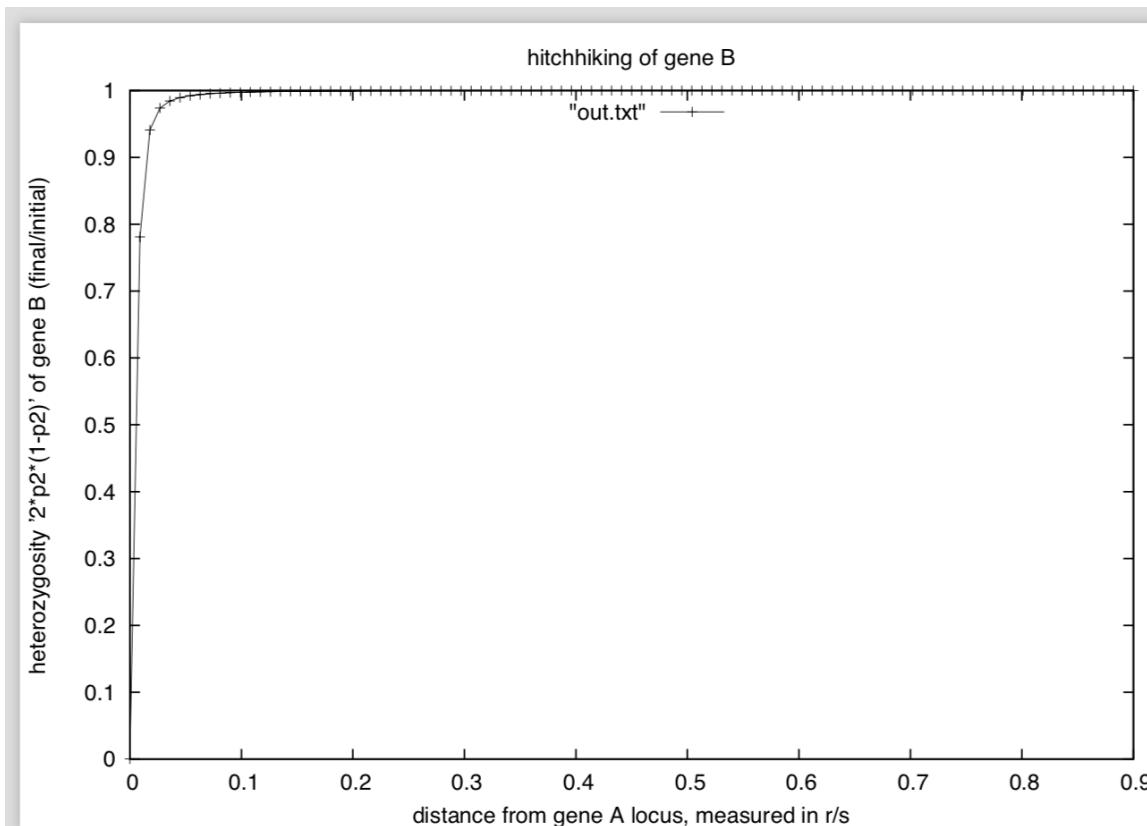


Figure 4.4: The ratio of the final to initial heterozygosity at a neutral locus as a function of the distance from the selected locus as measured by r/s . Negative values of r/s are left of the selected locus, positive values are to the right.

Python output for hitchhiking of gene B



- Figure shown for **fixed** selection coefficient $s = 0.1$. Thus x-axis is simply a **scaled** form of **recombination rate** r . Clearly recombination rate is correlated to (roughly proportional to) **base pair** distance – number of nucleotides separating genes A and B ; however, the correlation is rough, as shown by an earlier figure that graphed r^2 as a function of **marker distance**.

Approximate pseudocode to generate previous figure

```
def heterozygosity(r,s,N)
    x1 = 1.0/(2*N)      #one new A1 allele introduced
    x2 = 0
    x3 = 0.5 - 1.0/(2*N)
    x4 = 0.5
    p1 = x1+x2          #frequency of A1 allele
    p2 = x1+x3          #frequency of B1 allele
    H0 = 2*p2*(1-p2)    #H0 is initial heterozygosity of gene B
    while (1-p1)>epsilon: #while A1 not yet fixed
        x1 += deltaX1(x1,x2,x3,x4,r,s)
        x2 += deltaX2(x1,x2,x3,x4,r,s)
        x3 += deltaX3(x1,x2,x3,x4,r,s)
        x4 += deltaX4(x1,x2,x3,x4,r,s)
        p1  = x1+x2
        p2  = x1+x3      #x1+x3 is p2 defined by Prob[B1]
    return 2*p2*(1-p2) #heterozygosity of B locus

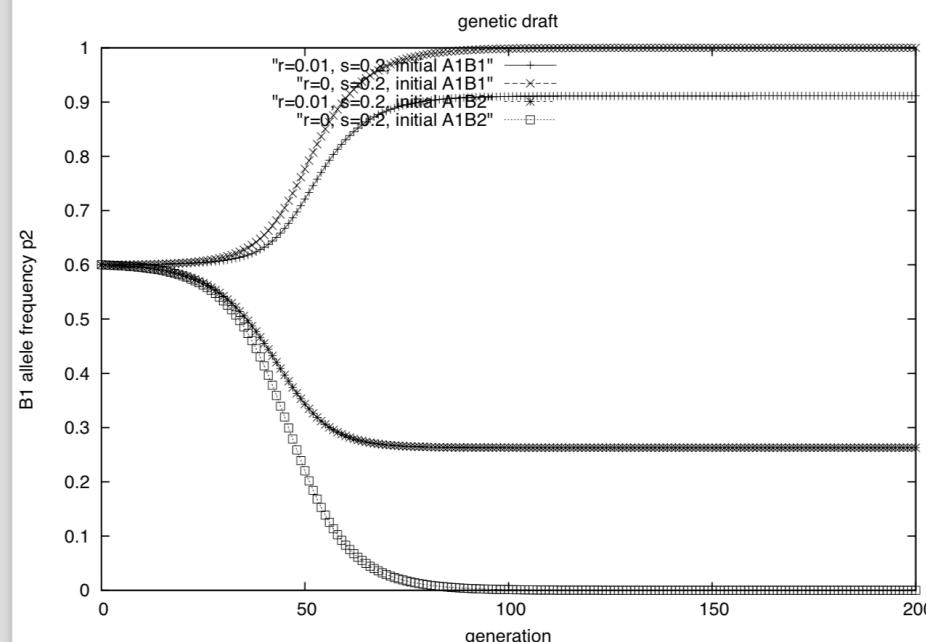
def main
    N      = 5000  #population size
    r      = 0      #initial recombination rate
    s      = 0.1   #fixed selection coefficient
    NUMSTEPS = 100
    for k = 1 to NUMSTEPS
        r  = r + s/NUMSTEPS
        val = heterozygosity(r,s,N)
        print "%s\t%s" % (r,val)
    ..
```

Genetic Draft

- Imagine that your lab is studying a particular gene X , which has several alleles with significant proportions in the population. Is one allele “fitter” than another? Is the locus X under positive selection? Or instead, is X possibly deleterious and potentially lethal in homozygous form? Or could X be located near a gene A , for which a particular allele A_1 is sweeping through the population, so that X is hitchhiking on A ?
- Example:** In the differentiation and evolution of *H. sapiens* from our closest primate relatives, did genes relating to hand dexterity undergo rapid evolution, subsequently driving the evolution of genes relating to growth of the cerebral cortex – or vice versa?
- One of the goals of population genetics is to develop a mathematically rigorous theory that explains how allele frequencies change over time, how genetic **sweeps** occur (when one allele becomes **fixed** in the population), how deleterious alleles are systematically removed from the population, etc.
- Genetic drift, pointwise mutation, selection, and recombination**, are **stochastic forces**, capable of modifying allele frequencies in the genome.
- Genetic draft** is the term used when gene **hitchhiking** is considered as a **stochastic force**, capable of modifying allele frequencies in the genome.
- Suppose that A_1 is a **new, advantageous** mutation at the A locus in an A_2A_2 homozygous population, and that the **neutral** B locus is located near the A locus. Assume that B is bi-allelic, with B_1 [resp. B_2] allele frequency equal to p_2 [resp. $q_2 = 1 - p_2$]. The new A_1 mutation (occurring in a single individual in the population of size N) could have occurred on a chromosome containing the B_1 allele, or instead on a chromosome containing the B_2 allele. The former occurs with probability p_2 , while the latter occurs with probability q_2 . Thus we have the following table.

| haplotype (gamete) | probability | x_1 | x_2 | x_3 | x_4 | $D = x_1x_4 - x_2x_3$ |
|--------------------|-------------|----------------|----------------|----------------------|----------------------|-----------------------|
| A_1B_1 | p_2 | $\frac{1}{2N}$ | 0 | $p_2 - \frac{1}{2N}$ | q_2 | $\frac{q_2}{2N}$ |
| A_1B_2 | q_2 | 0 | $\frac{1}{2N}$ | p_2 | $q_2 - \frac{1}{2N}$ | $-\frac{p_2}{2N}$ |

Hitchhiking of B1 (or B2) on A1



- Output from python code `geneticDraft.py` in DEMOS directory on web server. Population size $N = 200$, initial probability of allele B_1 is $p_2 = 0.6$, and heterozygosity factor $h = \frac{1}{2}$. Depending on whether an initial advantageous mutation A_1 is on the same chromosome (gamete) as B_1 [resp. B_2], as A_1 proceeds to **total fixation** (i.e. A_1 **sweeps** the population), the hitchhiking allele B_1 [resp. B_2] proceeds to **partial fixation** for $r/s > 0$ (**total fixation** for $r/s = 0$).

| haplotype (gamete) | probability | x_1 | x_2 | x_3 | x_4 | $D = x_1 x_4 - x_2 x_3$ |
|--------------------|-------------|----------------|----------------|----------------------|----------------------|-------------------------|
| $A_1 B_1$ | p_2 | $\frac{1}{2N}$ | 0 | $p_2 - \frac{1}{2N}$ | q_2 | $\frac{q_2}{2N}$ |
| $A_1 B_2$ | q_2 | 0 | $\frac{1}{2N}$ | p_2 | $q_2 - \frac{1}{2N}$ | $-\frac{p_2}{2N}$ |

- Recall that for the neutral locus B , which hitchhikes with A_1 , the B_1 allele frequency $p_2 = x_1 + x_3$, and we showed that

$$\Delta x_1 + \Delta x_3 = \frac{Ds[p_1 h + q_1(1 - h)]}{\bar{w}}$$

- For A_1B_1 haplotype, it follows that $D = \frac{q_2}{2N} > 0$, and so B_1 frequency will **increase**.
- In contrast, for the A_1B_2 haplotype, $D = -\frac{p_2}{2N} < 0$, and so B_1 frequency will **decrease**, and consequently, the B_2 frequency will **increase**.
- The reason that only **partial fixation** occurs with hitchhiking genes is that at some future generation, the **neutral, hitchhiking** gene B will be separated from the advantageous A_1 allele due to a **recombination** event. This is modeled mathematically in the **absence of selection** by our previous observation that linkage disequilibrium decreases every generation by factor $(1 - r)$ – so that $D^{(t)} = D_0(1 - r)^t$. As the current generation linkage disequilibrium value D decreases, so does the hijacking allele frequency $\Delta p_2 = \Delta x_1 + \Delta x_3$, as shown in the equation above.
- Note that all results in Chapter 3 (selection) and currently in Chapter 4 (two locus dynamics, recombination, hitchhiking) do **not** consider **genetic drift**. Indeed, we know that when population size N is large, the probability that a new mutation, with initial frequency $\frac{1}{2N}$, will **disappear** within 1 generation is $e^{-1} = \frac{1}{e} \approx 0.3679$.
- Accounting for drift in the simulation from the previous page is possible, but leads to complications in formulas, and it is much simpler to perform computer simulations. Nevertheless, the formulas described in Chapters 3 and 4 are (approximately) **valid** as soon as the initial frequency is $\gg \frac{1}{2N}$.

Fixation probability of neutral, advantageous and disadvantageous alleles

- Our ultimate goal is to describe **genetic draft** as a stochastic force having somewhat analogous properties to **genetic drift**. In order to do so, we need first to discuss **fixation probabilities** of an allele, usually a new mutation with initial frequency $\frac{1}{2N}$.
- Using the Kolmogorov forward equation (a.k.a. Fokker-Planck partial differential equation, and the derived **stationary distribution**, on pages 203-206 of Gillespie's text, there is a derivation in the case of **heterozygosity factor** $h = \frac{1}{2}$ that the **fixation probability** $\pi_1(p)$ of allele A_1 having current frequency p is equal to the following:

$$\pi_1(p) = \frac{1 - e^{-2Ns p}}{1 - e^{-2Ns}}$$

where N is population size, and $s \in [0, 1]$ is selection factor. Let's accept this on faith – don't be confused with an earlier result that $\pi_1(p) = p$ in the **absence of selection**. Both of these results are non-trivial, since genetic drift is considered.

- If we wish for a fixation of allele A_1 with probability $1 - \epsilon$, where $\epsilon = 0.01$, for instance, then writing $\pi_1(p)$ for the probability of fixation of A_1 with initial frequency p , we want

$$\pi_1(p) = \frac{1 - e^{-2Ns p}}{1 - e^{-2Ns}} \geq 1 - \epsilon$$

When N is large, $e^{-2Ns} \approx 0$ and

$$\begin{aligned}\pi_1(p) &\approx 1 - e^{-2Ns p} \\ \pi_1(p) &\geq 1 - \epsilon\end{aligned}$$

$$1 - e^{-2Ns p} \geq 1 - \epsilon \Leftrightarrow \epsilon \geq e^{-2Ns p} \Leftrightarrow \ln(\epsilon) \geq -2Ns p \Leftrightarrow 2Ns p \geq -\ln(\epsilon) \Leftrightarrow p \geq \frac{-\ln(\epsilon)}{2Ns}$$

Thus

$$\pi_1(p) \geq 1 - \epsilon \Leftrightarrow p \geq \frac{-\ln(\epsilon)}{2Ns}$$

- For example, if $\epsilon = 0.01$, population size $N = 1000$, selection factor $s = 0.1$, then it suffices for initial frequency $p \geq 0.02303$ for the corresponding allele to fix.
- Page 9 of Chapter 2 notes shows that the **probability of fixation** $\pi_1(p)$ of allele A_1 in a diploid population of size N , where initial frequency of A_1 is p , satisfies

$$\pi_1(p) = p$$

provided the locus A is **neutral** – i.e. in the presence of selection.

- Page 26 of Chapter 2 notes proves that the time T_2 to coalesce two remaining sequences equals $2N$. If mutation rate is u , then the expected number of mutations is

$$2Nu$$

- Page 18 of Chapter 2 notes shows that the **substitution rate** for a **neutral mutation** equals

$$\rho = 2Nu \cdot \pi_1(p) = 2Nup$$

- The substitution rate ρ equals the number of mutations multiplied by the **fixation rate**, so for a (single) new **neutral** mutation A_1 , the initial A_1 frequency $p = \frac{1}{2N}$, and

$$\rho = 2Nu \cdot \pi_1\left(\frac{1}{2N}\right) = 2Nu \cdot \frac{1}{2N} = u$$

- If A_1 is a new **advantageous** mutation with initial frequency $p = \frac{1}{2N}$, and if $h = \frac{1}{2}$, then the **fixation probability** of A_1 equals

$$\begin{aligned}\pi_1(p) &= \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}} \\ \pi_1\left(\frac{1}{2N}\right) &= \frac{1 - e^{-s}}{1 - e^{-2Ns}} = \frac{1 - (1 - s + \frac{s^2}{2!} - \frac{s^3}{3!} + \dots)}{1 - e^{-2Ns}} \approx \frac{s}{1 - e^{-2Ns}} \approx s \quad (\text{if } N \text{ is large}) \\ \rho &= 2Nu \cdot \pi_1\left(\frac{1}{2N}\right) \approx 2Nus\end{aligned}$$

- If A_2 is a new **deleterious** mutation with initial frequency $q = \frac{1}{2N}$, then the **fixation probability** of A_2 equals

$$\begin{aligned}\pi_2(q) &= 1 - \pi_1(1-q) = 1 - \left[\frac{1 - e^{-2Ns(1-q)}}{1 - e^{-2Ns}} \right] = 1 - \left[\frac{1 - e^{2Ns} \cdot e^{-2Ns}}{1 - e^{-2Ns}} \right] = \frac{1 - e^{-2Ns} - (1 - e^{2Ns} \cdot e^{-2Ns})}{1 - e^{-2Ns}} \\ &= \frac{-e^{-2Ns} + e^{2Ns} \cdot e^{-2Ns}}{1 - e^{-2Ns}} = \frac{e^{-2Ns} (e^{2Ns} - 1)}{1 - e^{-2Ns}} = \frac{e^{2Ns} - 1}{e^{2Ns} (1 - e^{-2Ns})} = \frac{e^{2Ns} - 1}{e^{2Ns} - 1}\end{aligned}$$

If $2Ns \gg 1$ and $q = \frac{1}{2N}$, then $\pi_2\left(\frac{1}{2N}\right) \approx 0$; however, if $2Ns \leq 1$, then there is a non-trivial probability that deleterious allele A_2 will fix in the population.

- Note that the equation

$$\pi_1(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}$$

does **not** imply that

$$\pi_1(q) = \frac{1 - e^{-2Nsq}}{1 - e^{-2Ns}}$$

since we are using the h, s model of selection with selection factor s , and this **assumes** genotype A_1A_1 has fitness 1, genotype A_1A_2 has fitness $1 - hs = 1 - \frac{s}{2}$ (since $h = \frac{1}{2}$ was assumed), and that genotype A_2A_2 has fitness $1 - s$. This is the reason for the slightly convoluted computation above.

Genetic draft and pseudo-hitchhiking model

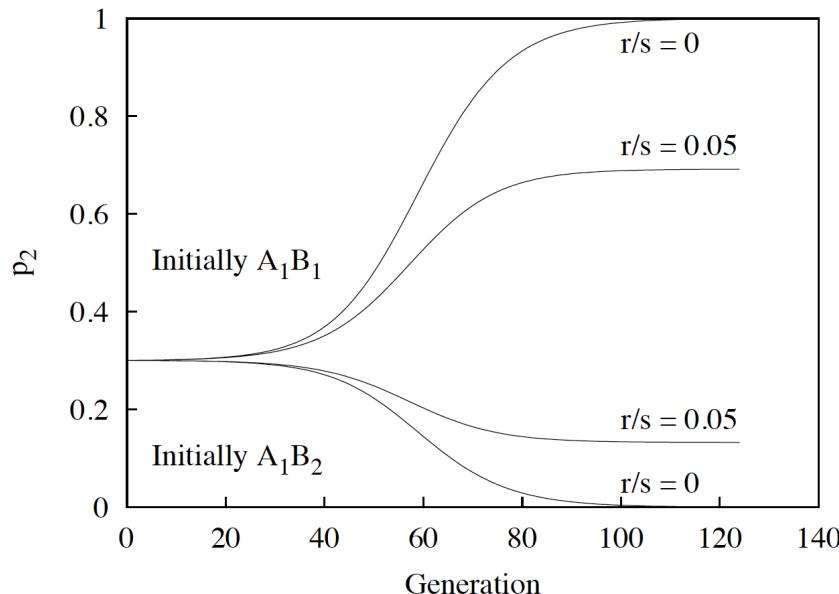


Figure 4.5: The frequency of the B_2 allele under different hitchhiking scenarios. For the upper two curves, the A_1 allele is initially linked to the B_1 allele; in the bottom two, it is linked to the B_2 allele. $s = 0.2$ for all trajectories.

- Genetic draft involves an initial **stochastic** element, involving the placement of new **advantageous** allele A_1 near a B_1 allele, or instead near a B_2 allele, subsequently followed by a **deterministic** process, whereby over generations the B_1 allele reaches an equilibrium frequency (as shown in the figure above).
- Our goal in this section is to provide a simplistic model for genetic draft, where it is assumed that the **time** to reach equilibrium frequency for B_1 is much **shorter** than the time for recombination to occur. For reasons of simplicity, assume that equilibrium is reached within **one generation**.

- If a recombination event between loci A and B does **not occur** ($r = 0$), then the hitchhiking allele of the B locus (either B_1 or B_2) will not be separated from the advantageous A_1 allele. In the absence of drift, A_1 will sweep through the population, reaching equilibrium frequency $p_1 = 1$. It follows that the hitchhiking allele of B will reach equilibrium frequency of 1 (if B_1 hitchhikes) or 0 (if B_2 hitchhikes).
- Let X be the **indicator** random variable whose value is the **equilibrium frequency** p_2 of allele B_1 . Since the probability that new mutation A_1 with initial frequency $\frac{1}{2N}$ will lie next to B_1 equals p_2 , we have the following:

$$X = \begin{cases} 1 & \text{with probability } p_2 \\ 0 & \text{with probability } q_2 = 1 - p_2 \end{cases}$$

The expected value and variance of X is well-known (as this corresponds to a coin flip with heads probability p_2), and is easily computed.

$$\begin{aligned} E[X] &= 1 \cdot p_2 + 0 \cdot (1 - p_2) = p_2 \\ E[X^2] &= 1^2 \cdot p_2 + 0^2 \cdot (1 - p_2) = p_2^2 \\ V[X] &= E[X^2] - (E[X])^2 = p_2 - p_2^2 = p_2(1 - p_2) = p_2q_2 \end{aligned}$$

- Now consider a model that **implicitly** accounts for **genetic drift**, by introducing the probability ρ that a new mutation A_1 is not lost due to genetic drift (recall that ρ is the probability of a new **substitution** A_1 , where ρ equals mutation rate u times probability of **fixation** in the population – i.e. not being lost due to genetic drift). Such substitutions are detectable in multiple sequence alignments.

$$X = \begin{cases} p_2 & \text{with probability } 1 - \rho \\ 1 & \text{with probability } \rho p_2 \\ 0 & \text{with probability } \rho q_2 = \rho(1 - p_2) \end{cases}$$

The new mutation A_1 is lost due to genetic drift with probability $1 - \rho$, in which case the B_1 allele frequency never changes from its initial value of p_2 (Hardy-Wright implies that allele frequencies never change in absence of drift and selection). On the other hand, if mutation A_1 is not lost due to drift, then it will ultimately reach equilibrium frequency of 1, in which case the hitchhiking allele of the B locus will also reach equilibrium frequency of 1.

- We compute the expected value and variance of X as follows.

$$E[X] = p_2 \cdot (1 - \rho) + 1 \cdot \rho p_2 + 0 \cdot \rho q_2 = p_2$$

$$E[X^2] = p_2^2 \cdot (1 - \rho) + 1^2 \cdot \rho p_2 + 0^2 \cdot \rho q_2 = p_2^2 - p_2^2 \rho + \rho p_2$$

$$V[X] = E[X^2] - (E[X])^2 = p_2^2 - p_2^2 \rho + \rho p_2 - p_2^2 = p_2 q_2 \rho$$

- On page 5 of the Probability Appendix, it is shown that the Wright-Fisher model has variance

$$\frac{pq}{2N}$$

- For a given population genetics model (Wright-Fisher, Moran, Avise, etc.), the **variance effective population size** N_e is defined to be that value such that $\frac{pq}{2N_e}$ is equal to the variance of the model.
- Letting $p = p_2$ and $q = q_2$, the previous model (with substitution probability of ρ for A_1) has variance $pq\rho$, so

$$\frac{pq}{2N_e} = pq\rho$$

and hence the **variance effective population size** N_e is given by

$$N_e = \frac{1}{2\rho}$$

- The final **pseudo-hitchhiking** model **implicitly** accounts for **genetic drift**, was well as **recombination**, which latter separates the hitchhiking allele of B locus from the advantageous mutation A_1 **before** the hitchhiking B -allele reaches frequency 1, instead reaching a frequency of **only** $y \in [0, 1]$.
- Suppose A_1 is a new, advantageous mutation, which “lands” near B_1 with probability p_2 , and “lands” near B_2 with probability $q_2 = 1 - p_2$. The frequency of that copy will increase from $\frac{1}{2N}$ to some asymptotic value, y , which depends on recombination rate r and selection factor s . The figure on page 38 of Chapter 4 notes illustrates the situation for the (partial) fixation of B_1 which hitchhikes on for the following parameters: recombination rate $r = 0.01$, selection coefficient $s = 0.2$, population size $N = 200$, and asymptotic value $y = 0.9114$ if a single new mutation A_1 is adjacent to B_1 .
- If A_1 “lands” near the B_1 allele in a single gamete, then proportion $p_2 - \frac{1}{2N}$ of the $2N$ gametes contain A_2B_1 , while proportion $1 - p_2 - \frac{1}{2N} = q_2 - \frac{1}{2N}$ of the $2N$ gameters contain A_2B_2 . Since the pseudohitchhiking model **implicitly** accounts for genetic drift, the $1 - \frac{1}{2N}$ initial gametes that contain allele A_2 are **reduced** by factor $(1 - y)$. This explains the **pseudo-hitchhiking** model:

$$X = \begin{cases} p_2 & \text{with probability } 1 - \rho \\ p_2(1 - y) + y & \text{with probability } \rho p_2 \\ p_2(1 - y) & \text{with probability } \rho q_2 = \rho(1 - p_2) \end{cases}$$

With probability $1 - \rho$, there is no new, advantageous **substitution** A_1 , in which case Hardy-Wright implies that the initial B_1 frequency of p_2 never changes. The term $p_2(1 - y) + y$ represents the proportion y of B_1 alleles that hitchhike with advantageous A_1 plus the proportion $p_2(1 - y)$ of non-hitchhiking B_1 alleles. Similarly, if new mutation A_1 “lands” next to a B_2 allele, then B_1 frequency becomes $p_2(1 - y)$.

- We compute the mean and variance of the **pseudo-hitchhiking** model.

$$\begin{aligned}
 E[X] &= (1 - \rho)p_2 + [p_2(1 - y) + y]\rho p_2 + [p_2(1 - y)]\rho q_2 \\
 &= p_2 - p_2\rho + p_2\rho[p_2 - p_2y + y] + q_2\rho[p_2 - p_2y] \\
 &= p_2 - p_2\rho + p_2^2\rho - p_2^2y\rho + p_2\rho y + p_2\rho q_2 - p_2q_2\rho y \\
 &= p_2 - p_2\rho + p_2^2\rho + p_2^2\rho y + p_2\rho y + p_2\rho - p_2^2\rho - p_2\rho y + p_2^2\rho y \\
 &= p_2
 \end{aligned}$$

$$\begin{aligned}
 E[X^2] &= p_2^2(1 - \rho) + [p_2^2(1 - y)^2 + y^2 + 2p_2y(1 - y)]\rho p_2 + p_2^2(1 - y^2)\rho - p_2^3(1 - y)^2\rho \\
 &= p_2^2 - p_2^2\rho + \rho p_2^3(1 - y)^2 + \rho p_2y^2 + 2\rho p_2^2y - 2\rho p_2^2y^2 + p_2^2\rho - 2p_2^2\rho y + p_2^2y^2\rho - [p_2^3\rho - 2p_2^3\rho y + p_2^3y^2\rho] \\
 &= p_2^2 - p_2^2\rho + \rho p_2^3 - 2y\rho p_2^3 + \rho p_2^3y^2 + \rho p_2y^2 + 2\rho p_2^2y - 2\rho p_2^2y^2 + p_2^2\rho - 2p_2^2\rho y + p_2^2y^2\rho - p_2^3\rho + 2p_2^3\rho y - p_2^3y^2\rho \\
 &= p_2^2 + \rho p_2y^2 - \rho p_2^2y^2
 \end{aligned}$$

$$\begin{aligned}
 V[X] &= p_2^2\rho p_2y^2 - \rho p_2^2y^2 - p_2^2 = \rho p_2y^2(1 - p_2) \\
 &= \rho p_2q_2y^2
 \end{aligned}$$

- Since the Wright-Fisher model has variance

$$\frac{pq}{2N}$$

and the **pseudo-hitchhiking** model has variance $\rho p_2 q_2 y^2$, we have

$$\frac{pq}{2N_e} = pqy^2 \rho$$

$$N_e = \frac{1}{2y^2 \rho}$$

- Assuming that genetic drift and genetic draft (hitchhiking) are independent events, each contributing equally to genome evolution, it follows that total variance is additive, so

$$V[X_1] = pq\rho y^2 \quad (\text{variance due to genetic draft})$$

$$V[X_2] = \frac{pq}{2N} \quad (\text{variance due to genetic drift})$$

$$V[X] = V[X_1] + V[X_2] = pq \left(\rho E[y^2] + \frac{1}{2N} \right)$$

$$\frac{pq}{2N_e} = pq \left(\rho E[y^2] + \frac{1}{2N} \right)$$

$$N_e = \frac{1}{2 \left(\rho E[y^2] + \frac{1}{2N} \right)} = \frac{1}{2 \left(\frac{2N\rho E[y^2]+1}{2N} \right)} = \frac{N}{2N\rho E[y^2] + 1}$$

$$\lim_{N \rightarrow \infty} N_e = \frac{1}{2\rho E[y^2]}$$

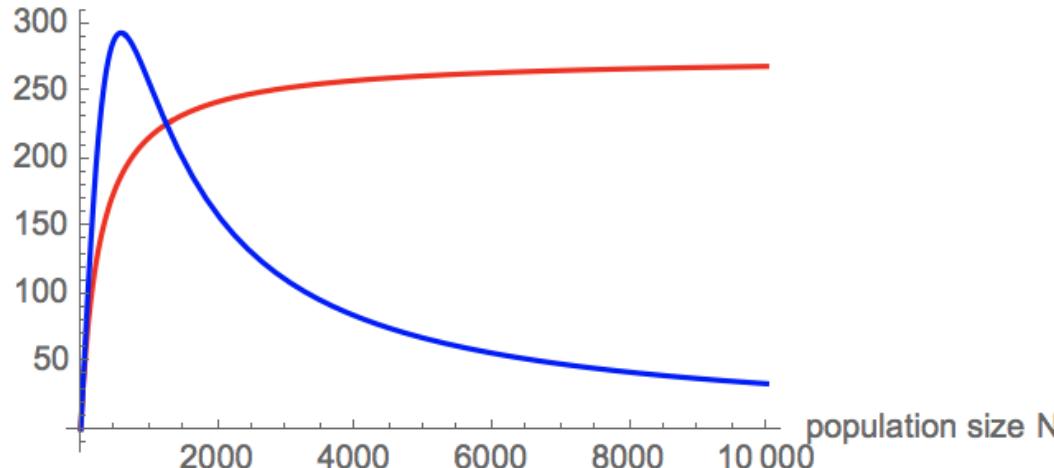
If new mutation A_1 were **neutral**, then its substitution rate $\rho = u$; however, A_1 is assumed to be **advantageous**, so its substitution rate is

$$\rho \approx 2Nus$$

where u is mutation rate, and s is selection factor. Next page shows a graph of effective population size N_e under both scenarios.

Effective population size

Effective population size plotted as function of N
/e population size N_e



- Reproduction of plot similar to Figure 4.6, generated by Mathematica code on following page. The **effective population size** N_e is plotted as a function of (real) population N , where

$$N_e = \frac{N}{2N\rho E[y^2] + 1}$$

- Two curves are shown: **constant** $\rho = 0.0025$, and the **substitution rate of advantageous mutations** A_1 is given by the function

$$\rho(N) = 2Nus$$

where N is (real) population size, $u = 0.00001$ is mutation rate, selection factor $s = 0.1$, and final frequency y of the hitchhiking allele (B_1 if gamete is A_1B_1 , B_2 if gamete is A_1B_2).

Mathematica code for previous figure

```
rho0 = 0.0025 (* hitchhiking constant rho *)
rho[n_] := 2 n u s;
y = 0.85;
s = 0.1;
u = 0.00001;
maxN = 10 000;
F[n_] := n / (1 + 2 n rho0 y^2);
G[n_] := n / (1 + 2 n rho[n] y^2);
g1 = Plot[F[n], {n, 1, maxN},
  AxesLabel → {"population size N", "effective population size Ne"},
  PlotLabel → "Effective population size plotted as function of N", PlotStyle → Red,
  PlotRange → All];
g2 = Plot[G[n], {n, 1, maxN},
  AxesLabel → {"population size N", "effective population size Ne"},
  PlotLabel → "Effective population size plotted as function of N",
  PlotStyle → Blue, PlotRange → All];
Show[g1, g2]
```

Genetic draft better at explaining why heterozygosity appears independent of population size

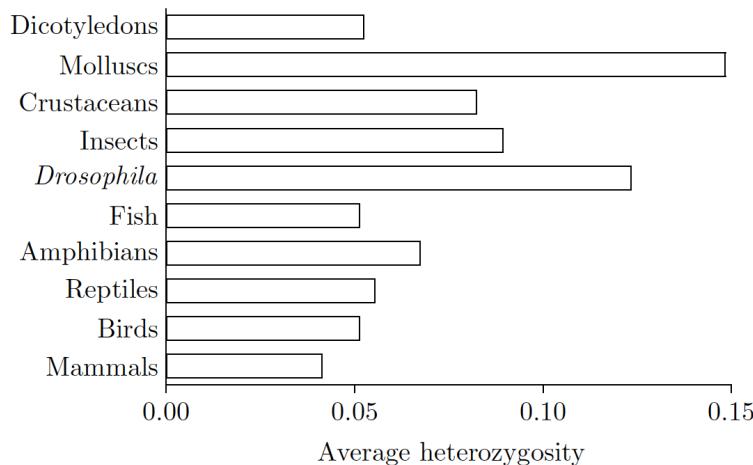


Figure 1.4: Estimates of average protein heterozygosities based on electrophoretic studies plotted from the data in Nevo et al. (1984).

- On pages 17-18 of Chapter 2 notes, we showed that under **genetic drift**, equilibrium **heterozygosity** \mathcal{H}^* satisfies

$$\mathcal{H}^* = \frac{4Nu}{1 + 4Nu}$$

and so clearly $\lim_{N \rightarrow \infty} \mathcal{H}^* = 1$. It follows that under **genetic drift**, the **larger** the population size, the **more** heterozygosity, ultimately approaching the limiting value of 1, the maximum possible heterozygosity value. In the bi-allelic case maximum heterozygosity of 1/2 occurs when $p = 1/2 = q$, so here we are considering the case of mutation at various sites in the coding region of a protein (infinite sites model).

- However, this does not agree with heterozygosities determined in populations of a number of species, as shown in the figure above. In real populations, **heterozygosity** values appear to be largely **independent** of population size, and are not close to the predicted value of 1, even for enormous populations – on page 39 of Gillespie’s text, the effective population sizes of some species, like *Drosophila*, might well be greater than 10^{10} .

- Under **genetic draft**, the **effective population size** N_e satisfies

$$N_e = \frac{N}{1 + 2N\rho E[y^2]}$$

which **decreases** as a function of (real) population size N . In the formula for \mathcal{H}^* above, replacing population size N by **effective population size** N_e under **genetic draft**, we have

$$\begin{aligned}\mathcal{H}^* &= \frac{4N_e u}{1 + 4N_e u} \\ &= \frac{4Nu}{1 + 2N\rho E[y^2]} \cdot \left[\frac{1}{1 + \frac{4Nu}{1+2N\rho E[y^2]}} \right] = \frac{4Nu}{1 + 2N\rho E[y^2]} \cdot \left[\frac{1 + 2N\rho E[y^2]}{(1 + 2N\rho E[y^2]) + 4Nu} \right] \\ &= \frac{4Nu}{1 + 2N\rho E[y^2] + 4Nu}\end{aligned}$$

- Clearly, under **genetic draft** (hitchhiking)

$$\lim_{N \rightarrow \infty} \mathcal{H}^* = \lim_{N \rightarrow \infty} \frac{4Nu}{1 + 2N\rho E[y^2] + 4Nu} = \lim_{N \rightarrow \infty} \frac{4Nu}{2N\rho E[y^2] + 4Nu} = \frac{2u}{\rho E[y^2] + 2u}$$

- For instance, assuming u is 10^{-8} per nucleotide per year, substitution rate ρ is equal to u (which is the case for neutral sites, as shown on page 41 of Chapter 4 notes), y is 0.8, and that $N = 10^{10}$, the predicted heterozygosity is

$$\mathcal{H}^* = \frac{4Nu}{1 + 4Nu} = 0.997506234414 \approx 0.9975$$

under genetic drift, while the predicted heterozygosity is

$$\mathcal{H}^* = \frac{4Nu}{1 + 4Nu + 4N\rho y^2} = 0.756143667297 \approx 0.7561$$

under genetic draft. Neither of these values agrees well with the value of ≈ 0.13 from the figure above for *Drosophila*; however, according to genetic draft does suggest (in the words of Gillespie) “a solution to the neutral theory’s problem that levels of variation in natural populations are remarkably similar across species in spite of great differences in population sizes”.

- What is meant is that **neutral theory** under **genetic drift** cannot easily explain the similar levels of diversity across different species, with different generation times and different (large) population sizes, while this is more easily explainable under **genetic draft**. For this reason, Gillespie argues that genetic draft may be a more important stochastic force than genetic drift.



Chapter 5: Nonrandom Mating

Nonrandom mating due to inbreeding and subdivision

- **inbreeding** increases population **homozygosity** and causes decreased fitness leading to inbreeding depression. Inbreeding may be due to reasons having to do with
 - ▷ royalty, caste system, religion, race, etc.
- **restricted migration** increases population **homozygosity** and causes decreased fitness (Wahlund's effect)
 - ▷ remote island populations (Iceland, New Guinea, Solomon Islands)
 - ▷ remote mountains or challenging climate (Nepal, Arctic, Andes Mountains)
- how to quantify increase of homozygosity due to nonrandom mating – either assortive mating or due to Wahlund's effect:
 - ▷ define F to be the probability that presence of allele A_1 is due to special conditions (inbreeding or subdivision)
 - ▷ $F = F_I$ due to inbreeding, and $F = F_{ST}$ due to Wahlund's effect
 - ▷ F_I is the probability that the two alleles (one on each chromosome of a chromosomal pair) of an individual at a given locus are **identical by descent**
 - ▷ add pqF to both homozygote frequencies A_1A_1 and A_2A_2 , and subtract $2pqF$ from heterozygote frequency A_1A_2

$$p^2 + pqF = p^2 + p(1 - p)F = p^2 - p^2F + pF = p^2(1 - F) + pF$$

$$q^2 + pqF = q^2 + q(1 - q)F = q^2 - q^2F + qF = q^2(1 - F) + qF$$

$$2pq - 2pqF = 2pq(1 - F)$$

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|--|-------------------|--------------|-------------------|
| frequency for random mating | p^2 | $2pq$ | q^2 |
| frequency for non-random mating | $p^2(1 - F) + pF$ | $2pq(1 - F)$ | $q^2(1 - F) + qF$ |

- Note that $\text{Prob}[A_1A_1] = p^2(1 - F) + pF$ can be explained as follows:
 - draw first allele randomly (with replacement): $\text{Prob}[A_1] = p$
 - second allele allele is A_1 due to one of the following circumstances
 - special conditions (inbreeding or geographic isolation) which occur with probability F
 - due to no special conditions, the second allele is A_1 , with probability $(1 - F)p$
- $\text{Prob}[A_1A_1|A_1] = F + (1 - F)p$
- $$\text{Prob}[A_1A_1] = \text{Prob}[A_1] \cdot \text{Prob}[A_1A_1|A_1] = p(F + (1 - F)p) = p^2(1 - F) + pF$$

Similar argument shows $\text{Prob}[A_2A_2] = q^2(1 - F) + qF$, and since the sum of genotype frequencies for A_1A_1 , A_1A_2 , A_2A_2 equals 1, we have

$$2pq(1 - F) = 1 - [pF + p^2(1 - F)] - [qF + q^2(1 - F)]$$

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|----------------------------------|-------------------|--------------|-------------------|
| frequency (HW) for random mating | p^2 | $2pq$ | q^2 |
| frequency for nonrandom mating | $p^2(1 - F) + pF$ | $2pq(1 - F)$ | $q^2(1 - F) + qF$ |

- Example:** Find p and F for a population in which the genotype frequencies of A_1A_1 , A_1A_2 , A_2A_2 are 0.377, 0.285, and 0.338, respectively.

```

:= ClearAll;
q = 1 - p;
eq1 = p^2 (1 - F) + p F == 0.377;
eq2 = 2 p q (1 - F) == 0.285;
eq3 = q^2 (1 - F) + q F == 0.338;
NSolve[{eq1, eq2, eq3}, {p, F}]

]:= {{p → 0.5195, F → 0.429132}}

```

Inbreeding



- Charles V (1500-1558), Holy Roman Emperor, from 1519 until 1556 (painting ca. 1515)
- **Inbreeding depression** is the reduced fitness in a given population as a result of inbreeding
- recall from page 43 of Chapter 3 notes that
$$\text{inbreeding depression} \Leftrightarrow \text{outbred } \bar{w} > \text{inbred } \bar{w} \Leftrightarrow 1 - 2pqhs - q^2s > 1 - qs \Leftrightarrow h < \frac{1}{2}$$
- Due to inbreeding among royalty, the facial bone structure of Emporer Charles V is sometimes believed to be an example of **inbreeding depression** – it is not, since the **Hapsburger jaw** is caused by a dominant allele.
- Frederick III (1463-1525), called *Friedrich der Weise*, Kurfürst von Sachsen (Elector from Saxony) from 1486 to 1525 – protected **Martin Luther** from Charles V turning him over to Church authorities to be burned as a heretic
- Image from <https://whyevolutionistrue.wordpress.com/2015/07/03/inbreeding-depression-in-man/>

Identity by descent (Gustave Malécot, 1911-1998)

- coefficient of kinship $f_{x,y} = \text{Prob}[X \equiv Y]$ where $X \equiv Y$ indicates that allele X from a fixed locus of individual x is **identical by descent** to allele Y from the same locus of individual y .
- Recall that two alleles are **identical by descent** if and only if they are copies of the **same** piece of DNA in a common ancestor.
- r_0, r_1, r_2 respectively denote the (relation) probability of sharing respectively 0,1 or 2 alleles between individuals x, y
- \bar{r} denotes the **expected number** of shared alleles at a given locus

$$\bar{r} = E[r] = 0 \cdot r_0 + 1 \cdot r_1 + 2 \cdot r_2$$

- coefficient of relatedness

$$r = \frac{\bar{r}}{2} = \frac{r_1}{2} + r_2$$

- coefficient of kinship

$$f_{x,y} = \frac{\bar{r}}{4} = \frac{r}{2}$$

| relationship | r_0 | r_1 | r_2 | avg \bar{r} | coeff relatedness r | coeff kinship $f_{x,y}$ |
|------------------|---------------|---------------|---------------|---------------|-----------------------|-------------------------|
| parent-offspring | 0 | 1 | 0 | 1 | $\frac{1}{2}$ | $\frac{1}{4}$ |
| full sibs | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | 1 | $\frac{1}{2}$ | $\frac{1}{4}$ |
| half sibs | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ |
| first cousins | $\frac{3}{4}$ | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ |

Some calculations

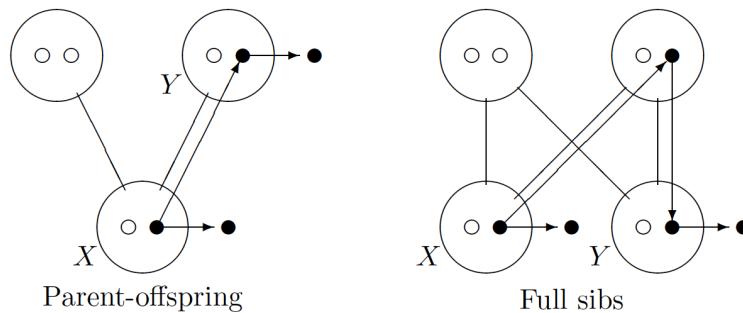


Figure 5.1: Two pedigrees used in the text to illustrate the calculation of the coefficient of kinship.

- Unless specifically stated, all computations assume the genes are located on **autosomes**.
- X is offspring, Y is Mom (or Dad).

$$f_{x,y} = \text{Prob}[a \equiv a' | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

$$\text{Prob}[a \text{ inherited from } Y] = \frac{1}{2}$$

$$\text{Prob}[a \equiv a' | a' \in \{a'_1, a'_2\} \text{ in } Y] = \frac{1}{2}$$

$$f_{x,y} = \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = \frac{1}{4}$$

$$r_0 = 0$$

$$r_1 = 1$$

$$r_2 = 0$$

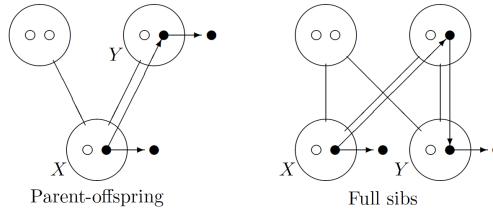


Figure 5.1: Two pedigrees used in the text to illustrate the calculation of the coefficient of kinship.

- **X and Y are full siblings**

$$f_{x,y} = \text{Prob}[a \equiv a' | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

$$\frac{1}{2} = \text{Prob}[a, a' \text{ come from same parent} | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

$$\frac{1}{2} = \text{Prob}[a \equiv a' | a, a' \text{ come from same parent}]$$

$$f_{x,y} = \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = \frac{1}{4}$$

$$\frac{1}{2} = \text{Prob}[X's \text{ maternal allele} \not\equiv \text{to } Y's \text{ maternal allele}]$$

$$\frac{1}{2} = \text{Prob}[X's \text{ paternal allele} \not\equiv \text{to } Y's \text{ paternal allele}]$$

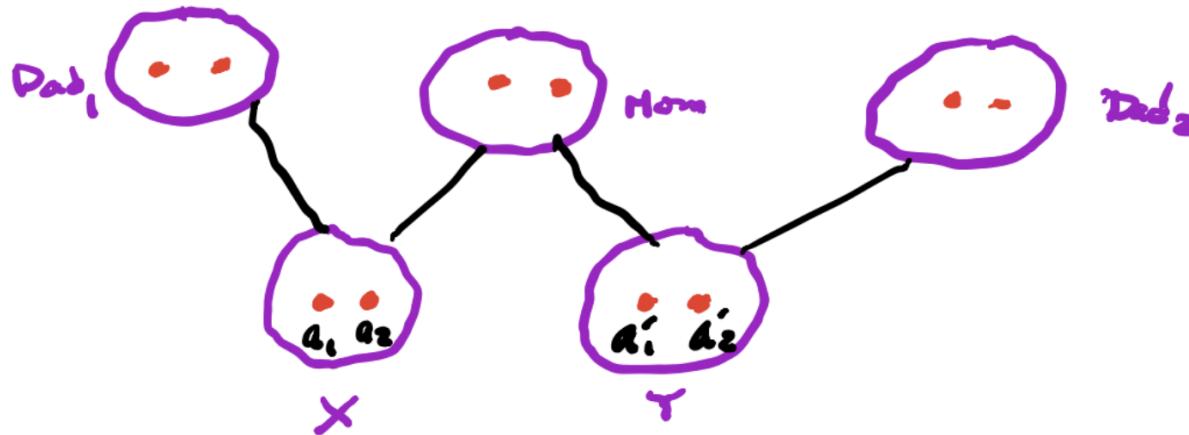
$$\frac{1}{4} = \text{Prob}[X's \text{ maternal allele} \not\equiv \text{to } Y's \text{ maternal allele}] \cdot \text{Prob}[X's \text{ paternal allele} \not\equiv \text{to } Y's \text{ paternal allele}]$$

$$r_0 = \frac{1}{4}$$

$$\frac{1}{4} = \text{Prob}[X's \text{ maternal allele} \equiv \text{to } Y's \text{ maternal allele}] \cdot \text{Prob}[X's \text{ paternal allele} \equiv \text{to } Y's \text{ paternal allele}]$$

$$r_2 = \frac{1}{4}$$

$$r_1 = 1 - r_0 - r_2 = \frac{1}{2}$$



- X and Y are **half siblings** – for specificity, suppose X, Y have same mother but different fathers

$$f_{x,y} = \text{Prob}[a \equiv a' | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

$$\frac{1}{2} = \text{Prob}[a \text{ comes from Mom} | a \in \{a_1, a_2\} \text{ in } X]$$

$$\frac{1}{2} = \text{Prob}[a' \text{ comes from Mom} | a' \in \{a'_1, a'_2\} \text{ in } Y]$$

$$\frac{1}{4} = \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = \text{Prob}[a, a' \text{ come from Mom} | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

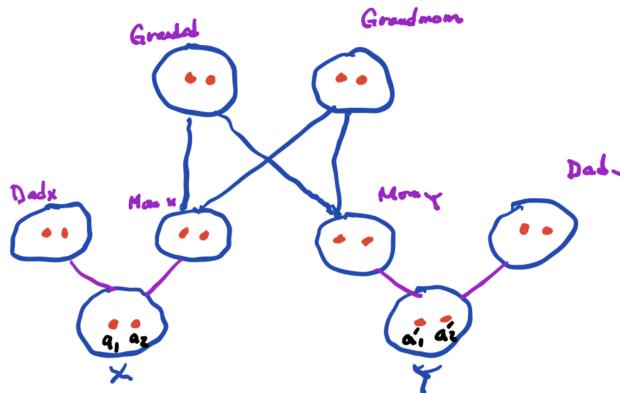
$$\frac{1}{2} = \text{Prob}[a \equiv a' | a, a' \text{ come from Mom}]$$

$$f_{x,y} = \left(\frac{1}{2}\right) \cdot \left(\frac{1}{4}\right) = \frac{1}{8}$$

$r_2 = 0$ since X, Y have different fathers

$r_1 = \frac{1}{2}$ since probability that X, Y choose same maternal allele is $1/2$

$$r_0 = 1 - r_1 - r_2 = \frac{1}{2}$$



- X and Y are **first cousins** – for specificity, suppose the mother of X is the sister of the mother of Y
- Select $a \in \{a_1, a_2\}$ from X , and $a' \in \{a'_1, a'_2\}$ from Y .

$$f_{x,y} = \text{Prob}[a \equiv a' | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

$$\frac{1}{2} = \text{Prob}[a \text{ comes from Mom of } X]$$

$$\frac{1}{2} = \text{Prob}[a' \text{ comes from Mom of } Y]$$

$$\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \text{Prob}[a \text{ comes from Mom of } X] \cdot \text{Prob}[a' \text{ comes from Mom of } Y]$$

$= \text{Prob}[a, a' \text{ are alleles found in the 4 alleles of Grandpa and Grandma}]$

$$\frac{1}{4} = \text{Prob}[a \equiv a' | a, a' \text{ are two randomly chosen alleles (with replacement) from alleles of Grandpa and Grandma}]$$

$$f_{x,y} = \left(\frac{1}{4}\right) \cdot \left(\frac{1}{4}\right) = \frac{1}{16}$$

$r_2 = 0$ since paternal allele of $X \not\equiv$ paternal allele of Y , as their fathers are unrelated

$r_1 = \frac{1}{4}$ since probability $f_{x,y} = \frac{1}{4}$ that two sisters (the mothers of X, Y) share at least one allele

$$r_0 = 1 - r_1 - r_2 = \frac{3}{4}$$

General recursive method to compute $f_{x,y}$

- Recall that
 - ▷ F_I is the probability that the two alleles (one on each chromosome of a chromosomal pair) of an individual at a given locus are **identical by descent**
 - ▷ the **coefficient of kinship** $f_{x,y} = f_{y,x}$ and is defined by

$$f_{x,y} = \text{Prob}[a \equiv a' | a \in \{a_1, a_2\} \text{ in } X, a' \in \{a'_1, a'_2\} \text{ in } Y]$$

- Let d, m denote the **dad** resp. **mom** of x . Then

$$f_{x,x} = \frac{1}{2} (1 + f_{d,m}) = \frac{1 + F_I}{2}$$

which is the probability that two randomly chosen alleles at the same locus of individual x (selected **with replacement**) are **identical by descent**.

- Let x, y be any two individuals, where x **is not an ancestor of** y , and let d, m denote the **dad** resp. **mom** of x . Then

$$f_{x,y} = \begin{cases} \frac{1+F_I}{2} & \text{if } x = y \\ \frac{1}{2}(f_{d,y} + f_{m,y}) & \text{if } x \neq y \end{cases}$$

- This elementary recursion was first described in

“A recursive algorithm for the calculation of identity coefficients”, by G. Karigl
Annals of Human Genetics, vol. 45, no. 3, pp. 299–305 (1981)

- We now use the recursive method to compute **coefficient of kinship** $f_{x,y}$ in the following table:

| parent-offspring | full sibs | half sibs | first cousins |
|------------------|---------------|---------------|----------------|
| $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ |

- parent-offspring:** x is child of mom m and dad d , and assume that inbreeding coefficient F_I of m is 0, and that the coefficient of kinship $f_{m,d}$ is 0. Then

$$f_{m,m} = \frac{1 + F_I}{2} = \frac{1}{2}$$

$$f_{x,m} = \frac{1}{2} (f_{m,m} + f_{m,d}) = \frac{1}{2} \left(\frac{1}{2} + 0 \right) = \frac{1}{4}$$

- full sibs:** x and y are offspring of mom m and dad d . Then

$$f_{x,y} = \frac{1}{2} (f_{m,y} + f_{d,y}) = \frac{1}{2} (f_{y,m} + f_{y,d}) = \frac{1}{2} \left(\frac{1}{4} + \frac{1}{4} \right) = \frac{1}{4}$$

- half sibs:** x and y have same mom m , but x has dad d_1 and y has dad d_2 , and assume there is no relation between y and d_1 , hence $f_{y,d_1} = 0$. Then

$$f_{x,y} = \frac{1}{2} (f_{m,y} + f_{d_1,y}) = \frac{1}{2} (f_{y,m} + f_{y,d_1}) = \frac{1}{2} \left(\frac{1}{4} + 0 \right) = \frac{1}{8}$$

- first cousins:** x has mom m_1 and dad d_1 , y has mom m_2 and dad d_2 , m_1 and m_2 are **sisters** (full sibs), and assume there is no relation between x and d_2 , and otherwise no relations, so $f_{x,d_2} = 0$, $f_{y,d_1} = 0$, $f_{m_1,d_2} = 0$, $f_{m_2,d_1} = 0$. Then

$$f_{x,y} = \frac{1}{2} (f_{m_1,y} + f_{d_1,y}) = \frac{1}{2} (f_{y,m_1} + f_{y,d_1}) = \frac{1}{2} (f_{y,m_1} + 0)$$

$$f_{y,m_1} = \frac{1}{2} (f_{m_2,m_1} + f_{d_2,m_1}) = \frac{1}{2} \left(\frac{1}{4} + f_{d_2,m_1} \right) = \frac{1}{2} \left(\frac{1}{4} + 0 \right) = \frac{1}{8}$$

$$f_{x,y} = \frac{1}{2} (f_{m_1,y} + f_{d_1,y}) = \frac{1}{2} \left(\frac{1}{8} + 0 \right) = \frac{1}{16}$$

Inbreeding coefficient F_I

- The **inbreeding coefficient** F_I is the probability that the two alleles (one on each chromosome of a chromosomal pair) of an individual at a given locus are **identical by descent**
- Wright formulated the genotype frequencies with F_I term in the following table:

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|--------------------------------|-----------------------|----------------|-----------------------|
| frequency for random mating | p^2 | $2pq$ | q^2 |
| frequency for inbreeding F_I | $p^2(1 - F_I) + pF_I$ | $2pq(1 - F_I)$ | $q^2(1 - F_I) + qF_I$ |
| fitness | 1 | 1-hs | 1-s |

where selection coefficient $0 \leq s \leq 1$, and the parameter for homozygous effect h satisfies

| heterozygosity effect | type of dominance | type of selection |
|-----------------------|--|--|
| $h = 0$ | A_1 dominant, A_2 recessive | directional evolution, converging to $p = 1$ |
| $h = 1$ | A_2 dominant, A_1 recessive | directional evolution, converging to $p = 1$ |
| $0 < h < 1$ | incomplete dominance | directional evolution, converging to $p = 1$ |
| $h < 0$ | overdominance (heterozygote is fittest) | balancing selection (e.g. sickle cell trait) |
| $h > 1$ | underdominance (homozygotes are fittest) | disruptive selection |

Mean fitness with inbreeding

- mean fitness \bar{w} satisfies

$$\begin{aligned}\bar{w} &= 1 \cdot (p^2(1 - F_I) + pF_I) + (1 - hs) \cdot (2pq(1 - F_I)) + (1 - s) \cdot (q^2(1 - F_I) + qF_I) \\&= p^2(1 - F_I) + pF_I + 2pq(1 - F_I) - 2pqhs(1 - F_I) + q^2(1 - F_I) + qF_I - s(q^2(1 - F_I) + qF_I) \\&= (p^2 + 2pq + q^2)(1 - F_I) + pF_I - 2pqhs(1 - F_I) + qF_I - sq^2(1 - F_I) - sqF_I \\&= (1 - F_I) + F_I(p + q) - 2pqhs + F_I(2pqhs) - sq^2 + sq^2F_I - sqF_I \\&= 1 - 2pqhs - sq^2 + F_I(2pqhs + sq^2 - sq) = 1 - (2pqhs + sq^2) + F_I(2pqhs + sq(q - 1)) \\&= 1 - (2pqhs + sq^2) + F_I(2pqhs - sq(1 - q)) = 1 - (2pqhs + sq^2) + F_I(2pqhs - spq)\end{aligned}$$

$$\bar{w} = 1 - a - bF_I$$

$$a = 2pqhs + sq^2$$

$$b = spq - 2pqhs = 2spq\left(\frac{1}{2} - h\right)$$

- In cases of **incomplete dominance**, where $0 < h < 1$, in real populations, **heterozygosity effect** $h < \frac{1}{2}$, since expected fitness \bar{w} of an outbred population is greater than the expected fitness \bar{w} of an inbred population exactly when $h < \frac{1}{2}$, as explained on page 43 of Chapter 3 notes. Thus in real populations, in most cases,

$$b = 2spq\left(\frac{1}{2} - h\right) > 0$$

the value of b is positive, so that expected fitness \bar{w} is a **decreasing linear** function of **inbreeding coefficient** F_I , resulting in **inbreeding depression**.

Estimation of overall viability for multiplicative epistasis

- Assuming **multiplicative epistasis**, the survival (overall fitness) S is the product of both genetic factors as well as of nongenetic factors (accidental death, natural catastrophes)

$$S = \prod_i (1 - a_i - b_i F_i) \cdot \prod_j (1 - x_j)$$

where x_j is the probability of the j th nongenetic factor causing death

- Recall that

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots$$

For $x > 0$, the error is in using a first order approximation $e^{-x} \approx 1 - x$ is

$$e^{-x} - (1 - x) = \frac{x^2}{2!} - \frac{x^3}{3!} + \dots < \frac{x^2}{2}$$

so if $x = 0.01$, the error is less than $\frac{0.0001}{2} = 0.00005$.

- Thus for x, a_i, b_i positive and close to 0, since $0 \leq F_I \leq 1$,

$$\begin{aligned}
S &= \prod_i (1 - a_i - b_i F_I) \cdot \prod_j (1 - x_j) \\
&\approx \prod_i e^{-a_i - b_i F_I} \cdot \prod_j e^{-x_j} \\
&= \exp\left(\sum_i -a_i\right) \cdot \exp\left(\sum_i -b_i F_I\right) \cdot \exp\left(\sum_j -x_j\right) \\
&= e^{-A - BF_I} \\
A &= \sum_i a_i + \sum_j x_j \\
B &= \sum_i b_i
\end{aligned}$$

- Recall that

$$\log(S) \approx -A - BF_I$$

Given values of S from a sample of individuals, the method of least squares (linear regression, supported in Excel, Mathematica, etc.) yields estimates of A, B .

- Using least squares estimates of A, B we see that survival (overall fitness) is a **linearly decreasing** function of the amount of inbreeding F_I

$$S = e^{-A - BF_I} \approx e^{-A} \cdot (1 - BF_I)$$

as shown in the figure on the next slide.

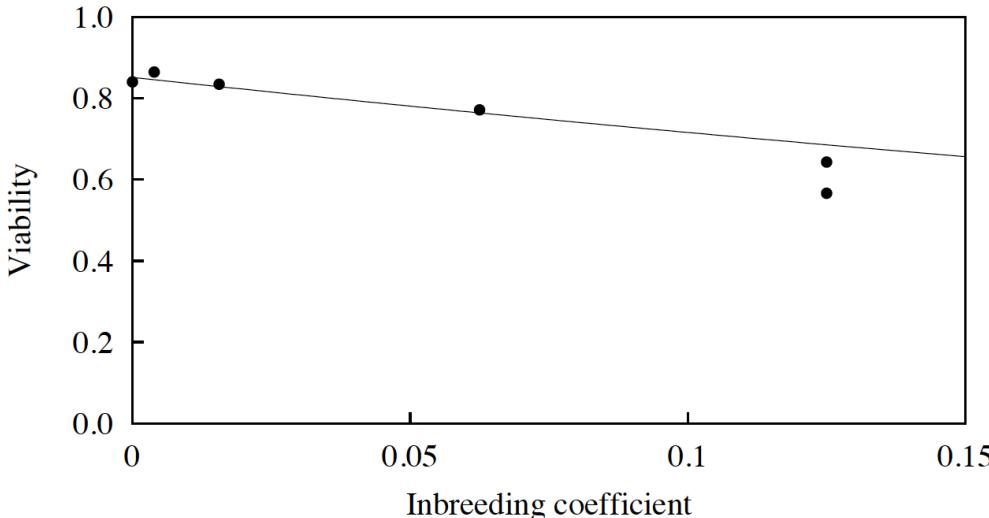
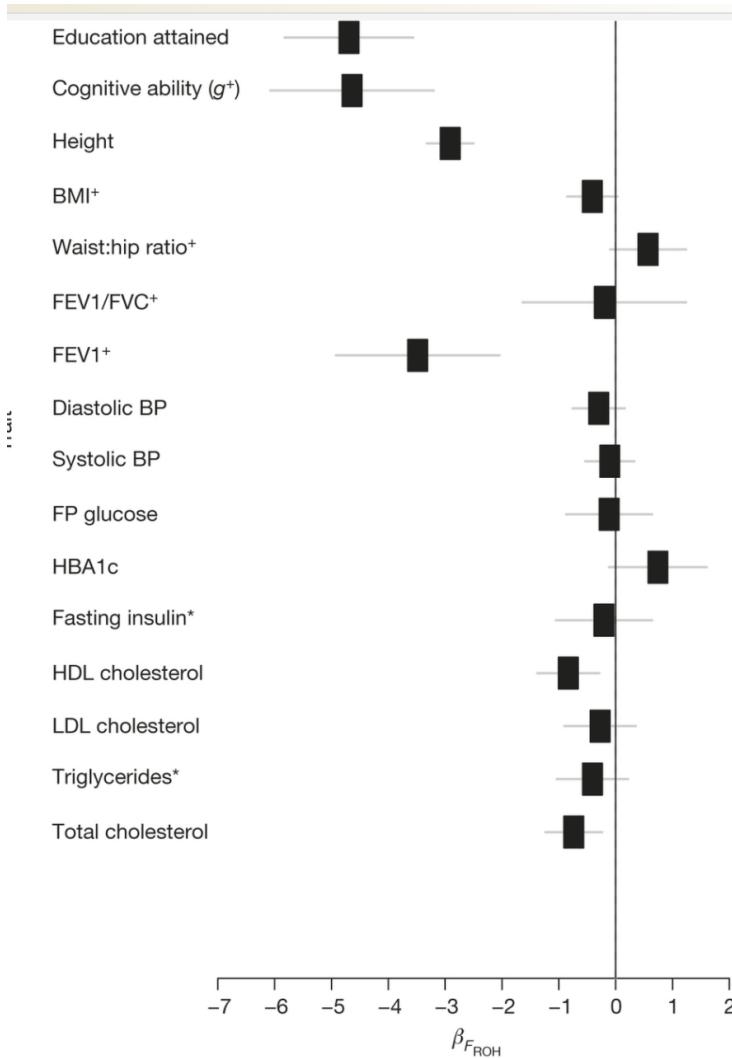


Figure 5.3: The viabilities of young children as a function of their inbreeding coefficients. The curve is the function $\exp(-A - BF_I)$ with $A = 0.1612$ and $B = 1.734$. The data are from Morton et al. (1956).

- Note that viability (fitness/survival) of individuals having highest inbreeding coefficient F_I is **lower** than the regression line, which suggests that defective alleles may act **synergistically**, rather than **independently**
- A **negative** slope in the graph of viability (fitness) as a function of **inbreeding coefficient** is a quantitative measure of **inbreeding depression**.



- The slopes m of regression equations $y = mx + b$ for the **fitness** of 16 phenotypic characters as a function of **inbreeding coefficient F_I** . Note that all slopes are near 0, except for those for **educational attainment, cognitive ability, height, and FEV1+**, where FEV1+ is a measure of the volume of air expelled by the lungs – these 4 slopes are **negative**, where educational attainment and cognitive ability both have a slope of -6 .
- Image from <https://whyevolutionistrue.wordpress.com/2015/07/03/inbreeding-depression-in-man/>

Self-fertilization (selfing)

- Self-fertilization, or **selfing**, commonly occurs in plants, which produce both male and female gametes.
- Assume that allele frequencies in the initial generation (**generation 0**) are p, q . In Chapter 1, we saw that in the absence of drift, mutation and selection, the genotype frequencies in generation 1 and all future generations are the following.

$$\text{Prob}[A_1A_1] = p^2$$

$$\text{Prob}[A_1A_2] = 2pq$$

$$\text{Prob}[A_2A_2] = q^2$$

- Homozygote genotype frequency for A_1A_1 [resp. A_2A_2] is p^2 [resp. q^2], since this is the probability of selecting two A_1 [resp. A_2] gametes – here, selection is **with replacement**, or equivalently in an infinite population.
- Heterozygote genotype frequency is $2pq$, since the probability of selecting an A_1 gamete followed by selecting an A_2 gamete is pq , while the probability of selecting an A_2 gamete followed by selecting an A_1 gamete is also pq .
- Clearly a selfing A_1A_1 [resp. A_2A_2] homozygote can **only** have genotype A_1A_1 [resp. A_2A_2].
- In a selfing A_1A_2 heterozygote, by Mendelian segregation, the A_1 gamete frequency **in that individual** is $\frac{1}{2}$, and similarly the A_2 gamete frequency **in that individual** is $\frac{1}{2}$. It follows that in a selfing heterozygote, $25\% = \frac{1}{2} \cdot \frac{1}{2}$ of the offspring have genotype A_1A_1 , similarly $25\% = \frac{1}{2} \cdot \frac{1}{2}$ of the offspring have genotype A_2A_2 , while $50\% = 2 \cdot \frac{1}{2} \cdot \frac{1}{2}$ of the offspring have genotype A_1A_2 .

- Thus, in the absence of mutation, selection, and recombination, **selfing reduces heterozygosity**, as shown in the following table of genotype frequencies for generations 0, 1, 2, ...

| generation \genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|----------------------|--|-----------------------|--|
| 0 | $a_{1,1}$ | $a_{1,2}$ | $a_{2,2}$ |
| 1 | $a_{1,1} + \frac{a_{1,2}}{4}$ | $\frac{a_{1,2}}{2}$ | $a_{2,2} + \frac{a_{1,2}}{4}$ |
| 2 | $a_{1,1} + \frac{a_{1,2}}{4} + \frac{a_{1,2}}{8}$ | $\frac{a_{1,2}}{4}$ | $a_{2,2} + \frac{a_{1,2}}{4} + \frac{a_{1,2}}{8}$ |
| : | : | : | : |
| n | $a_{1,1} + \frac{a_{1,2}}{2} \cdot \sum_{i=1}^n \frac{1}{2^i}$ | $\frac{a_{1,2}}{2^n}$ | $a_{2,2} + \frac{a_{1,2}}{2} \cdot \sum_{i=1}^n \frac{1}{2^i}$ |

- As generation $n \rightarrow \infty$, we have

$$Prob[A_1A_1] = a_{1,1} + \frac{a_{1,2}}{2}$$

$$Prob[A_1A_2] = 0$$

$$Prob[A_2A_2] = a_{2,2} + \frac{a_{1,2}}{2}$$

- If the initial genotype frequencies are at Hardy-Wright equilibrium, then we have the following table.

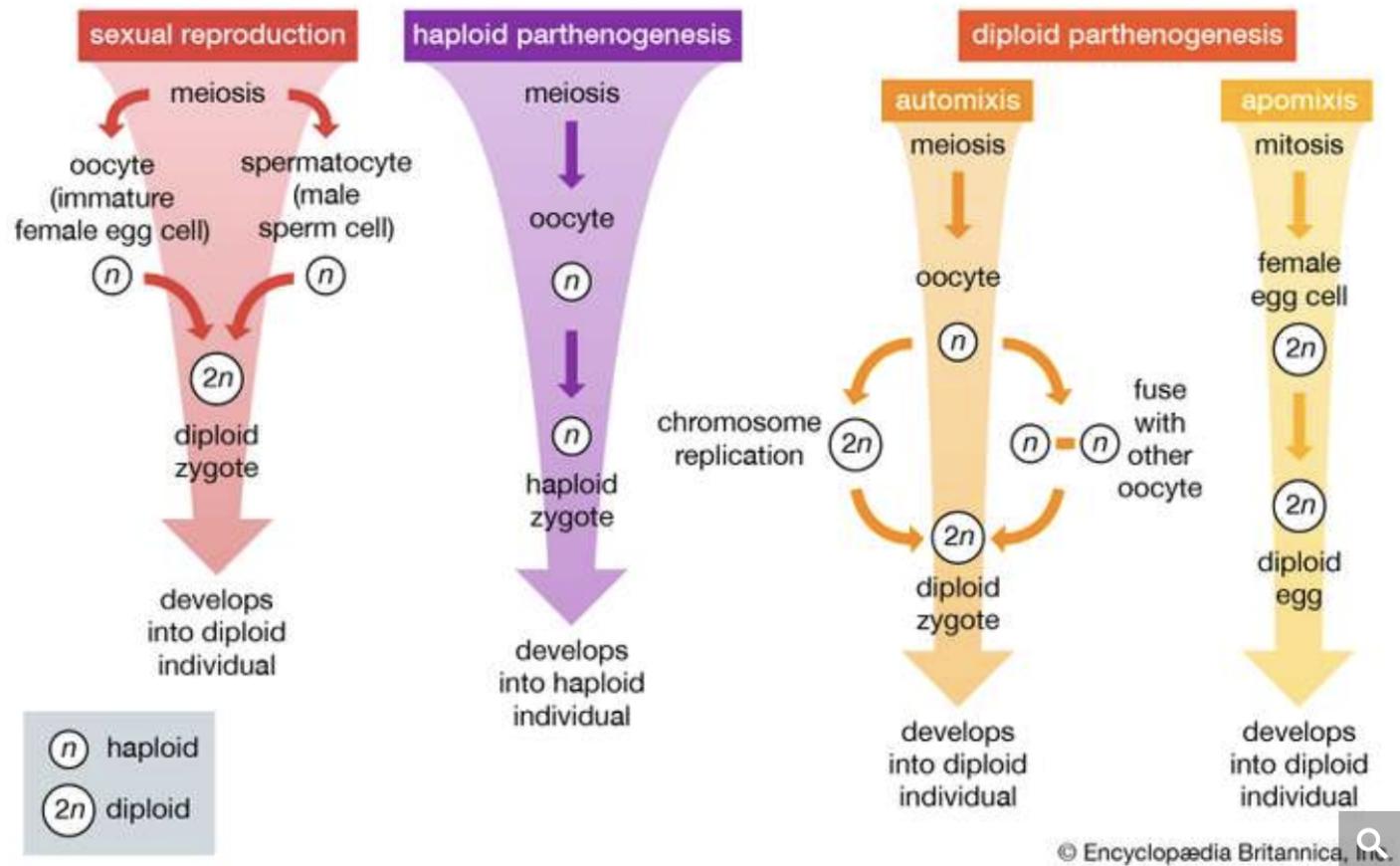
| generation \genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|----------------------|--|------------------|--|
| 0 | p^2 | $2pq$ | q^2 |
| 1 | $p^2 + \frac{2pq}{4}$ | $\frac{2pq}{2}$ | $q^2 + \frac{2pq}{4}$ |
| 2 | $p^2 + \frac{pq}{2} + \frac{pq}{4}$ | $\frac{pq}{2}$ | $q^2 + \frac{pq}{2} + \frac{pq}{4}$ |
| : | : | : | : |
| n | $p^2 + \frac{1}{2} \cdot \sum_{i=1}^n \frac{2pq}{2^i}$ | $\frac{pq}{2^n}$ | $q^2 + \frac{1}{2} \cdot \sum_{i=1}^n \frac{2pq}{2^i}$ |

- A_1 allele frequencies in successive generations satisfy the following:

$$\begin{aligned}
 p' &= \left[p^2 + \frac{pq}{2} \right] + \frac{pq}{2} = p^2 + pq = p(p+q) = p \\
 p'' &= \left[p^2 + \frac{pq}{2} + \frac{pq}{4} \right] + \frac{pq}{4} = p^2 + pq = p(p+q) = p \\
 &\vdots \quad \vdots \\
 p^{(n)} &= \left[p^2 + \frac{1}{2} \cdot \sum_{i=1}^n \frac{2pq}{2^i} \right] + \frac{2pq}{2^n} = p^2 + pq = p(p+q) = p
 \end{aligned}$$

- In the absence of selection, mutation and drift, **self-fertilization does not change allele frequencies**, but does reduce **heterozygosity** – indeed, in the n th generation, the heterozygosity is $\frac{2pq}{2^n}$.

The process of sexual reproduction versus several forms of parthenogenesis



- Parthenogenesis is asexual reproduction, in contrast to self-fertilization, which is sexual reproduction using sperm and egg cells from the same individual (image from Encyclopedia Britannica).

- Recall that **inbreeding coefficient** F_I is the probability that two alleles of an individual at a given locus are **identical by descent**
- In populations where mixed selfing and outcrossing occur (in the same individuals, as in plants), selfing changes F_I

α = probability of selfing

αF_I = probability of selfing and that the 2 alleles are identical by descent in the (selfing) parent

$1 - F_I$ = probability that two alleles of the (selfing) parent not identical by descent

$\alpha(1 - F_I) \cdot \frac{1}{2}$ = probability that selfing occurs, that two alleles of the (selfing) parent not identical by descent, and that alleles X, Y are copies of the same allele in (selfing) parent

$$F'_I = \alpha \left[F_I + \frac{1 - F_I}{2} \right]$$

- At **equilibrium**, the inbreeding coefficient F^* satisfies $F_I = F'_I$, so

$$\begin{aligned} F_I = F'_I &\Leftrightarrow F_I = \alpha \left[F_I + \frac{1 - F_I}{2} \right] \Leftrightarrow F_I - \left(\alpha F_I + \frac{\alpha(1 - F_I)}{2} \right) = 0 \\ &\Leftrightarrow F_I(1 - \alpha) - \frac{\alpha}{2} + \frac{\alpha F_I}{2} = 0 \Leftrightarrow F_I \left(1 - \alpha + \frac{\alpha}{2} \right) = \frac{\alpha}{2} \\ &\Leftrightarrow F_I \left(1 - \frac{\alpha}{2} \right) = \frac{\alpha}{2} \Leftrightarrow F_I = \frac{\frac{\alpha}{2}}{1 - \frac{\alpha}{2}} \Leftrightarrow F_I = \frac{\alpha}{2 - \alpha} \end{aligned}$$

- The incremental change in inbreeding coefficient satisfies

$$\begin{aligned} \Delta F_I = F'_I - F_I &= \alpha \left[F_I + \frac{1 - F_I}{2} \right] - F_I = \alpha F_I + \frac{\alpha}{2} - \frac{\alpha F_I}{2} - F_I \\ &= F_I \left(\alpha - \frac{\alpha}{2} - 1 \right) + \frac{\alpha}{2} = \frac{\alpha}{2} + F_I \left(\frac{\alpha}{2} - 1 \right) = \frac{\alpha}{2} - F_I \left(1 - \frac{\alpha}{2} \right) \\ &= \frac{\alpha}{2} - F_I \left(\frac{2 - \alpha}{2} \right) = \frac{2 - \alpha}{2} \left(\frac{\alpha}{2 - \alpha} - F_I \right) \end{aligned}$$

- **Equilibrium inbreeding coefficient:** $F_I^* = \frac{\alpha}{2-\alpha}$ obtained by solving equation $\Delta F_I = 0$, or as we did above by setting $F_I = F'_I$.
- Note that F_I **increases under selfing**, and **decreases under outcrossing**. Indeed, if probability of selfing $\alpha = 1$, then

$$\Delta F_I = \frac{2-\alpha}{2} \left(\frac{\alpha}{2-\alpha} - F_I \right) = \frac{2-1}{2} \left(\frac{1}{2-1} - F_I \right) = \frac{1}{2} (1 - F_I)$$

so F_I approaches 1 exponentially fast, at the same rate that Achilles approaches the tortoise in Zeno's paradox. On the other hand, if probability of selfing $\alpha = 0$, then

$$\Delta F_I = \frac{2-\alpha}{2} \left(\frac{\alpha}{2-\alpha} - F_I \right) = \frac{2-0}{2} \left(\frac{0}{2-0} - F_I \right) = -F_I$$

so after one generation, the new inbreeding coefficient F_I is zero.

- Another example: if $\alpha = \frac{1}{2}$, then

$$F_I^* = \frac{1/2}{2 - 1/2} = \frac{1/2}{3/2} = \frac{1}{3}$$

Change in A_1 allele frequency under selection and inbreeding

- On page 127 (before Problem 5.5), Gillespie asserts that “allele frequencies do not change under any system of inbreeding. Genotype frequencies do change, and they change in such a way that there are fewer heterozygotes than are seen in an outbreeding population”. What is meant is that in the case of no mutation, no selection, allele frequencies do not change under any system of inbreeding, which we saw on pages 19-20 of these notes. Here is a **another proof** of that assertion.

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|---|-----------------------|----------------|-----------------------|
| frequency with inbreeding coefficient F_I | $p^2(1 - F_I) + pF_I$ | $2pq(1 - F_I)$ | $q^2(1 - F_I) + qF_I$ |

$$\begin{aligned}
 p' &= p^2(1 - F_I) + pF_I + pq(1 - F_I) \\
 &= p[p(1 - F_I) + F_I + q(1 - F_I)] \\
 &= p[(p + q)(1 - F_I) + F_I] = p \cdot 1 = p
 \end{aligned}$$

- However in the presence of selection, allele frequencies certainly do change, and in a later figure, we will see that the major allele p takes **longer to fix** in **outbreeding** populations than in **selfing** populations!
- Recall that in the **presence of selection**, but **absence of mutation and absence of selfing**, we have the following:

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|---------------------------------------|------------|---------------|--------------|
| frequency under random mating fitness | p^2 1 | $2pq$ 1-hs | q^2 1-s |

with

$$p' = \frac{p^2 w_{1,1} + pq w_{1,2}}{\bar{w}} = \frac{p^2 w_{1,1} + pq w_{1,2}}{p^2 w_{1,1} + 2pq w_{1,2} + q^2 w_{2,2}} = \frac{p^2 + pq(1 - hs)}{p^2 + 2pq(1 - hs) + q^2(1 - s)}$$

$$\Delta_s p = p' - p = \frac{pq s [ph + q(1 - h)]}{1 - 2pqhs - q^2 s} \quad (\text{see pages 10-11 of Chapter 3 notes})$$

Mutation-selection balance equation in absence of selfing

- On page 31 of Chapter 3 notes, it is shown that the change in A_1 allele frequency due to **mutation** is

$$\Delta_u p \approx -u$$

where u is the mutation rate per generation.

- It follows that in the **presence of mutation and selection**, but **absence of selfing**, we have the **mutation-selection balance** equation described on pages 30-31 of Chapter 3 notes, given by

$$\begin{aligned}\Delta_s p + \Delta_u p &= 0 \\ \frac{pq s [ph + q(1-h)]}{1 - 2pqhs - q^2s} - u &= 0 \\ q^* \approx \frac{u}{hs} \quad (\text{mutation-selection balance equation in absence of selfing})\end{aligned}$$

Mutation-selection balance equation in presence of selfing

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|--|--------------------------|------------------------|------------------------------|
| frequency under mixed selfing/outcrossing fitness | $p^2(1 - F_I) + pF$ 1 | $2pq(1 - F_I)$ 1-hs | $q^2(1 - F_I) + qF_I$ 1-s |

- In the case described by the table above,

$$\begin{aligned}
 \bar{w} &= [p^2(1 - F) + pF] \cdot 1 + [2pq(1 - F)] \cdot (1 - hs) + [q^2(1 - F) + qF] \cdot (1 - s) \\
 &= 1 - (1 - F)(2pqhs + sq^2) - Fsq \quad \text{derivation on slide 13, "Mean fitness with inbreeding"} \\
 p' &= \frac{p^2(1 - F) + pF + pq(1 - F) \cdot (1 - hs)}{1 - (1 - F)(2pqhs + sq^2) - Fsq} \\
 \Delta_s p = p' - p &= \frac{[p^2(1 - F) + pF] \cdot 1 + pq(1 - F) \cdot (1 - hs)}{1 - (1 - F)(2pqhs + sq^2) - Fsq} - p \\
 &= \frac{[p^2(1 - F) + pF] \cdot 1 + pq(1 - F) \cdot (1 - hs) - p[1 - (1 - F)(2pqhs + sq^2) - Fsq]}{\bar{w}}
 \end{aligned}$$

- Concentrate now on the **numerator** of the previous expression, which we now denote by ν and which can be written in the form $\nu = \alpha + \beta$, where α regroups all terms that **include** the factor F_I , and β regroups all terms that **do not include** the factor F_I . We have

$$\begin{aligned}
 \nu &= p^2 - p^2F_I + pF_I + pq - pqhs - pqF_I + pqhsF_I - [p - 2p^2qhs + psq^2 - 2p^2qhsF_I - pq^2sF_I + pqsF_I] \\
 \alpha &= F_I [p - pq + pqhs - p^2 - 2p^2qhs - pq^2s + pqs] \\
 \beta &= p^2 + pq - pqhs - p + 2p^2qhs + pq^2s
 \end{aligned}$$

- Noting that all terms found in the expression equal to β are found with the **opposite sign** (i.e. multiplied by -1) in the expression within square brackets for α , we have

$$\begin{aligned}
 \nu &= (1 - F_I) [p^2 + pq - pqhs - p + 2p^2qhs + pq^2s] + F_I pqs = (1 - F_I) [p(p + q - qhs - 1 + 2pqhs + q^2s)] + F_I pqs \\
 &= (1 - F_I) [p(-qhs + 2pqhs + q^2s)] + F_I pqs = (1 - F_I) [pqs(-h + 2ph + q)] + F_I pqs \\
 &= (1 - F_I) [pqs(h\{2p - 1\} + q)] + F_I pqs = (1 - F_I) [pqs(h\{p - (1 - p)\} + q)] + F_I pqs \\
 &= (1 - F_I) [pqs(ph - qh + q)] + F_I pqs \\
 &= (1 - F_I) [pqs(ph + q(1 - h))] + F_I pqs
 \end{aligned}$$

- Thus we have finally established that in the presence of **inbreeding coefficient** F_I , the change in A_1 allele frequency due to selection is given by

$$\begin{aligned}
 \Delta_s p &= \frac{\nu}{\bar{w}} \\
 &= \frac{(1 - F) [pqs(ph + q(1 - h))] + Fpqs}{\bar{w}}
 \end{aligned}$$

- For the result we have just established, Gillespie simply states the result 7 lines from the bottom of page 129, preceded by the statement: “By now you should have little trouble in establishing that . . .”. Journal articles **never** include such algebraic derivations, so this is yet another reminder that whenever you read journal articles, you should do so with a pen and paper nearby.

- Compare the current equation for $\Delta_s p$, in the **presence** of inbreeding coefficient F_I , with the result for $\Delta_s p$ in the **absence** of inbreeding, given on slide 24 and also covered in Chapter 3.

$$\Delta_s p = \frac{pq[s(ph + q(1 - h))]}{\bar{w}} = \frac{pq[s(ph + q(1 - h))]}{1 - 2pqhs - q^2s} \quad (\text{w/o inbreeding})$$

$$\Delta_s p = \frac{(1 - F_I)[pq(ph + q(1 - h))] + F_I pq s}{\bar{w}} = \frac{(1 - F_I)[pq(ph + q(1 - h))] + F_I pq s}{1 - (1 - F_I)(2pqhs + sq^2) - F_I sq} \quad (\text{with inbreeding})$$

- If $p \approx 1$, then $\bar{w} \approx 1$ in both outbreeding ($F_I = 0$) and inbreeding ($F_I > 0$), and we have

$$\Delta_s p = \frac{pq[s(ph + q(1 - h))]}{1 - 2pqhs - q^2s} \approx \frac{qsh}{1} = qsh \quad (\text{w/o inbreeding})$$

$$\Delta_s p = \frac{(1 - F_I)[pq(ph + q(1 - h))] + F_I pq s}{1 - (1 - F_I)(2pqhs + sq^2) - F_I sq} \approx \frac{(1 - F_I)qsh + F_I qs}{1} = (1 - F_I)qsh + F_I qs \quad (\text{with inbreeding})$$

- Recall that on page 31 of Chapter 3 notes, we showed that for $p \approx 1$, the change in A_1 allele frequency due to mutation $A_1 \xrightarrow{u} A_2$ is approximately

$$\Delta_u p = -u$$

- For $p \approx 1$, **mutation-selection equilibrium** occurs when

$$\begin{aligned}\Delta_s p + \Delta_u p &= 0 \\ &\approx qsh - u \quad (\text{w/o inbreeding}) \\ &\approx (1 - F_I)qsh + F_Iqs \quad (\text{with inbreeding})\end{aligned}$$

hence when $p \approx 1$, the **minor allele** frequency q^* at **mutation-selection equilibrium** is given by

$$0 = qsh - u \quad (\text{w/o inbreeding})$$

$$q^* = \frac{u}{sh} \quad (\text{mutation-selection balance equation with inbreeding})$$

$$0 = (1 - F_I)qsh + F_Iqs \quad (\text{with inbreeding})$$

$$u = qsh - F_Iqsh + F_Iqs = q(sh - F_Ish + F_Is) \quad (\text{with inbreeding})$$

$$q^* = \frac{u}{sh(1 - F_I) + F_Is} \quad (\text{mutation-selection balance equation with inbreeding})$$

- Let q_O^* denote the equilibrium minor allele frequency for mutation-selection balance **w/o inbreeding**, hence under **outbreeding**.

- Let q_I^* denote the equilibrium minor allele frequency for mutation-selection balance **under inbreeding**.



$$q_O^* = \frac{u}{sh}$$

$$q_I^* = \frac{u}{sh(1 - F_I) + F_Is} = \frac{u}{s[h + F_I(1 - h)]}$$

- Thus in the case of **incomplete dominance** (Darwinian directed selection) where $0 < h < 1$, it is clear that q_I^* for **inbreeding** populations is **less than** q_O^* for **outbreeding** populations (i.e. populations with no inbreeding). Thus in the presence of mutation, **major allele** equilibrium frequency is greater in **self-fertilizing** populations than in **outbreeding** populations. This is shown in a figure later.

What happens when an outbreeding population suddenly shifts to selfing?

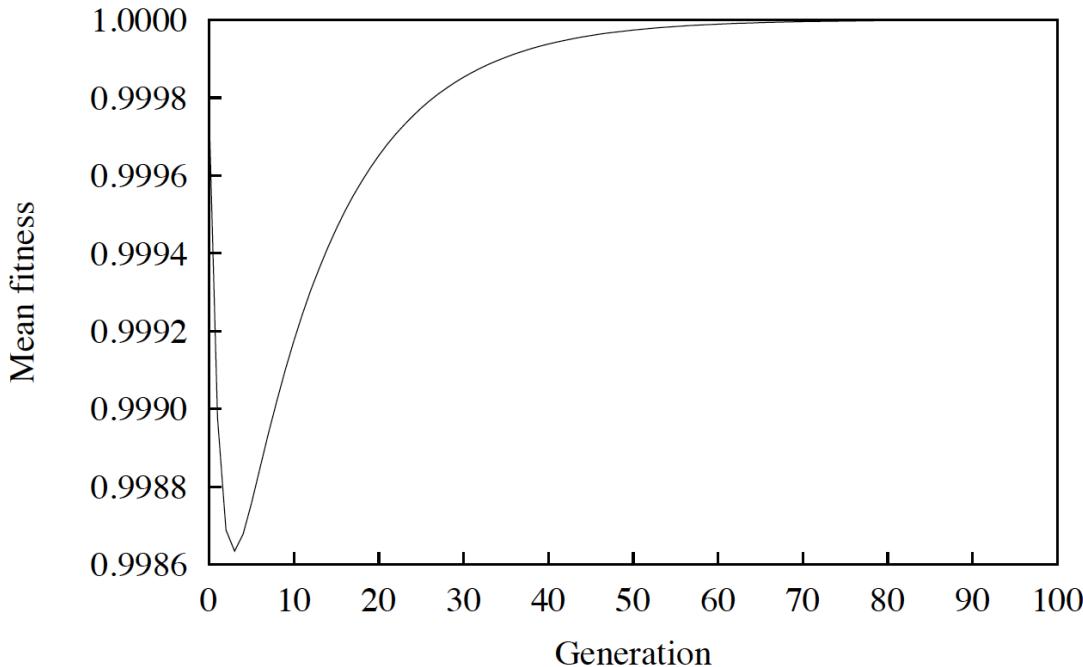
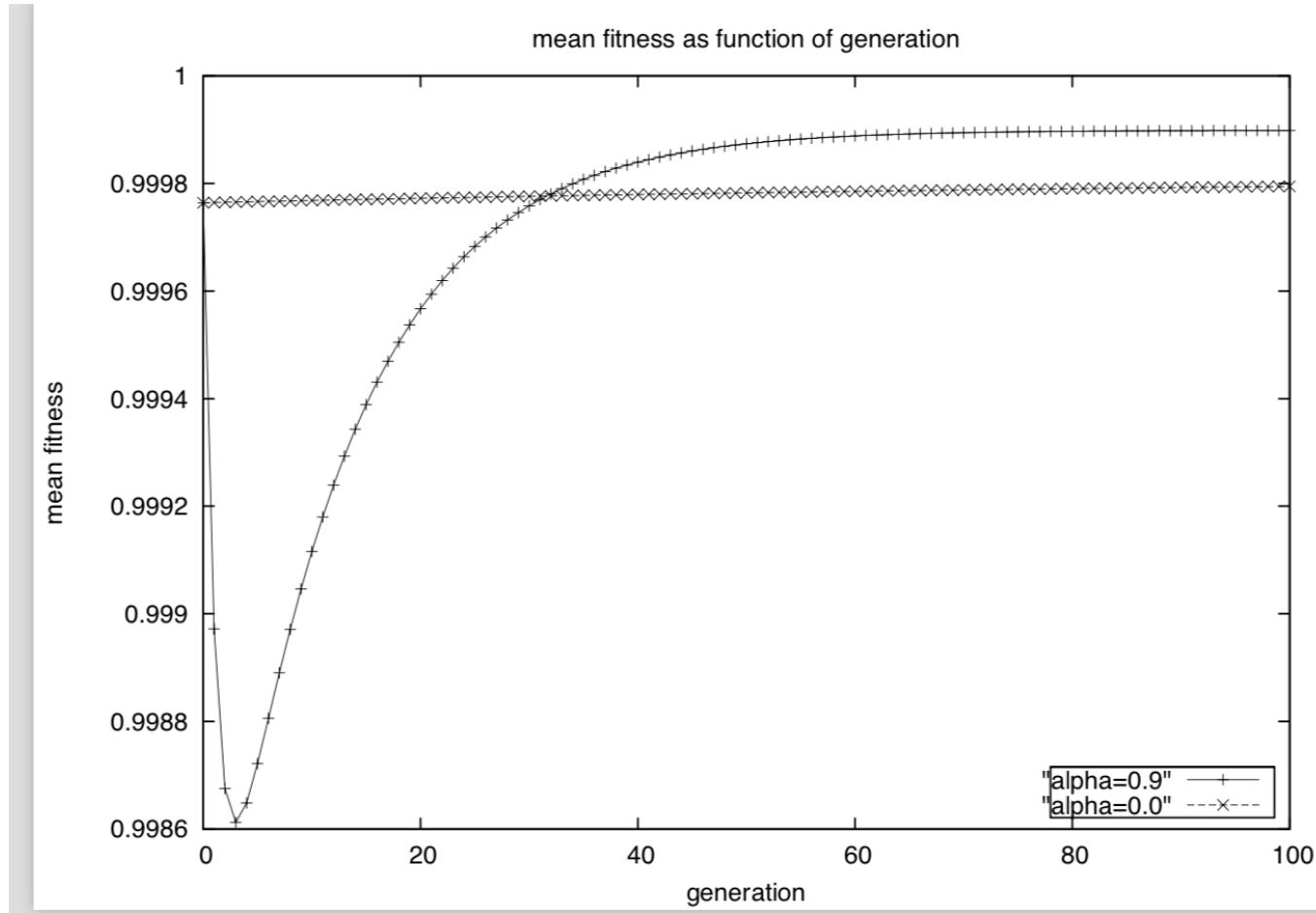


Figure 5.5: The mean fitness of a population that began selfing at generation 0. The parameters are $s = 0.1$, $h = 0.05$, $u = 10^{-4}$, $\alpha = 0.9$ and the initial population was in Hardy-Weinberg equilibrium with $q = u/h s$.

- What Gillespie doesn't stress here is that, having reached **mutation-selection balance** $q^* = \frac{u}{hs}$, upon shifting to selfing, mutation rate drops to 0, so there is **no mutation** any more! Since this is unrealistic, the next slide illustrates the situation when $u = 0.00001$ remains unchanged.



- Reproducing data for Figure 5.5 using program `selfing.py` in DEMOS directory of class website, where parameters are $s = 0.1$, $h = 0.05$, $u = 10^{-4}$, $\alpha = 0.9$, and the initial population was in Hardy-Weinberg equilibrium with $q = u/h s = 0.02$, with initial inbreeding coefficient $F_I = 0$. Initial mean population fitness $\bar{w} = 0.99898$.

- There is an initial **drop** in mean fitness when switching from outbreeding to self-fertilization, due to **inbreeding depression**; however, soon the mean fitness of the selfing population increases and ultimately **exceeds** the mean fitness of the outbreeding population. Outbreeding fitness never changes from the initial value of $\bar{w} = 0.99898$, since the major allele frequency at **mutation-selection equilibrium (w/o selfing)** $p^* = 1 - q^* = 1 - 0.02 = 0.98$. In contrast, the major allele frequency at **mutation-selection equilibrium (with selfing)** satisfies

$$F^* = \frac{\alpha}{2 - \alpha} = \frac{0.9}{2 - 0.9} \approx 0.818182$$

$$q^* = \frac{u}{(1 - F^*)hs + F^*s} = \frac{0.0001}{(1 - 0.818182)(0.05)(0.1) + 0.818182 \cdot 0.1} \approx 0.00121$$

$$p^* = 1 - q^* = 1 - 0.00121 \approx 0.99880$$

Portion of program selfing.py

```
def main(h,s,alpha):
    u      = 0.0001 #mutational probability u
    F      = 0.0      #initial fitness - due to selfing, F changes
    q      = u/(h*s) #initial value of q
    gen   = 0
    p      = 1-q
    wBar  = 1 - (1-F)*(2*p*q*h*s + s*q**2) - F*s*q    #mean fitness
    p0    = p          #initial value of p under assumption of NO SELFING
    q0    = 1-p0
    wBar0= 1 - (2*p0*q0*h*s + s*q0**2)      #mean fitness when F=0
    L    = [] #list of (x,y) coordinates to plot figure
    for gen in range(NUMGEN+1):
        L.append( (gen,p,wBar,p0,wBar0) )
        #-----update p,F, wBar -----
        #Note: p is updated using current mean fitness
        num  = (p**2*(1 - F) + p*F) + p*q*(1 - F)*(1 - h*s) #p^2*fit + pq*fit
        p    = num/wBar #numerator over denominator
        q    = 1-p
        F    = alpha*(F+(1-F)/2)
        wBar = 1 - (1-F)*(2*p*q*h*s + s*q**2) - F*s*q
        #-----update p0,wBar0 when F=0 -----
        num0 = p0**2 + p0*q0*(1-h*s)      #p^2*fit + pq*fit when F=0
        p0   = num0/(1 - 2*p0*q0*h*s - s*(1-p0**2))
        q0   = 1-p0
        wBar0= 1 - (2*p0*q0*h*s + s*q0**2)
        for (x,y,z,u,v) in L:
            print "%s\t%s\t%s\t%s\t%s" % (x,y,z,u,v)
```

Under what conditions is selfing a good strategy?

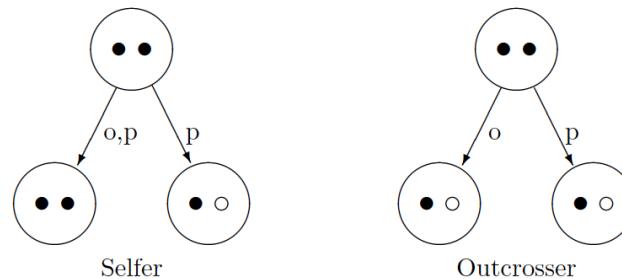


Figure 5.4: The gametes produced by a selfer and an outcrosser. The *p* to the right of an arrow indicates that the parent's contribution came from pollen; an *o* indicates it came from an ovule. The filled circles represent gametes from the illustrated parents; the open circles represent gametes chosen at random from the gamete pool.

- Figure above shows that individuals in **mixed selfing/outcrossing** populations produce
 - ▷ 2 copies of their own genes in each offspring produced by self-fertilization
 - ▷ 1 copy of their own genes in each offspring produced by outcrossingwhile individuals in **pure outcrossing** populations produce only 1 copy of their own genes in each offspring.
- Under the **selfish gene** hypothesis (cf Richard Dawkins), whose goal is to maximize the number of its copies in a population, a mixed selfing/outcrossing strategy may appear optimal – selfing to reduce genetic load and to maximize number of gene copies, and outcrossing to minimize the loss of fitness due to inbreeding depression
- Why don't all organisms use mixed selfing/outbreeding strategy?
- Goal of this section is to provide a formula that determines **when** self-fertilization appears to be a better strategy than outcrossing

- Although more gene copies produced, overall fitness of offspring is less. Recall from slide with header “Inbreeding” that we showed that overall fitness (viability) S satisfies

$$\begin{aligned}
 S &= \prod_i (1 - a_i - b_i F_I) \cdot \prod_j (1 - x_j) \\
 &\approx \prod_i e^{-a_i - b_i F_I} \cdot \prod_j e^{-x_j} \\
 &= \exp\left(\sum_i -a_i\right) \cdot \exp\left(\sum_i -b_i F_I\right) \cdot \exp\left(\sum_j -x_j\right) \\
 &= e^{-A - BF_I} \\
 A &= \sum_i a_i + \sum_j x_j \\
 B &= \sum_i b_i \\
 S &= e^{-A - BF_I} \\
 \log S &= -A - BF_I
 \end{aligned}$$

- The reason for defining terms A, B is to describe a clear **linear** dependence of **fitness** S , in which there is a contribution A that does not multiply by **inbreeding coefficient** F_I , as well as a contribution B that is multiplied by **inbreeding coefficient** F_I .
- From the figure on the previous page, a total of 3 gene copies for the selfing/outbreeding individual (left image) – 2 of the 3 copies (due to selfing) each contribute fitness $e^{-A - BF_I}$ and the remaining of the 3 copies (due to outbreeding) contributes fitness e^{-A} . Thus for selfing/outbreeding individual, overall survival (fitness, viability) is

$$2e^{-A - BF_I} + e^{-A}$$

- From the figure two slides ago, a total of 2 gene copies for the outbreeding individual (right image) – each of 2 copies contributes fitness of e^{-A} , so overall survival (fitness, viability) is

$$2e^{-A}$$

- It follows that **selfing/outbreeding** strategy better than **outbreeding** strategy if and only if

$$\begin{aligned} 2e^{-A-BF_I} + e^{-A} > 2e^{-A} &\Leftrightarrow 2e^{-A-BF_I} > e^{-A} \Leftrightarrow 2e^{-BF_I} > 1 \Leftrightarrow e^{-BF_I} > \frac{1}{2} \\ &\Leftrightarrow -BF_I > \ln(2^{-1}) = -\ln(2) \approx -0.6931 \end{aligned}$$

- $F_I = \frac{1}{2}$ for the selfed offspring of an outbred individual, since probability is $\frac{1}{2}$ that each gene (chromosome) of the offspring is a copy (identical by descent) of the same maternal chromosome. Thus in such cases, **selfing is advantageous** (from the perspective of a greater contribution to overall fitness) if and only if

$$-BF_I > -0.6931 \Leftrightarrow B < 2 \cdot 0.6931 \approx 1.3862$$

- mutation rate u decreases major allele frequency

$$\Delta_u p \approx -u$$

since $A_1 \xrightarrow{u} A_2$.

- As shown on pages 27-28 of these notes), provided that $p \approx 1$, selection s increases major allele frequency by the increment

$$\Delta_s p \approx (1 - F_I)qhs + F_Iqs$$

- equilibrium occurs when

$$\begin{aligned} \Delta_u p + \Delta_s p = 0 &\Leftrightarrow -u + (1 - F_I)qhs + F_Iqs = 0 \Leftrightarrow u = q[(1 - F_I)hs + F_I s] \\ &\Leftrightarrow q^* = \frac{u}{(1 - F_I)hs + F_I s} \end{aligned}$$

where we now denote the minor allele frequency at equilibrium by q^*

- mean fitness at equilibrium is obtained by replacing q by q^* in the defining equation for mean fitness

$$\begin{aligned}
 \bar{w} &= 1 - (1 - F_I)(2pqsh + q^2s) - F_Iqs \\
 &\approx 1 - (1 - F_I)2pqsh - F_Iqs \\
 \bar{w}^* &\approx 1 - (1 - F_I)2pq^*sh - F_Iq^*s \\
 &= 1 - (1 - F_I)2psh \cdot \frac{u}{(1 - F_I)hs + F_I s} - F_I s \cdot \frac{u}{(1 - F_I)hs + F_I s} \\
 \bar{w}^* &\approx 1 - u \cdot \frac{2(1 - F_I)h + F_I}{(1 - F_I)h + F_I}
 \end{aligned}$$

- recall that **genetic load at equilibrium** is defined by

$$\begin{aligned}
 L &= \frac{w_{\max} - \bar{w}^*}{w_{\max}} \\
 &\approx \frac{1 - \left[1 - u \cdot \frac{2(1 - F_I)h + F_I}{(1 - F_I)h + F_I} \right]}{1} \\
 &= u \cdot \frac{2(1 - F_I)h + F_I}{(1 - F_I)h + F_I}
 \end{aligned}$$

- If $F_I = 0$ then the genetic load in **outbreeding** is

$$\begin{aligned}
 L &\approx \frac{1 - \left[1 - u \cdot \frac{2(1 - 0)h + 0}{(1 - 0)h + 0} \right]}{1} \\
 &= u \cdot \frac{2h}{h} = 2u
 \end{aligned}$$

- If $F_I = 1$ then the genetic load in **complete inbreeding** is

$$\begin{aligned}
 L &\approx \frac{1 - \left[1 - u \cdot \frac{2(1 - 1)h + 1}{(1 - 1)h + 1} \right]}{1} \\
 &= u \cdot \frac{1}{1} = u
 \end{aligned}$$

- It follows that **inbreeding** has **half the genetic load of outbreeding!**

Subdivision increases homozygosity

| Genotype: | A_1 | A_1A_1 | A_1A_2 | A_2A_2 |
|-----------------------------|-------|----------|----------|----------|
| Frequency in patch 1: | 1/4 | 1/16 | 3/8 | 9/16 |
| Frequency in patch 2: | 3/4 | 9/16 | 3/8 | 1/16 |
| Frequency in species: | 1/2 | 5/16 | 3/8 | 5/16 |
| Hardy-Weinberg frequencies: | 1/2 | 1/4 | 1/2 | 1/4 |

- Assume a (plant) species divided into 2 plots of land, with an equal (large) number of individual plants in each plot, and that rate of growth, and other parameters are identical for both plots. Assume each subpopulation has reached Hardy-Weinberg equilibrium, so that the genotype equilibrium frequencies are as given by

$$Prob[A_1A_1] = p_i^2$$

$$Prob[A_1A_2] = 2p_iq_i$$

$$Prob[A_2A_2] = q_i^2$$

in which $p_1 = \frac{1}{4}$ (so $q_1 = \frac{3}{4}$) and $p_2 = \frac{3}{4}$ (so $q_2 = \frac{1}{4}$).

- The overall A_1 allele frequency p , obtained by combining both plots is given by the **weighted** sum of A_1 allele frequencies from each plot

$$\begin{aligned} \text{Prob}[A_1] &= \frac{1}{2} \cdot p_1 + \frac{1}{2} \cdot p_2 \\ &= \left(\frac{1}{2} \cdot \frac{1}{4}\right) + \left(\frac{1}{2} \cdot \frac{3}{4}\right) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2} \end{aligned}$$

so the genotype frequencies of the cultivation would be

$$\begin{aligned} \text{Prob}[A_1A_1] &= p^2 = 1/4 \\ \text{Prob}[A_1A_2] &= 2pq = 1/2 \\ \text{Prob}[A_2A_2] &= q^2 = 1/4 \end{aligned}$$

if there were no subdivisions.

- There is an increase of **homozygosity** in the entire population due to the subdivision. Note that both **inbreeding** and **subdivision reduce heterozygosity**, where genotype frequencies are given in the following table, in which F_{ST} is the **subdivision** analogue of the **inbreeding coefficient** F_I .
- Assume **subdivision i** contains **proportion** c_i of the entire population, and that each subpopulation is in Hardy-Weinberg equilibrium, so that $p_i = \text{Prob}[A_1A_1] + \frac{1}{2}\text{Prob}[A_1A_2]$ in the i th subdivision. In the table, we'll refer to subdivision as **plot**, since this is easily visualized for plant cultivation.

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|---|-----------------------------|---------------------------|------------------------------------|
| HW equil in plot i | p_i^2 | $2p_iq_i$ | q_i^2 |
| Avg genotype freq in population | $\sum_i c_i p_i^2$ | $\sum_i 2c_i p_i q_i$ | $\sum_i c_i q_i^2$ |
| HW equil with included F_{ST} genotype fitness | $p^2(1 - F_{ST}) + pF$ 1 | $2pq(1 - F_{ST})$ 1-hs | $q^2(1 - F_{ST}) + qF_{ST}$ 1-s |

- Equate heterozygosities from previous table

$$2pq(1 - F_{ST}) = \sum_i c_i 2p_i q_i$$

$$\begin{aligned} F_{ST} &= 1 - \frac{\sum_i c_i 2p_i q_i}{2pq} = \frac{2pq - (1 - \sum_i c_i p_i^2 - \sum_i c_i q_i^2)}{1 - p^2 - q^2} \\ &= \frac{(1 - p^2 - q^2) - 1 + \sum_i c_i p_i^2 + \sum_i c_i q_i^2}{1 - (p^2 - q^2)} \\ &= \frac{\sum_i c_i (p_i^2 + q_i^2) - (p^2 + q^2)}{1 - (p^2 - q^2)} \end{aligned}$$

$G_T = p^2 + q^2$ total homozygosity without subdivision

$G_S = \sum_i c_i (p_i^2 + q_i^2)$ total homozygosity for population with subdivision

$$F_{ST} = \frac{G_S - G_T}{1 - G_T}$$

- G_T is homozygosity of entire population, i.e. more specifically, G_T is probability that two alleles drawn at random (with replacement) from the entire species are identical by state.
- G_S is the average of homozogysities of all subpopulations, i.e. more specifically, G_S is average of all subdivision probabilities p_i , where p_i is the probability that two alleles drawn at random (with replacement) from the i th subpopulation are identical by state.

- From the previous page, we have shown that F_{ST} satisfies

$$\begin{aligned} F_{ST} &= \frac{\sum_i c_i(p_i^2 + q_i^2) - (p^2 + q^2)}{1 - (p^2 - q^2)} \\ &= \frac{(\sum_i c_i p_i^2 - p^2) + (\sum_i c_i q_i^2 - q^2)}{2pq} \end{aligned}$$

- Our goal is to show that as long as there is **variation** in the allele frequency p_i across subdivisions, then $F_{ST} > 0$ and so overall heterogeneity will be **decreased**.
- We do that by showing that $\sum_i c_i p_i^2 - p^2$ is the **variance** of a new random variable X , that $\sum_i c_i q_i^2 - q^2$ is the **variance** of a new Y , that $Y = 1 - X$, so $V[Y] = V[X]$.
- Suppose that you have a k -faced die, where the i th face has p_i spots, for which $\sum_{i=1}^k p_i = 1$. Suppose as well that the die is **loaded**, so that it comes up on face i with probability c_i , and $\sum_{i=1}^k c_i p_i = p$. This is summarized by the random variable X , where

$$X = \begin{cases} p_1 & \text{with probability } c_1 \\ p_2 & \text{with probability } c_2 \\ \vdots & \vdots \\ p_k & \text{with probability } c_k \end{cases}$$

- Similarly, define the random variable Y , where

$$Y = \begin{cases} q_1 & \text{with probability } c_1 \\ q_2 & \text{with probability } c_2 \\ \vdots & \vdots \\ q_k & \text{with probability } c_k \end{cases}$$

for which $q_i = 1 - p_i$.

- Clearly the expectation satisfies

$$E[X] = \sum_{i=1}^k c_i p_i = p$$

$$E[Y] = \sum_{i=1}^k c_i q_i = \sum_{i=1}^k c_i(1 - p_i) = \sum_{i=1}^k c_i - \sum_{i=1}^k c_i p_i = 1 - p = q$$

- Similarly, the **second moment** and **variance** satisfy

$$E[X^2] = \sum_{i=1}^k c_i p_i^2$$

$$V[X] = E[X^2] - (E[X])^2 = \left(\sum_{i=1}^k c_i p_i^2 \right) - p^2$$

$$E[Y^2] = \sum_{i=1}^k c_i q_i^2$$

$$V[Y] = E[Y^2] - (E[Y])^2 = \left(\sum_{i=1}^k c_i q_i^2 \right) - q^2$$

- Noting that $Y = 1 - X$, we have

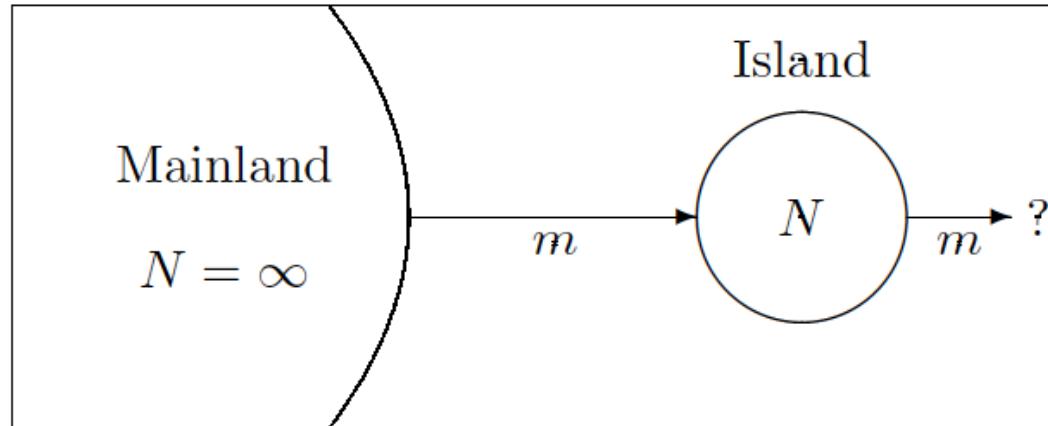
$$V[Y] = V[1 - X] = V[1] + (-1)^2 \cdot V[X] = 0 + 1 \cdot V[X] = V[X]$$

- Thus

$$\begin{aligned} F_{ST} &= \frac{\left(\sum_i c_i p_i^2 - p^2\right) + \left(\sum_i c_i q_i^2 - q^2\right)}{2pq} \\ &= \frac{V[X] + V[Y]}{2pq} \\ &= \frac{2V[X]}{2pq} \end{aligned}$$

- Thus, as long as there is **variation** in the allele frequency p_i across subdivisions, then $F_{ST} > 0$ and so overall heterogeneity will be **decreased**.
- Wahlund's effect is the fact that there is a decrease of heterozygosity due to subdivision

Wright's island model of migration



- assumptions
 - ▷ **Haploid gametes** migrate (despite species being diploid), with $2Nm$ migrants each generation – this is our usual convention of considering a population of N individuals instead as a pool of $2N$ gametes.
 - ▷ N is large, but finite, so presumably, the mainland population is regenerated to ensure constant population size of N – otherwise, mainland population would eventually have dwindled to zero population
 - ▷ Only genetic drift (due to finite population) and migration are considered, **without** either selection or mutation.
- Recall that **homozygosity** \mathcal{G} is defined as the probability that two **distinct** gametes from the pool of $2N$ gametes are **identical by descent**, i.e. copies of the same DNA. On pages 10-11 of Chapter 2 notes, we showed that value of \mathcal{G} differs from that of homozygosity $G = 2pq$ (or more generally $\sum_i 2p_i q_i$) by only a small amount:

$$G - \mathcal{G} \approx \frac{pq}{N}$$

- The formal investigation of drift + migration with migration probability m is essentially identical to that of drift + mutation with mutation probability u , which we covered on pages 16-17 of Chapter 2 notes.
- When establishing residence on a new uninhabited island, a pioneer colony is generally rather homogeneous (e.g. inhabitants of Iceland) – in contrast mainland population is very heterogeneous – indeed, if the mainland has achieved equilibrium heterozygosity $\mathcal{H}^* = \frac{4Nu}{1+4Nu}$ under mutation and drift, then for N large, heterozygosity could be close to 1.
- Migration from the heterogeneous mainland to the homogeneous island will **increase** the heterozygosity of the island, or equivalently **decrease** the homozygosity of the island.
- Our goal now is to formally model the effect of migration from the mainland into the island, and the ensuing **equilibrium** homozygosity and heterozygosity of the island.
- The mathematical model will answer the question: How many persons need to migrate on average each generation for the equilibrium heterozygosity of the island to be essentially equal to that of the mainland?
- In the presence of drift, but without mutation, selection or migration,

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

- the first term represents the probability that two randomly chosen distinct gametes from next generation are both copies of the same gamete in the current generation – here, you should recall that genetic drift follows the Wright-Fisher model, which produces the next generation by $2N$ independent selections of a gamete from the current generation (**selection with replacement**)
- the right term represents the probability that the two randomly chosen distinct gametes from next generation are **not** both copies of the same gamete in the current generation, and that nevertheless are identical by descent, hence their possibly distant ancestors were copies of the same ancient gamete

- Now consider the effect on **island homozygosity** for both **drift** and **migration from the mainland**, but without mutation or selection. Select two randomly chosen distinct gametes from next generation. The probability that neither is a new migrant is $(1 - m)^2$. Above, we saw that when considering only drift, we had $\mathcal{G}' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \mathcal{G}$, so clearly

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

$$(1 - m)^2 = 1 - 2m + m^2 \approx 1 - 2m$$

$$\mathcal{G}' \approx (1 - 2m) \cdot \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \mathcal{G} \right]$$

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

$$\mathcal{G}' \approx \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \mathcal{G} \right] - 2m\mathcal{G}$$

$$\mathcal{H}' = 1 - \mathcal{G}' \approx 1 - \frac{1}{2N} - \mathcal{G} + \frac{\mathcal{G}}{2N} + 2m\mathcal{G} = (1 - \mathcal{G}) - \frac{1 - \mathcal{G}}{2N} + 2m\mathcal{G}$$

$$= (1 - \mathcal{G}) \left[1 - \frac{1}{2N} \right] + 2m\mathcal{G} = \mathcal{H} \left[1 - \frac{1}{2N} \right] + 2m(1 - \mathcal{H})$$

$$\Delta\mathcal{H} = \mathcal{H}' - \mathcal{H} = \mathcal{H} - \frac{\mathcal{H}}{2N} + 2m(1 - \mathcal{H}) - \mathcal{H} = -\frac{\mathcal{H}}{2N} + 2m - 2m\mathcal{H}$$

$$\Delta\mathcal{H} = 0 \Leftrightarrow -\frac{\mathcal{H}}{2N} + 2m - 2m\mathcal{H} = 0 \Leftrightarrow \frac{\mathcal{H}}{2N} + 2m\mathcal{H} = 2m \Leftrightarrow \mathcal{H} \left(2m + \frac{1}{2N} \right) = 2m \Leftrightarrow \mathcal{H} \left(\frac{4Nm + 1}{2N} \right) = 2m \Leftrightarrow \mathcal{H} = \frac{4Nm}{4Nm + 1}$$

$$\Delta\mathcal{H} = 0 \Leftrightarrow \mathcal{H} = \frac{4Nm}{4Nm + 1}$$

- **Equilibrium** values of heterozygosity and homozygosity on the island are thus

$$\mathcal{G}^* = \frac{1}{4Nm + 1}$$

$$\mathcal{H}^* = \frac{4Nm}{4Nm + 1}$$

- Since

$$\mathcal{G}^* = \frac{1}{4Nm + 1}$$

it follows that homozygosity, as measured by \mathcal{G} , changes quickly from near 1 when the number of migrants $2Nm$ is small, to very small values near 0 in the case of mass migration from the (highly heterogeneous) mainland to the island.

- On page 135, Gillespie states:

When $4Nm \gg 1$, the island population's homozygosity is like that of the mainland. Thus, when the fraction of migrants is greater than about $\frac{1}{4N}$, the effects of isolation become unimportant and the island appears to be infinite in size, as indicated by its value of \mathcal{G} . This observation becomes compelling when expressed in numbers of migrants rather than in fraction of migrants.

Isolation disappears if $m > 1/4N$, or, equivalently, if $2Nm > \frac{1}{2}$. The absolute number of diploid migrants each generation is $2Nm$ (strictly, the number of pairs of haploid genomes, as gametes rather than zygotes migrate in our model).

If more than one individual migrates every other generation, then the effects of isolation become unimportant. Surprisingly, this statement is independent of the population size of the island. One might have thought that more migration would be required to make large islands like the mainland. However, in large islands drift is a weak force, so less migration is needed to balance drift. The message is clear: very little migration can make a subdivided species appear like one large randomly mating species when neutral alleles are involved.”

Example problems

| Genotype: | A_1 | A_1A_1 | A_1A_2 | A_2A_2 |
|-----------------------------|-------|----------|----------|----------|
| Frequency in patch 1: | 1/4 | 1/16 | 3/8 | 9/16 |
| Frequency in patch 2: | 3/4 | 9/16 | 3/8 | 1/16 |
| Frequency in species: | 1/2 | 5/16 | 3/8 | 5/16 |
| Hardy-Weinberg frequencies: | 1/2 | 1/4 | 1/2 | 1/4 |

- **Problem 5.8:** Verify that the expression for F_{ST} using the variance in allele frequencies gives the correct value for the table above.

■ Solution:

$$p_1 = \frac{1}{4}$$

$$p_2 = \frac{3}{4}$$

$$p = \left(\frac{1}{2} \cdot \frac{1}{4} \right) + \left(\frac{1}{2} \cdot \frac{3}{4} \right) = \frac{1}{2}$$

$$F_{ST} = \frac{\left(\sum_i c_i p_i^2 - p^2 \right) + \left(\sum_i c_i q_i^2 - q^2 \right)}{2pq}$$

$$\sum_i c_i p_i^2 = \left(\frac{1}{2} \cdot \left(\frac{1}{4} \right)^2 \right) + \left(\frac{1}{2} \cdot \left(\frac{3}{4} \right)^2 \right) = \frac{1}{32} + \frac{9}{32} = \frac{10}{32} = \frac{5}{16}$$

$$\sum_i c_i p_i^2 - p^2 = \frac{5}{16} - \left(\frac{1}{2} \right)^2 = \frac{1}{16}$$

$$\sum_i c_i q_i^2 = \left(\frac{1}{2} \cdot \left(\frac{3}{4} \right)^2 \right) + \left(\frac{1}{2} \cdot \left(\frac{1}{4} \right)^2 \right) = \frac{5}{16}$$

$$\sum_i c_i q_i^2 - q^2 = \frac{5}{16} - \left(\frac{1}{2} \right)^2 = \frac{1}{16}$$

$$F_{ST} = \frac{\left(\sum_i c_i p_i^2 - p^2 \right) + \left(\sum_i c_i q_i^2 - q^2 \right)}{2pq} = \frac{\frac{2}{16}}{2 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{4}$$

- $F_{ST} = 1/4$ is also computed using the following Mathematica code:

```
= q = 1 - p;  
eq1 = p^2 (1 - F) + p F == 5/16;  
eq2 = 2 p q (1 - F) == 3/8;  
eq3 = q^2 (1 - F) + q F == 5/16;  
NSolve[{eq1, eq2, eq3}, {p, F}]  
= {{p → 0.5, F → 0.25}}
```

■ **Problem 5.10:** Calculate F_{ST} for the data from the following table.

| Country | pi | qi |
|----------|-------|-------|
| England | 0.637 | 0.363 |
| Italy | 0.661 | 0.339 |
| West | 0.701 | 0.299 |
| Thailand | 0.746 | 0.254 |
| Japan | 0.724 | 0.276 |
| Nigeria | 0.942 | 0.058 |
| Canadian | 0.556 | 0.444 |
| Papua | 0.880 | 0.120 |

1. Assume that all of the subdivisions are the same size so $c_i = \frac{1}{8}$.
2. Assume that $c_1 = c_2 = c_3 = c_4 = 0.2$ and $c_5 = c_6 = c_7 = c_8 = 0.05$.

■ **Solution:** For $c_1 = c_2 = c_3 = c_4 = c_5 = c + 6 = c_7 = c_8 = \frac{1}{8}$, we have the solution.

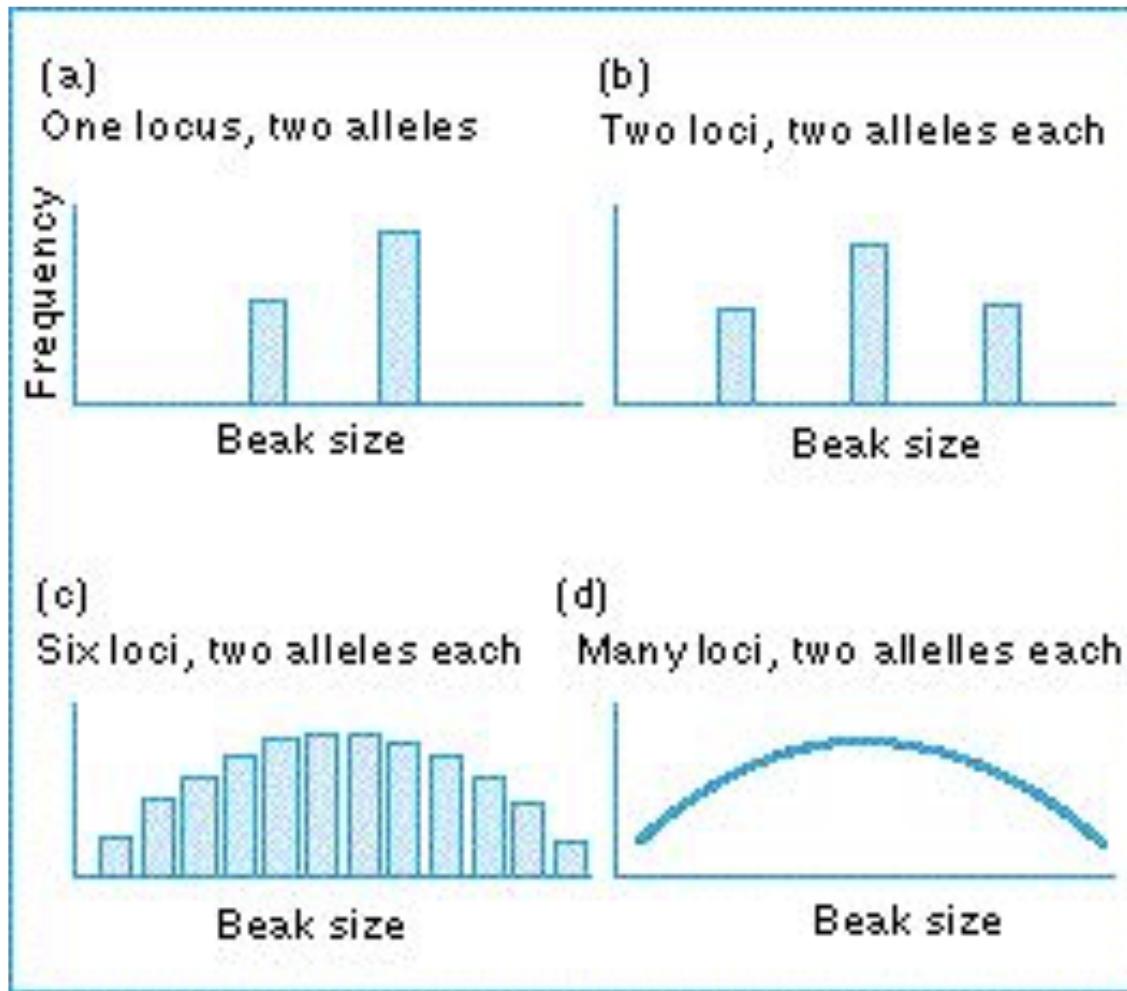
| Country | pi | qi | ci | $(pi)^2$ | $(qi)^2$ |
|-------------------|---------------|-------|-------|----------|----------|
| England | 0.637 | 0.363 | 0.125 | 0.405769 | 0.131769 |
| Italy | 0.661 | 0.339 | 0.125 | 0.436921 | 0.114921 |
| West | 0.701 | 0.299 | 0.125 | 0.491401 | 0.089401 |
| Thailand | 0.746 | 0.254 | 0.125 | 0.556516 | 0.064516 |
| Japan | 0.724 | 0.276 | 0.125 | 0.524176 | 0.076176 |
| Nigeria | 0.942 | 0.058 | 0.125 | 0.887364 | 0.003364 |
| Canadian | 0.556 | 0.444 | 0.125 | 0.309136 | 0.197136 |
| Papua | 0.880 | 0.120 | 0.125 | 0.774400 | 0.014400 |
| p | 0.7309 | | | | |
| q | 0.2691 | | | | |
| sum ci*(pi)^2-p^2 | 0.0140 | | | | |
| sum ci*(qi)^2-q^2 | 0.0140 | | | | |
| 2pq | 0.3934 | | | | |
| F_ST | 0.0713 | | | | |

■ **Solution:** For $c_1 = c_2 = c_3 = c_4 = 0.2$ and $c_5 = c_6 = c_7 = c_8 = 0.05$, we have the solution.

| Country | pi | qi | ci | $(pi)^2$ | $(qi)^2$ |
|-------------------------|---------------|-------|-------|----------|----------|
| England | 0.637 | 0.363 | 0.200 | 0.405769 | 0.131769 |
| Italy | 0.661 | 0.339 | 0.200 | 0.436921 | 0.114921 |
| Nest | 0.701 | 0.299 | 0.200 | 0.491401 | 0.089401 |
| Thailand | 0.746 | 0.254 | 0.200 | 0.556516 | 0.064516 |
| Japan | 0.724 | 0.276 | 0.050 | 0.524176 | 0.076176 |
| Nigeria | 0.942 | 0.058 | 0.050 | 0.887364 | 0.003364 |
| Canadian | 0.556 | 0.444 | 0.050 | 0.309136 | 0.197136 |
| Papua | 0.880 | 0.120 | 0.050 | 0.774400 | 0.014400 |
| p | 0.7041 | | | | |
| q | 0.2959 | | | | |
| $\sum ci^*(pi)^2 - p^2$ | 0.0071 | | | | |
| $\sum ci^*(qi)^2 - q^2$ | 0.0071 | | | | |
| pq | 0.4167 | | | | |
| $\therefore ST$ | 0.0342 | | | | |



Chapter 6: Quantitative genetics



Why do many characters have a normal distribution?

(Image from https://www.blackwellpublishing.com/ridley/tutorials/Quantitative_genetics6.asp)

Why do many characters have a normal distribution?

- From **Evolution** by Mark Ridley, 3rd Edition:

The reason is as follows. Imagine first that beak size was controlled by a single pair of Mendelian alleles at one locus, with one dominant to the other, AA and Aa long and aa short: in this case, the population would contain two categories of individuals.

Imagine now that it was controlled by two loci with two alleles each. Beak size might now have a background value (say, 1 cm) plus the contribution of the two loci, with an A or a B adding 0.1 cm. If A and B were dominant to a and b , then an aabb individual would have a 1 cm beak, AAbb , Aabb , aaBB , and aaBb 1.1 cm, and AABB , AaBB , AABb , and AaBb 1.2 cm. The figure on the left is the frequency distribution if all alleles had frequency one-half and the two loci were in linkage equilibrium. The distribution now has three categories and has become more spread out.

It becomes still more spread out if it is influenced by six loci and it starts to look normal for 12 loci. Thus, when a large enough number of genes influence a character, it will have a continuous, normal frequency distribution. The normal distribution can result either if there are a large number of alleles at each of a small number of loci influencing the characters, or if there are fewer alleles at a larger number of loci. In this tutorial, we shall mainly discuss the theory of quantitative genetics as if there were many loci, each with a small number of alleles. This may well be the genetic system underlying many continuously varying characters; however, the theory applies equally well when there are a few (even one) loci and many alleles at each.

- quantitative genetics concerns inheritance of **quantitative characters**, such as
 - height, weight, blood pressure, pulse, etc.
- traditional focus of quantitative genetics
 - correlation** of quantitative traits between relatives
 - response to selection**, with applications to breeding
- quantitative genetics concerns **variation** rather than **mean** value of a trait

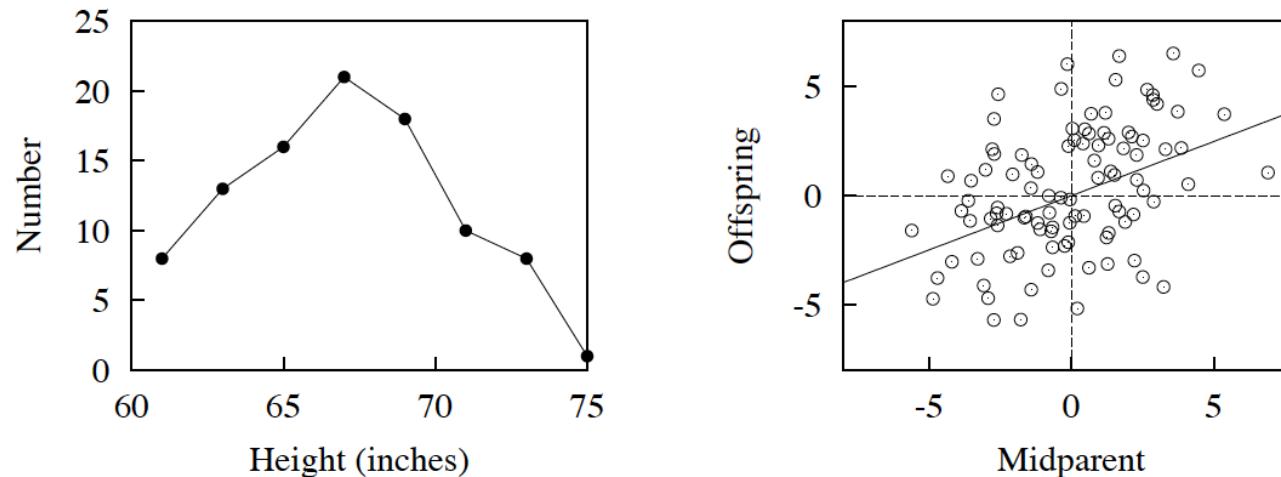


Figure 6.1: The left-hand figure is a histogram of the number of students of a particular height in an evolution class at UC Davis. The right-hand figure graphs the deviation of a student's height from the population mean against the deviation of the student's parents' average height from the population mean. The correlation coefficient is 0.476.

- Note that the figure on the left is approximately a **normal distribution**, as expected by Fisher's theorem that the genetic contribution to traits affected by many di-allelic loci, each of small effect, is approximately a **normal distribution**. We'll prove this at the end of the chapter.

- heritability (h^2): amount of phenotypic (observable) variation in a population that is due to genetic differences, expressed as **ratio** of variation due to genetic differences over total variation – value of heritability $0 \leq h^2 \leq 1$
- nature versus nurture:** twin studies in humans used to determine the importance of genes versus environment
- heritability of 0.8 does **not** mean that a trait is 80% caused by genetic factors
- heritability of 0.8 **does** mean that 80% of the variability in the trait in a population is due to genetic differences among people
- P denotes **phenotype** (quantifiable trait), expressed as **deviation** from population mean
- X_m denotes additive effect of **maternally derived** allele, expressed as **deviation** from population mean
- X_p denotes additive effect of **paternally derived** allele, expressed as **deviation** from population mean
- E denotes additive effect of **environmental influences**, expressed as **deviation** from population mean

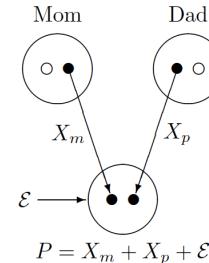


Figure 6.2: The additive model of inheritance for parents and offspring.

additive model

$$P = X_m + X_p + E$$

- heritability, denoted by h^2 , is defined by

$$h^2 = \frac{V_A}{V_A + V_E} \quad (\text{additive variance (from each parent) over variance from parents and environment})$$

$$V_A = V_m + V_p \quad (V_m \text{ maternal variance, } V_p \text{ paternal variance})$$

$$V_E = \text{variance due to environmental influences}$$

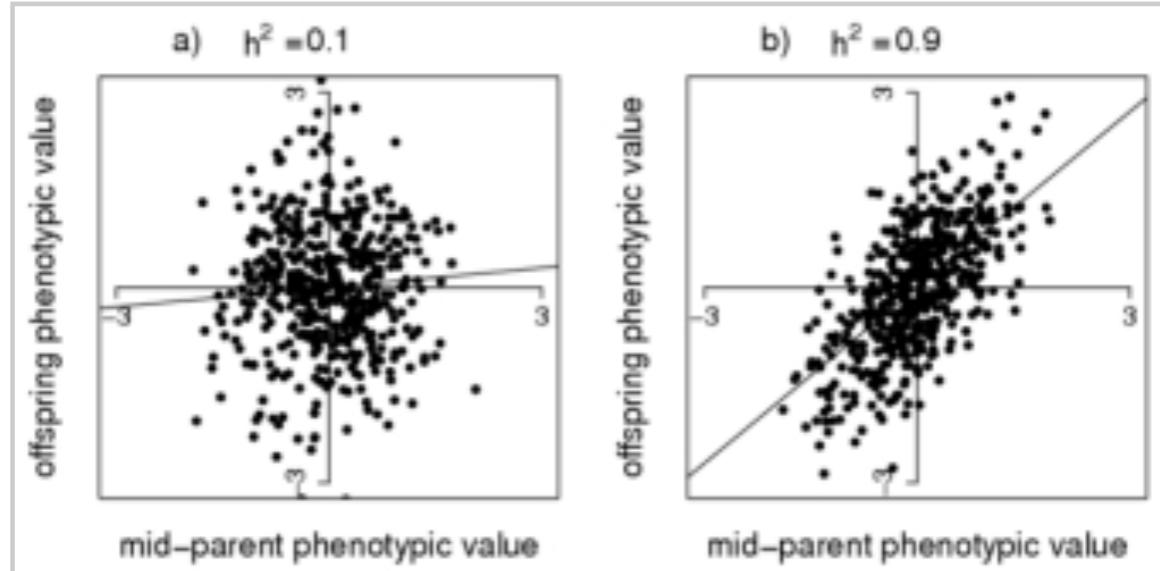


Figure 1: Heritability estimation.

Low (panel a) and high (panel b) heritability can be estimated from the regression (h^2) of offspring phenotypic values vs. the average of parental phenotypic values.

© 2008 **Nature Education** All rights reserved.



<https://www.nature.com/scitable/topicpage/estimating-trait-heritability-46889/>

- h^2 : slope of regression line in **scatter plot of offspring phenotypic deviation from mean as a function of midparent phenotypic deviation from mean**

- we will soon show that **heritability** h^2 is equal to the **slope** of the **regression line (least squares fit)** for scatter plot data for **offspring deviations from the mean** as a function of **midparent deviations from the mean**, given by

$$h^2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{E[XY] - \mu_x \mu_y}{E[X^2] - \mu_x^2}$$

where X is the r.v. that expresses the **deviation** from midparent mean, and Y is the r.v. that expresses the **deviation** from offspring mean

- Note:** points in scatter plot form “cloud” around regression line, due to **Mendelian segregation** and to **environmental influences**. Such deviation is measured by the **sum of squared residuals**
- Recall that **expectation is additive**, regardless of the type of random variable or probability distribution. Moreover, expectations are **zero**, since expectation means **average deviation from the mean** for measurements that are assumed to be **normally distributed**

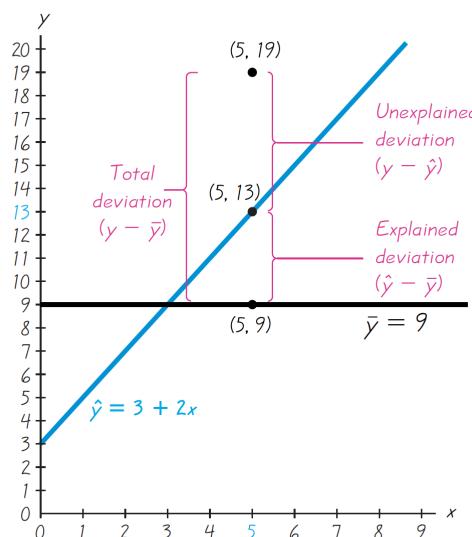
$$\begin{aligned} E[P] &= E[X_m + X_p + \mathcal{E}] = E[X_m] + E[X_p] + E[\mathcal{E}] = 0 \\ E[X_m] &= 0 \quad \text{average maternal deviation from mean} \\ E[X_p] &= 0 \quad \text{average paternal deviation from mean} \\ E[\mathcal{E}] &= 0 \quad \text{average deviation from mean due to environmental factors} \end{aligned}$$



$$\begin{aligned} V[P] &= V[X_m + X_p + \mathcal{E}] = V[X_m] + V[X_p] + V[\mathcal{E}] + \text{Cov}(X_m, X_p) + \text{Cov}(X_m, \mathcal{E}) + \text{Cov}(X_p, \mathcal{E}) \\ \text{Cov}(X_m, X_p) &= 0 \quad \text{random mating assumed (but perhaps tall people tend to marry tall persons)} \\ \text{Cov}(X_m, \mathcal{E}) &= 0 \\ \text{Cov}(X_p, \mathcal{E}) &= 0 \\ V[P] &= V[X_m] + V[X_p] + V[\mathcal{E}] \\ V_A &= V[X_m] + V[X_p] \\ V_E &= V[\mathcal{E}] \end{aligned}$$

- suppose Karen, a junior at BC, is 162 cm tall, while the average height of BC students is 160 cm – then $Y = 162 - 160 = 2$
- Suppose that Karen's father, Bob, is 190 cm tall. According to Wikipedia, average male height in the United States is 175.3 cm (5 ft 9 in), so Bob's deviation from the mean is $190 - 175.3 = 14.7$.
- Suppose that Karen's mother, Alice, is 166 cm tall. According to Wikipedia, average female height in the United States is 161.5 cm (5 ft 3.5 in), so Alice's deviation from the mean is $166 - 161.5 = 4.5$.
- Midparent deviation is $\frac{14.7+4.5}{2} = 9.6$, i.e. $X = 9.6$, and the point $(X, Y) = (9.6, 2)$ is in the scatterplot
- We will soon prove that heritability h^2 equals the slope m of the regression line $y = mx + b$, where the y -intercept $b = 0$ for our data (because values are in terms of deviation from mean). Before ever having met Karen, we can **estimate** her height by

$$\hat{y} = m \cdot x + b = 9.6 \cdot \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$



Why is h^2 the regression line slope for a offspring/midparent scatter plot?

- assume P is a single parent (to fix the discussion, say, a mother), and O is the offspring of the parent

$$P_p = X_m + X_{m'} + \mathcal{E}_p$$

$$P_o = X_m + X_p + \mathcal{E}_o$$

$$\text{Var}(P_p) = V_p$$

$$\begin{aligned} \text{Cov}(P_p, P_o) &= \text{Cov}(X_m, X_m) + \text{Cov}(X_m, X_p) + \text{Cov}(X_m, \mathcal{E}_o) + \text{Cov}(X_{m'}, X_m) + \text{Cov}(X_{m'}, X_p) + \text{Cov}(X_{m'}, \mathcal{E}_o) + \\ &\quad \text{Cov}(\mathcal{E}_p, X_m) + \text{Cov}(\mathcal{E}_p, X_p) + \text{Cov}(\mathcal{E}_p, \mathcal{E}_o) \\ &= \text{Cov}(X_m, X_m) + \text{Cov}(\mathcal{E}_p, \mathcal{E}_o) = \text{Var}(X_m) = \frac{V_A}{2} \end{aligned}$$

- there may be a relation between offspring and parent environment, so it is a big assumption that $\text{Cov}(\mathcal{E}_p, \mathcal{E}_o) = 0$
- recall from the “Probability Appendix” that correlation coefficient ρ satisfies

$$\rho_{x,y} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

and so

$$\begin{aligned} \text{Corr}(P_p, P_o) &= \frac{\text{Cov}(P_p, P_o)}{\sqrt{\text{Var}(P_p) \cdot \text{Var}(P_o)}} \\ &= \frac{V_A}{2} \cdot \frac{1}{\sqrt{\text{Var}(P_p) \cdot \text{Var}(P_p)}} \\ &= \frac{V_A}{2 \text{Var}(P_p)} = \frac{V_A}{2V_p} \\ &= \frac{h^2}{2} \end{aligned}$$

- note that this derivation assumes that $\text{Var}(P_p) = \text{Var}(P_o)$, which seems reasonable – i.e. we assume that the **phenotypic variation of any set of individuals is the same**
- on the other hand (book does not state this), from “Probability Appendix”, it is clear that the **slope m** of the least squares regression equation $y = mx + b$ equals $\frac{\text{Cov}(P_p, P_o)}{\text{Var}(P)}$, which may be useful in the case that $\text{Var}(P_p) \neq \text{Var}(P_o)$
- let M denote the **midparent** deviation from the mean for maternal (m) and paternal (p) parent

$$\text{Cov}(P_m, P_o) = \frac{V_A}{2}$$

$$\text{Cov}(P_p, P_o) = \frac{V_A}{2}$$

$$P_M = \frac{P_m + P_p}{2}$$

$$\begin{aligned}\text{Cov}(P_M, P_o) &= \text{Cov}\left(\frac{P_m + P_p}{2}, P_o\right) = \frac{1}{2} \cdot \text{Cov}(P_m + P_p, P_o) = \frac{1}{2} \cdot (\text{Cov}(P_m, P_o) + \text{Cov}(P_p, P_o)) \\ &= \frac{1}{2} \cdot \left(\frac{V_A}{2} + \frac{V_A}{2}\right) \\ &= \frac{V_A}{2}\end{aligned}$$

■ **variance of the midparent**

$$\begin{aligned}
 Var(P_M) &= Var\left(\frac{P_m + P_p}{2}\right) = \frac{1}{2^2} \cdot Var(P_m + P_p) = \frac{1}{4} \cdot (Var(P_m) + V(P_p) + 2Cov(P_m, P_p)) \\
 &= \frac{1}{4} \cdot (Var(P_m) + V(P_p)) \quad \text{assuming random mating hypothesis} \\
 &= \frac{1}{4} \cdot (Var(P_m) + V(P_m)) \quad \text{variance of maternal deviation equals that of paternal deviation from mean} \\
 &= \frac{Var(P_m)}{2} = \frac{V_p}{2}
 \end{aligned}$$

■ **slope m of the regression of offspring deviation on midparent deviation** is according to “Probability Appendix”

$$m = \frac{Cov(P_M, P_o)}{Var(P_M)} = \frac{V_A/2}{V_p/2} = \frac{V_A}{V_P} = h^2$$

| Species | Character | Heritability |
|----------------------------|--------------------|--------------|
| Honeybee | oxygen consumption | 0.15 |
| <i>Eurytemora herdmani</i> | length | 0.12 |
| Cricket | wing length | 0.74 |
| Flour beetle | fecundity | 0.36 |
| Red-backed salamander | vertebral count | 0.61 |
| Darwin's finch | weight | 0.91 |
| Darwin's finch | bill length | 0.85 |

Table 6.1: Heritability estimates determined by parent-offspring correlations for a variety of traits and species taken from a paper by Mousseau and Roff (1987).

- more generally, for a relative Y of X , whose probability of sharing k alleles with X is given by r_k , for $k = 0, 1, 2$

$$P_X = X_m + X_p + \mathcal{E}_X$$

$$P_Y = Y_m + Y_p + \mathcal{E}_Y$$

$$\text{Cov}(P_X, P_Y) = \text{Cov}(X_m, Y_m) + \text{Cov}(X_m, Y_p) + \text{Cov}(X_p, Y_m) + \text{Cov}(X_p, Y_p)$$

$$= r_0 \cdot 0 + r_1 \cdot \frac{V_A}{2} + r_2 \cdot 2 \cdot \frac{V_A}{2}$$

$$\text{Cov}(P_X, P_Y) = \left(\frac{r_1}{2} + r_2\right) \cdot V_A = rV_A \quad \text{coefficient of relatedness } r = \left(\frac{r_1}{2} + r_2\right)$$

$$\text{Corr}(P_X, P_Y) = \frac{\text{Cov}(P_X, P_Y)}{\sqrt{\text{Var}(P_X) \cdot \text{Var}(P_Y)}} = \frac{\text{Cov}(P_X, P_Y)}{\sqrt{\text{Var}(P_X) \cdot \text{Var}(P_X)}} = \frac{rV_A}{V_{P_X}} = rh^2$$

- Last line assumes that $\text{Var}(P_X) = \text{Var}(P_Y)$. Without this assumption, we can always use the **slope** of the regression $y = mx + b$ of Y on X

$$m = \beta = \frac{\text{Cov}(P_X, P_Y)}{\text{Var}(P_X)} = rh^2$$

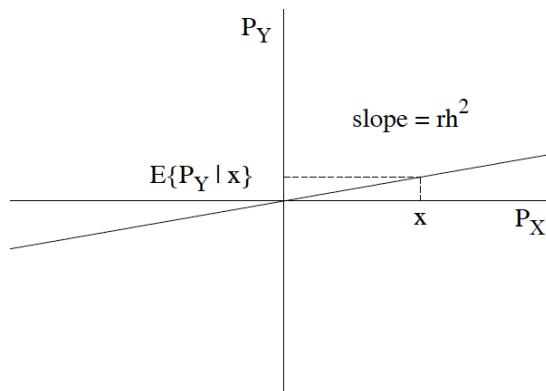


Figure 6.3: The use of regression to find the expected value of the phenotype of relative Y given that the phenotype of relative X is x .

| | Parent-offspring | Midparent-offspring | General |
|------------------------|------------------|---------------------|---------|
| Covariance | $V_A/2$ | $V_A/2$ | rV_A |
| Correlation | $h^2/2$ | $h^2/\sqrt{2}$ | rh^2 |
| Regression coefficient | $h^2/2$ | h^2 | rh^2 |

Table 6.2: A summary of the measures of resemblance between pairs of relatives.

- similar computations establish covariation ($\text{Cov}(X, Y)$), Pearson linear correlation coefficient (r), and regression line slope (m , also denoted β , and called *regression coefficient* in book)

Response to selection

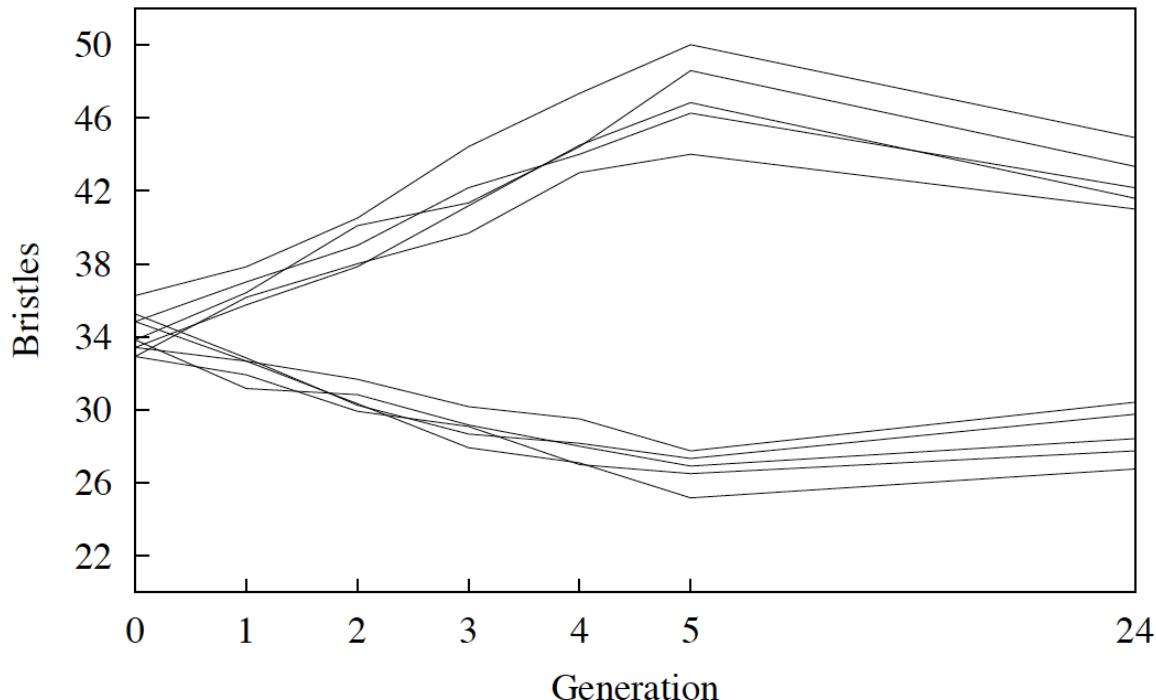


Figure 6.4: The results of a selective breeding experiment for abdominal bristles in *Drosophila*. The upper lines give the number of bristles during five generations of selection for a greater number of bristles in five replicate lines. From generations 6 to 24 there was no selection. The five lower lines give the results of selection for fewer bristles. The data are from Clayton et al. (1957).

- **Goal:** give mathematically precise definition of **selection differential**, denoted S , and of **selection response**, denoted R , and to prove

$$R = h^2 S$$

- **Note:** S is selection differential, and not selection coefficient s , encountered earlier in Chapter 2
- Consider [only] upper set of lines in figure above. Values obtained by retrieving 200 flies from current generation (100 male and 100 female), counting bristles, and selecting 20 males and 20 females, having the greatest number of bristles. Let $n = 20$, and

$$x_1, \dots, x_n$$

denote the midparent bristle values. More precisely,

$$\begin{aligned} m_i &= \text{number of bristles of male in } i\text{th couple} \\ \bar{m} &= \text{avg number of bristles of all 100 males} \\ f_i &= \text{number of bristles of female in } i\text{th couple} \\ \bar{f} &= \text{avg number of bristles of all 100 females} \\ x_i &= \frac{(m_i - \bar{m} + (f_i - \bar{f}))}{2} \quad i\text{th midparent value} \end{aligned}$$

- selection differential S defined as average of the $n = 20$ midparent deviations

$$S = \sum_{i=1}^n \frac{x_i}{n}$$

- recall that h **heritability** h^2 is slope of regression of offspring deviation from mean on midparent deviation from mean

$$y = h^2 \cdot x + b$$

$$b = 0$$

so the expected value of the offspring deviation from mean (estimated value using regression equation) is

$$E[P_O | P_M = x_i] = h^2 x_i$$

- selection response** R defined by

$$\begin{aligned} R &= \frac{1}{n} \sum_{i=1}^n E[P_O | P_M = x_i] \\ &= h^2 \sum_{i=1}^n \frac{x_i}{n} \\ &= h^2 S \end{aligned}$$

- if heritability $h^2 = \frac{1}{2}$, then the **response of one generation of selection** is to **move the population mean halfway** between its value in the parental generation and the mean value of the selected parents

■ According to Plomin et al.

Plomin, R.; Pedersen, N. L.; Lichtenstein, P.; McClearn, G. E. (1994). "Variability and stability in cognitive abilities are largely genetic later in life", *Behavior Genetics*. 24 (3): 207-15

heritability h^2 for IQ in adults is between 0.7 and 0.8, so in a Huxley-esque Brave New World, if n brainy couples are "allowed" to procreate, where $\sum_{i=1}^n \frac{x_i}{n} = 100$, then their offspring can be expected to have an average IQ between 121 and 124

$$\bar{m} = 100$$

$$\bar{f} = 100$$

$$121 = 100 + 0.7 \cdot 30$$

$$124 = 100 + 0.8 \cdot 30$$

■ The following quotation is taken from <https://library.cshl.edu/special-collections/eugenics>:

The ERO [Eugenics Records Office of Cold Spring Harbor Laboratory] was devoted to the collection and analysis of American family genetic and traits history records. These eugenics studies collected information such as inborn physical, mental and temperamental properties to enable the family to trace the segregation and recombination of inborn or heritable qualities. The family study files include individual analysis cards, field worker reports, pedigree charts, and special trait studies. Davenport was president of the American Society of Zoologists and in 1910 he founded the Eugenics Record Office at Cold Spring Harbor, and appointed Harry H. Laughlin to direct it. H. H. Laughlin became a spokesman for the programmatic side of the previous eugenics movement, lobbying for eugenic legislation to restrict immigration and sterilize "defectives," educating the public on eugenic health, and disseminating eugenic ideas widely. The Record Office formally came under the aegis of the Carnegie Institution of Washington in 1918.

Far too few biologists know that Cold Spring Harbor Lab was heavily involved in the **eugenics** movement, itself completely discredited under the Nazi "Endlösung" (final solution), whose goal was to eradicate "inferior races" from our species! Leading intellectuals at the time accepted eugenics, both in the US, Great Britain, Germany and throughout Europe. The epilogue of Crichton's **State of Fear** summarizes the past acceptance of eugenics.

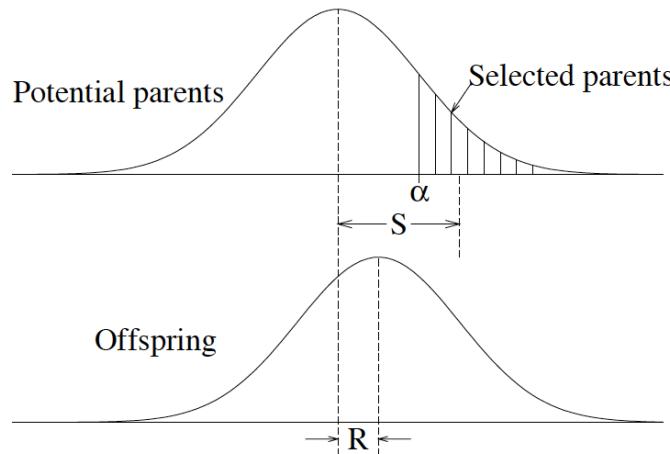


Figure 6.5: The response to selection.

- Derivation below assumes that the quantitative trait is **normally distributed** with mean μ , hence **deviation** from mean ($x - \mu$) is normally distributed with mean 0. Top figure shows the **midparent deviation from mean**.
- The cross-hatched area to the right of α in the top figure is the proportion A of selected individuals, assuming that all individual having midparent deviation greater than α are selected to breed.

$$\sigma^2 = V_P \quad (\text{variance of population midparent deviations})$$

$$A = \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx$$

$$z = \frac{x - \mu}{\sigma} = \frac{x}{\sigma} \quad Z\text{-score corresponding to } x$$

$$A = \int_{\alpha/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz$$

- The **probability density** $p(x)$ for the selected individuals (those having midparent deviation $x \geq \alpha$) is thus the truncated normal density for $x \geq \alpha$ given by

$$p(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)}{A} \quad A \text{ is a normalization constant, to ensure } \int p(x)dx = 1$$

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

$$S = E[X] = \frac{1}{A\sqrt{2\pi}\sigma} \cdot \int_{\alpha}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \quad \text{selection differential for individuals } \geq \alpha$$

- Most breeding programs select a fixed proportion of those males and females having the best traits (top 20% for fly bristles), instead of taking individuals, whose trait $\geq \alpha$. We need to describe how to **compute** the α that corresponds to a given proportion, which we now do. We start by making the dependence on α explicit, and so repeat some of what we just did.
- Given any $\alpha \in \mathbb{R}$ (α may be positive or negative), let $A(\alpha)$ denote the **area** to the right of α the midparent deviation distribution

$$\begin{aligned} A(\alpha) &= \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\ &= \int_{\alpha/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz \quad \text{where } z = \frac{x}{\sigma} \text{ denotes Z-score of } x \end{aligned}$$

$r =$ **proportion** of midparent deviation area to right (for pos selection), 20% for fly bristles

$A^{-1}(r) = \alpha$ nonanalytic inverse function of A , such that $A(\alpha) = r$ (computer evaluation)



$$S(\alpha) = \frac{1}{A(\alpha)\sqrt{2\pi}\sigma} \cdot \int_{\alpha}^{\infty} x \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \quad \text{selection differential when selecting individuals } \geq \alpha$$

$$\begin{aligned} &= \frac{\int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} x \exp\left(\frac{-x^2}{2\sigma^2}\right) dx}{\int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx} \\ &= \frac{\int_{\alpha}^{\infty} x \exp\left(\frac{-x^2}{2\sigma^2}\right) dx}{\int_{\alpha}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx} \end{aligned}$$

$$z = \frac{x}{\sigma} \quad dz = \frac{dx}{\sigma}$$

$$x = \sigma z \quad dx = \sigma dz$$

$$\begin{aligned} S(\alpha) &= \frac{\int_{\alpha/\sigma}^{\infty} \sigma z \exp\left(\frac{-z^2}{2}\right) \sigma dz}{\int_{\alpha/\sigma}^{\infty} \exp\left(\frac{-z^2}{2}\right) \sigma dz} \\ &= \sigma \cdot \left\{ \frac{\int_{\alpha/\sigma}^{\infty} z \exp\left(\frac{-z^2}{2}\right) dz}{\int_{\alpha/\sigma}^{\infty} \exp\left(\frac{-z^2}{2}\right) dz} \right\} \end{aligned}$$

$$i\left(\frac{\alpha}{\sigma}\right) = \sigma \cdot \left\{ \frac{\int_{\alpha/\sigma}^{\infty} z \exp\left(\frac{-z^2}{2}\right) dz}{\int_{\alpha/\sigma}^{\infty} \exp\left(\frac{-z^2}{2}\right) dz} \right\}$$

$$S(\alpha) = \sigma \cdot i\left(\frac{\alpha}{\sigma}\right)$$

is **intensity of selection**, as a function of α/σ

- Now we describe the **intensity of selection**, as a function of proportion, as we set out to do.

$r =$ **proportion** of midparent deviation area to right (for pos selection), 20% for fly bristles

$A^{-1}(r) = \alpha$ nonanalytic inverse function of A , such that $A(\alpha) = r$ (computer evaluation)

$$S(A^{-1}(r)) = \sigma \cdot i\left(\frac{\alpha}{\sigma}\right) = \sigma \cdot i\left(\frac{A^{-1}(r)}{\sigma}\right)$$

$i(r)$ is **intensity of selection**, as a function of **proportion** r (in book, r called p)

- Summarizing this section, if r is the proportion of high-scoring individuals selected, $i(r)$ the intensity of selection from figure below, and σ is the population standard deviation of midparent deviations, then then

$$S = \sigma \cdot i(r) \quad \text{selection differential}$$

$$R = h^2 \cdot \sigma \cdot i(r) \quad \text{reponse to selection}$$

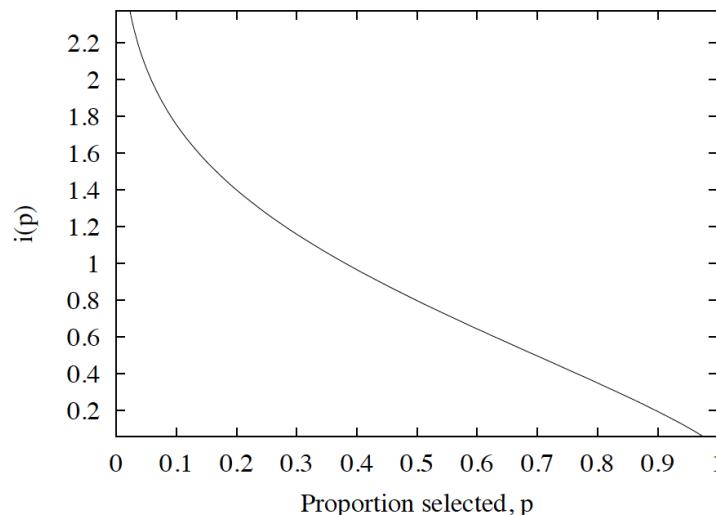


Figure 6.6: The intensity of selection, i , as a function of the proportion of individuals used as parents, p .

Dominance

- **Goal:** Prove **Fisher's result** that the genetic contribution to traits affected by many **di-allelic loci, each of small effect**, is approximately a **normal distribution**
- phenotype P depends on additive contribution from each parent (X_m maternal, X_p paternal contribution), together with the **dominance relationship** X_{mp} between maternal and paternal contribution, together with environment \mathcal{E}

$$P = X_m + X_p + X_{mp} + \mathcal{E}$$

- **Imprinted gene:** expression is determined by the parent who contributed the allele – for example, placental growth controlled by paternal allele to the detriment of the female bearing a child. Note that imprinted genes violate the usual rule of inheritance, in which both alleles in a heterozygote are equally expressed.
- **additive effects** will be chosen to maximize contribution to phenotype, while ensuring that additive effects are **uncorrelated** to dominance effects
- **genotypic value:** contribution of **genotype** to phenotype
- **additive value:** contribution of **allele A_1 or A_2** to phenotype

| | | | |
|-----------------|-------------|-----------------------|-------------|
| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
| frequency | p^2 | $2pq$ | q^2 |
| genotypic value | $a_{1,2}$ | $a_{1,2}$ | $a_{2,2}$ |
| additive value | $2\alpha_1$ | $\alpha_1 + \alpha_2$ | $2\alpha_2$ |

- genotypic and additive contribution to phenotype is in terms of **deviation from mean**, so **expected genotypic contribution** is zero

$$E_G = p^2 a_{1,1} + 2pqa_{1,2} + q^2 a_{2,2} = 0 \quad \text{expected value for genotypic contribution to phenotype}$$

$$\begin{aligned} E_A &= p^2(2\alpha_1) + 2pq(\alpha_1 + \alpha_2) + q^2(2\alpha_2) \\ &= 2\alpha_1[p(p+q)] + 2\alpha_2[q(p+q)] = 2\alpha_1p + 2\alpha_2q = 0 \quad \text{expected value for additive contribution to phenotype} \end{aligned}$$

- Note that from the last line, it follows that $(\alpha_1p + \alpha_2q) = 0$, which we'll use in deriving the formula for V_A .
- (population average) genetic variance V_G and **additive variance** V_A satisfy the following

$$\begin{aligned} V_G &= p^2 a_{1,1}^2 + 2pqa_{1,2}^2 + q^2 a_{2,2}^2 \\ p^2 a_{1,1} + 2pqa_{1,2} + q^2 a_{2,2} &= 0 \\ p^2(2\alpha_1) + 2pq(\alpha_1 + \alpha_2) + q^2(2\alpha_2) &= 0 \end{aligned}$$

Explanation: genotypic contributions to phenotype are actually stochastic values (random variables), $X_{1,1}$, $X_{1,2}$, $X_{2,2}$, respectively with expectations $a_{1,1}$, $a_{1,2}$ and $a_{2,2}$ and variances $a_{1,1}^2$, $a_{1,2}^2$ and $a_{2,2}^2$. Similarly, additive contributions to phenotype are actually stochastic values (random variables), $Y_{1,1}$, $Y_{1,2}$, $Y_{2,2}$, respectively with expectations $(2\alpha_1)$, $(\alpha_1 + \alpha_2)$, $(2\alpha_2)$ and variances $(2\alpha_1)^2$, $(\alpha_1 + \alpha_2)^2$, $(2\alpha_2)^2$.

$$\begin{aligned} V_G &= p^2 a_{1,1}^2 + 2pqa_{1,2}^2 + q^2 a_{2,2}^2 \\ V_A &= p^2(2\alpha_1)^2 + 2pq(\alpha_1 + \alpha_2)^2 + q^2(2\alpha_2)^2 \\ &= 2\alpha_1^2 p^2 + 2\alpha_1^2 p^2 + 2\alpha_1^2 pq + 4pq\alpha_1\alpha_2 + 2\alpha_2^2 pq + 2\alpha_2^2 q^2 + 2\alpha_2^2 q^2 \\ &= 2\alpha_1^2[p^2 + pq] + 2\alpha_2^2[q^2 + pq] + 2[\alpha_1^2 p^2 + 2pq\alpha_1\alpha_2 + \alpha_2^2 q^2] \\ &= 2\alpha_1^2 p + 2\alpha_2^2 q + 2(\alpha_1 p + \alpha_2 q)^2 = 2\alpha_1^2 p + 2\alpha_2^2 q + 2 \cdot 0 = 2(\alpha_1^2 p + \alpha_2^2 q) \end{aligned}$$

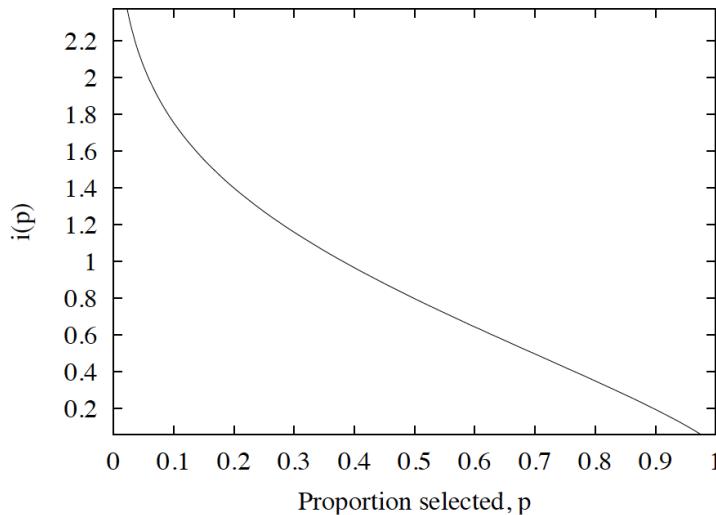


Figure 6.6: The intensity of selection, i , as a function of the proportion of individuals used as parents, p .

- **Goal:** Determine α_1, α_2 in terms of $p, q, a_{1,1}, a_{1,2}, a_{2,2}$ by using the method of **least squares** – let the function $F(\alpha_1, \alpha_2)$ represent the **expected sum of squared distances** between genotypic values and additive values

$$F(\alpha_1, \alpha_2) = p^2(a_{1,1} - 2\alpha_1)^2 + 2pq(a_{1,2} - [\alpha_1 + \alpha_2])^2 + q^2(a_{2,2} - 2\alpha_2)^2$$

- As in our derivation of the (linear) regression equation, compute partial derivatives of F , set these equal to 0, and solve.

$$\begin{aligned}\frac{\partial F(\alpha_1, \alpha_2)}{\partial \alpha_1} &= -4 \left[p^2(a_{1,1} - 2\alpha_1) + pq(a_{1,2} - \alpha_1 - \alpha_2) \right] = 0 \\ \frac{\partial F(\alpha_1, \alpha_2)}{\partial \alpha_2} &= -4 \left[q^2(a_{2,2} - 2\alpha_2) + pq(a_{1,2} - \alpha_1 - \alpha_2) \right] = 0\end{aligned}$$

- we have 2 equations in 2 unknowns, so by elimination of variables, we have

$$\begin{aligned}\alpha_1 &= pa_{1,1} + qa_{1,2} \\ \alpha_2 &= qa_{2,2} + pa_{1,2}\end{aligned}$$

so

$$\begin{aligned}V_A &= 2pq \left[p(a_{1,1} - a_{1,2}) + q(a_{1,2} - a_{2,2}) \right]^2 \\ V_D &= V_G - V_A = p^2q^2(a_{1,1} - 2a_{1,2} + a_{2,2})^2 \\ V_G &= V_A + V_D\end{aligned}$$

- Recall that the **covariance** between random variables X, Y is

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

In our case, X [resp. Y] is the **genotypic** value of the first [resp. second] parent. Since genotypic values are **deviations** from the mean, their expectation is 0, which we've used before. Given n pairs of parents, it follows that covariance is given by

$$\begin{aligned}Cov(X, Y) &= E[XY] - E[X]E[Y] = \sum_i \sum_j p_{i,j}x_iy_j - (\sum_i p_i x_i) \cdot (\sum_j p_j y_j) \\ &= \sum_i \sum_j p_{i,j}x_iy_j - 0 \cdot 0 = \sum_i \sum_j p_{i,j}x_iy_j\end{aligned}$$

The following table is used for this computation.

| Genotypes | | Values | | Frequency |
|-----------|----------|----------|----------|-----------------------------|
| X | Y | X | Y | |
| A_1A_1 | A_1A_1 | a_{11} | a_{11} | $p^2[r_0p^2 + r_1p + r_2]$ |
| A_1A_1 | A_1A_2 | a_{11} | a_{12} | $p^2[r_02pq + r_1q]$ |
| A_1A_1 | A_2A_2 | a_{11} | a_{22} | $p^2[r_0q^2]$ |
| A_1A_2 | A_1A_1 | a_{12} | a_{11} | $2pq[r_0p^2 + r_1p/2]$ |
| A_1A_2 | A_1A_2 | a_{12} | a_{12} | $2pq[r_02pq + r_1/2 + r_2]$ |
| A_1A_2 | A_2A_2 | a_{12} | a_{22} | $2pq[r_0q^2 + r_1q/2]$ |
| A_2A_2 | A_1A_1 | a_{22} | a_{11} | $q^2[r_0p^2]$ |
| A_2A_2 | A_1A_2 | a_{22} | a_{12} | $q^2[r_02pq + r_1p]$ |
| A_2A_2 | A_2A_2 | a_{22} | a_{22} | $q^2[r_0q^2 + r_1q + r_2]$ |

Table 6.6: All possible pairs of relatives needed for the calculation of the covariance between relatives.



$$\begin{aligned}
 Cov(X, Y) &= \left(\sum_{i=1}^2 \sum_{j=1}^2 \right) \left(\sum_{k=1}^2 \sum_{\ell=1}^2 \right) Prob[\text{genotype 1 is } A_iA_j \text{ genotype 2 is } A_kA_\ell] a_{i,j} a_{k,\ell} \\
 &= \{p^2[r_0p^2 + r_1p + r_2] \cdot a_{1,1}a_{1,1}\} + \{p^2[r_02pq + r_1q] \cdot a_{1,1}a_{1,2}\} + \dots
 \end{aligned}$$

- In $\text{Cov}(X, Y)$, group together terms involving r_0 . This yields

$$r_0 \cdot (p^2 a_{1,1} + 2pqa_{1,2} + q^2 a_{2,2})^2$$

Recalling that

$$E_H = p^2 a_{1,1} + 2pqa_{1,2} + q^2 a_{2,2} = 0$$

it follows that there is zero contribution from the factor of r_0 above. The coefficient of r_1 is

$$p\alpha_1^2 + q\alpha_2^2 = \frac{V_A}{2}$$

The coefficient of r_2 is

$$p^2 a_{1,1}^2 + 2pqa_{1,2}^2 + q^2 a_{2,2}^2 = V_G$$

so

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{r_1}{2} \cdot V_A + r_2 \cdot V_G V_G &= V_A + V_D \\ \text{Cov}(X, Y) &= \frac{r_1}{2} \cdot V_A + r_2 \cdot V_A + r_2 \cdot V_D \\ &= (\frac{r_1}{2} + r_2) V_A + r_2 V_D = rV_A + r_2 V_D \end{aligned}$$

where we have used the **coefficient of relatedness** r from page 123 of the book, defined as follows

$$\begin{aligned} \bar{r} &= 0 \cdot r_0 + 1 \cdot r_1 + 2 \cdot r_2 & \bar{r} \text{ is expected number of shared alleles by descent} \\ &= r_1 + 2r_2 \end{aligned}$$

$$r = \frac{\bar{r}}{2} \quad \text{definition of coefficient of relatedness}$$

■ Explanation of line 5:

| genotype 1 A_1A_2 | genotype 2 A_1A_2 | genotypic value 1 $a_{1,2}$ | genotypic value 2 $a_{1,2}$ | probability $2pq(r_02pq + r_1/2 + r_2)$ |
|------------------------|------------------------|--------------------------------|--------------------------------|--|
|------------------------|------------------------|--------------------------------|--------------------------------|--|



rows 1-3 \Rightarrow offspring genotype A_1A_1

rows 4-6 \Rightarrow offspring genotype A_1A_2

rows 7-9 \Rightarrow offspring genotype A_2A_2

left factor \Rightarrow first parent probability in population at Hardy-Weinberg equilibrium

right factor \Rightarrow second parent probability, taking into account possible alleles shared with first parent by descent, and of

● derivation of line 5:

- ▷ $2pq$ is probability first parent genotype is A_1A_2
- ▷ r_02pq : r_0 is probability that second parent (a relative of first) shares zero alleles by descent with first. Multiply r_0 by $2pq$, which is probability that second parent is A_1A_2 chosen at random from HW-equilibrium
- ▷ $r_1/2$: r_1 is probability that second parent (a relative of first) shares one allele by descent with first. The shared allele can be either A_1 or A_2 , each with probability $\frac{1}{2}$. If the shared allele is A_1 , then the probability that second parent also has allele A_2 is q , the relative frequency of allele A_2 in the population. If the shared allele is A_2 , then the probability that second parent also has allele A_1 is p , the relative frequency of allele A_1 in the population. Hence the probability that second parent has genotype A_1A_2 , assuming that he/she shares one allele from first, is equal to $\frac{q}{2} + \frac{p}{2} = \frac{1}{2}$. Finally, the probability contributed in this case is $r_1 \cdot \frac{1}{2}$.
- ▷ r_2 : r_2 is probability that second parent (a relative of first) shares two alleles by descent with first, whose genotype is A_1A_2 . In this case, the second parent has genotype A_1A_2 with probability 1, so the probability contributed in this case is $r_2 \cdot 1$.

Concluding remarks on Chapter 6 - Dominance

- By least squares, we have computed V_D in terms of $p, q, a_{1,1}, a_{1,2}, a_{2,2}$. Method of least squares ensures that V_D is not correlated with V_A .

$$P = X_m + X_p + X_{m,p} + \mathcal{E}$$
$$V_P = V_A + V_D + V_E$$

- Until now, we've focused on one gene; however, if we assume genes are independent (multiplicative epistasis), then

$$P = \sum_i X_m(i) + X_p(i) + X_{m,p}(i) + \mathcal{E}$$
$$V_A = \sum_i V_A(i)$$

- Central Limit Theorem asserts that if X_i are independent, identically distributed random variables, with finite variances, the the sum tends toward a normal distribution. Explicitly, if

$$S_n = \sum_{i=1}^n X_i$$
$$\lim_{n \rightarrow \infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim \mathcal{N}(0, 1)$$

- In the situation at hand, it infers that the sum of many independent genes that contribute to a given phenotype, then the phenotypic values tend to be normally distributed in the population – this result was first proved by **Fisher**.

■ **narrow sense heritability:**

$V_P = V_A + V_D + V_I + V_E$ V_I is additional variance due to interaction between genes in non-multiplicative epistasis

$$h^2 = \frac{V_A}{V_P}$$

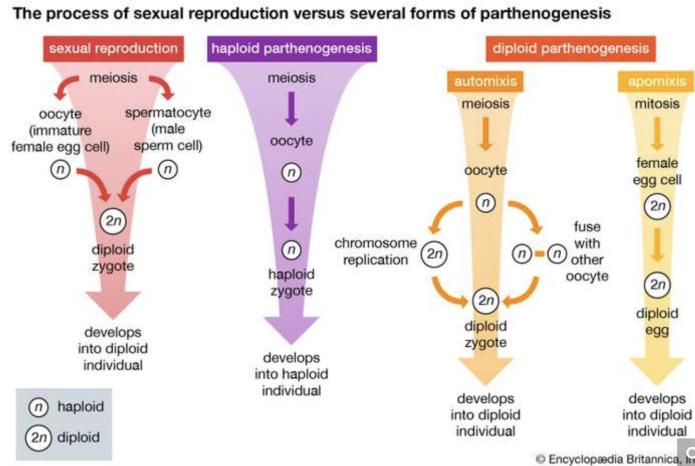
■ **broad sense heritability:**

$$H^2 = \frac{V_A + V_D + V_I}{V_P}$$



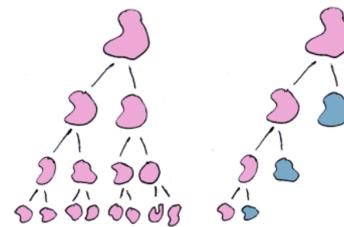
Chapter 7: Sexual reproduction

Parthenogenesis



- **Parthenogenesis** means the development of an embryo from an egg cell without fertilization (in contrast to self-fertilization or selfing from Chapter 5).
- Various forms of **parthenogenesis**
 - ▷ **haploid parthenogenesis:** development of haploid embryo from the (usual) haploid egg of a diploid (or polyploid) individual – examples are in *Lilium spp.* (lillies), *Orchis maculata* (purple orchid), *Nicotiana tabacum* (tobacco plant), etc.
 - ▷ **diploid parthenogenesis:** development of diploid embryo from meiotically unreduced gametophyte.
- **apomixis:** usually means production of a meiotically unreduced gametophyte.
- **endomitosis:** replication of chromosomes in the absence of cell or nuclear division (e.g. in salivary glands of *Drosophila*).

- In Chapter 5, we saw that self-fertilizing populations have half the genetic load of outcrossing populations.
- In this section, we will see that
 - ▷ the disadvantages of sexual reproduction over parthenogenesis include a **twofold cost** of sexual reproduction over parthenogenesis



- the advantages of sexual reproduction over parthenogenesis include
 - ▷ sexually reproducing populations require **less time to fix** advantageous mutations
 - ▷ sexually reproducing populations have more efficient removal of **deleterious mutations** using **Muller's ratchet**

2-fold cost of sexual reproduction

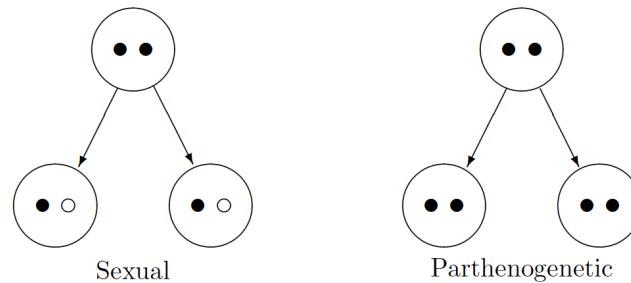


Figure 7.1: A comparison of the consequences of sexual versus parthenogenetic reproduction. The female parent is illustrated in both cases. The filled circles in the offspring come from the mother; the father's genetic contribution is indicated by the open circles in the sexual species.

- if a female diploid organism produces on average 2 offspring, then parthogenesis ensures twice as much genetic material will be passed on than sexual reproduction as shown in figure above

Advantage caused by segregation

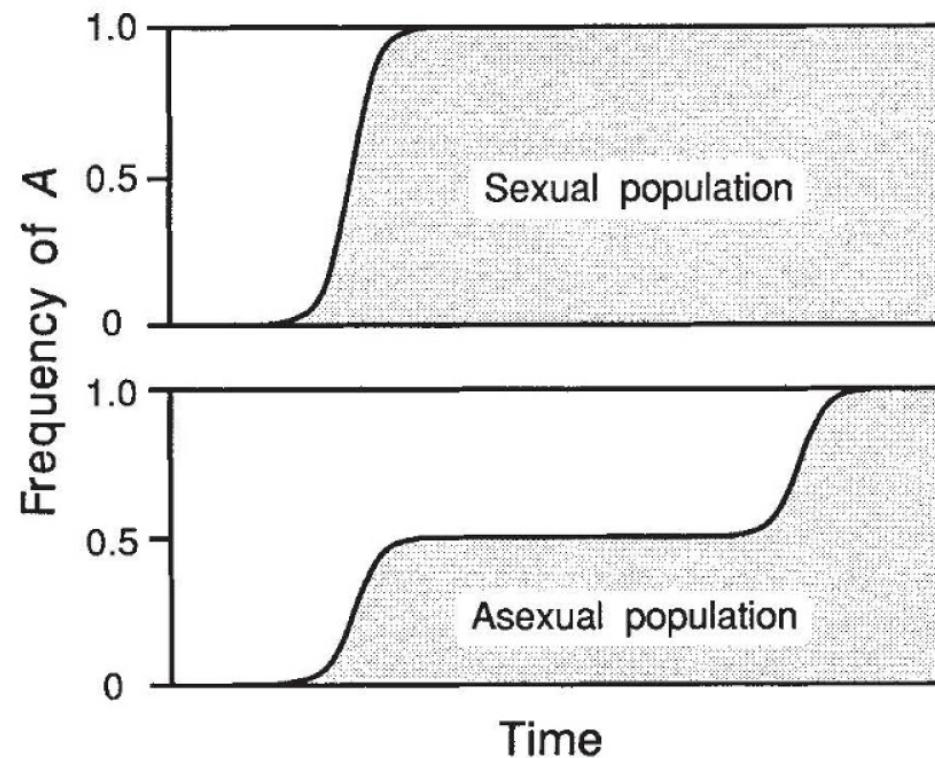


FIG. 1 Schematic illustration of the fates of an advantageous mutation A spreading through sexual and asexual diploid populations. The locus in the asexual population is fixed in a heterozygous state until a mutation appears at the second allele. The second mutation takes on average twice as long to appear, as only one allele is available to mutate in the heterozygote.

(Image from “Genetic segregation and the maintenance of sexual reproduction”, Nature 339 (1989) 300-301)

- Substitution of a new advantageous mutation in a parthenogenetic species requires **two mutations** in a lineage, while substitution in a sexual species requires only **one mutation**.
- Advantageous alleles can sweep through a sexually reproducing population much faster than in a population practicing parthenogenesis due both to **segregation** and **crossover**. Figure on previous slide depicts the reason for the advantage due to segregation.
- Mathematical argument for faster sweep of a new advantageous mutation in a sexually reproducing population depends on **fixation probability**, reviewed in the next slide

Fixation probability of a new advantageous mutation

- At the bottom of page 41 in Chapter 4 notes, we showed that for the h, s fitness model, if heterozygosity effect $h = \frac{1}{2}$, then

$$\pi_1(p) = \frac{1 - e^{-2Ns p}}{1 - e^{-2Ns}}$$
$$\pi_1\left(\frac{1}{2N}\right) = \frac{1 - e^{-s}}{1 - e^{-2Ns}} = \frac{1 - (1 - s + \frac{s^2}{2!} - \frac{s^3}{3!} + \dots)}{1 - e^{-2Ns}} \approx \frac{s}{1 - e^{-2Ns}} \approx s \quad (\text{if } N \text{ is large})$$

This result can be generalized for arbitrary $0 < h < 1$ to the following:

$$\pi\left(\frac{1}{2}\right) \approx 2s(1 - h)$$

- The **selective advantage of heterozygotes** over A_2A_2 homozygotes is defined to be the difference of heterozygote fitness and A_2A_2 homozygote fitness:

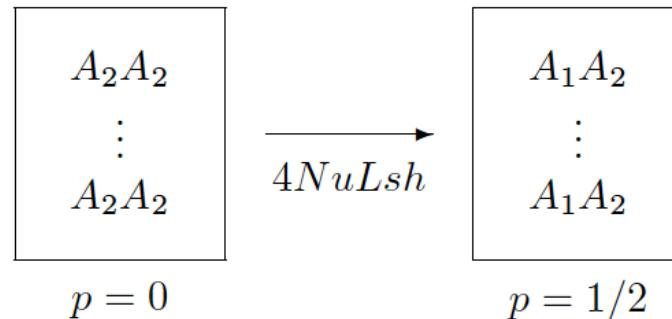
$$(1 - hs) - (1 - s) = s - sh = s(1 - h)$$

- Thus the fixation probability of a new **advantageous** mutation is equal to twice the **selective advantage** of heterozygotes over A_2A_2 homozygotes, or in Gillespie's words:

"The probability of not being lost due to the action of genetic drift is twice the selective advantage of the heterozygote."

We'll assume the result above without proof (since it depends on the Kolmogorov forward equation) – see Section 3.9 of Gillespie's text for more discussion.

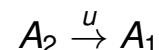
Transformation from homozygous A_2A_2 to heterozygous A_1A_2



- Goal of this section is to provide a mathematical argument that fixation of new advantageous mutations occurs much more slowly in parthenogenesis than for sexual reproduction, a result due to Kirkpatrick and Jenkins (1989).
- This result requires a slight modification of the h, s fitness model given as follows:
- As in the usual h, s model, A_1A_1 homozygotes are fitter than A_1A_2 heterozygotes (provided $0 < h < 1$), which are fitter than A_2A_2 homozygotes. However, instead of respective fitnesses $1, 1 - hs$ and $1 - s$, the new model has fitnesses

| genotype | A_1A_1 | A_1A_2 | A_2A_2 |
|----------|----------|----------|----------|
| fitness | $1 + s$ | $1 + hs$ | 1 |

- Suppose A_1 is new advantageous mutation which arises with mutation rate u in an otherwise homozygous A_2A_2 population, i.e.



- From previous slide, the probability of not being lost due to genetic drift is twice the selective advantage of heterozygote A_1A_2 , hence is

$$2[(1 + hs) - 1] = 2hs$$

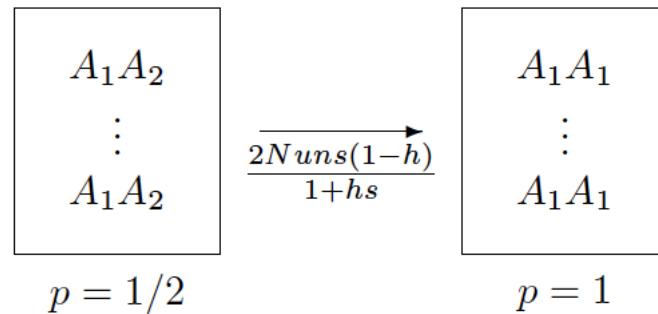
- 2Nu is number of A_1 alleles that arise due to mutation in population of size N
- The number of A_1 alleles that arise and are not lost due to genetic drift is

$$2Nu \cdot 2hs = 4Nush$$

assuming all loci are equally likely to be mutated.

- L is proportion of homozygous loci that experience this type of selection (some loci could possibly be less prone to mutation or are more likely to be “corrected” back to A_2 when mutated)
- 4NushL is number of A_1 alleles that arise and are not lost due to genetic – this is the conversion rate from homozygote A_2A_2 to heterozygote A_1A_2

Transformation from heterozygous A_1A_2 to homozygous A_1A_1



- Probability of mutation $A_2 \rightarrow A_1$ not being lost due to genetic drift, when transforming the heterozygote A_1A_2 to the homozygote A_1A_1 is twice the selective advantage of A_1A_1 over that of A_1A_2 , which equals

$$(1 + s) - (1 + hs) = s(1 - h)$$

- The fitness advantage of A_1A_1 homozygotes over A_1A_2 heterozygotes is by definition

$$(1 + s) - (1 + hs) = s(1 - h)$$

However, with no explanation, on page 172, Gillespie states with no additional definition or explanation that the selection coefficient of an A_1A_1 homozygote relative to the A_1A_2 heterozygote, call it s' , is found by solving

$$\begin{aligned} 1 + s' &= \frac{1 + s}{1 + hs} \\ s' &= \frac{1 + s}{1 + hs} - 1 = \frac{1 + s - (1 + hs)}{1 + hs} = \frac{s - hs}{1 + hs} \\ &= \frac{s(1 - h)}{1 + hs} \\ 2s' &= \frac{2s(1 - h)}{1 + hs} \end{aligned}$$

- Gillespie now implicitly assumes that the fixation probability of A_1A_1 homozygotes from a homogeneously uniform A_1A_2 heterozygote population is twice the selection coefficient, rather than twice the selective advantage as we showed a few slides ago. Why? Apparently to justify the resulting equilibrium value of

$$n^* = \frac{2hL(1 + hs)}{1 - h}$$

- Note that the selective advantage is

$$s(1 - h)$$

while the selective coefficient is

$$s(1 - h) \cdot 11 + hs$$

Generally, s is small (recall $s = 0.1$ is considered large), so there is not much difference between these values. Nevertheless, if we use the selective advantage in place of the selective coefficient, we would obtain the different equilibrium value n^* given by

$$n^* = \frac{2hL}{1 - h}$$

which differs from the value in Kirkpatrick and Jenkins (1989) by the missing factor $(1 + hs)$.

- For that reason, in the following, following Gillespie, we use the selection coefficient s' in place of selective advantage when arguing that the fixation probability of an A_1 mutation to transform an A_1A_2 heterozygous population to A_1A_1 homozygote is twice the selection coefficient.
- Note that the expected number of mutations $A_2 \rightarrow A_1$ in a completely heterozygous population is Nu , rather than $2Nu$, since only $N A_2$ alleles exist in population of size N which is entirely heterozygous.
- The expected number of mutations $A_2 \rightarrow A_1$ that arise in completely heterozygous population, and are not removed by genetic drift, is given by

$$Nu \cdot 2s' = \frac{2Ns(1 - h)}{1 + hs}$$

- If only proportion n of the sites are heterozygous, then the expected number of mutations $A_2 \rightarrow A_1$ that arise in proportion n heterozygous population, and are not removed by genetic drift, is given by

$$nNu \cdot 2s' = \frac{2nNs(1 - h)}{1 + hs}$$

- Equilibrium occurs when the expected number of loci that are homozygous A_2A_2 , which are converted to heterozygous A_1A_2 , is equal to the expected number of loci that are heterozygous A_1A_2 , which are converted to heterozygous A_1A_1

$$\begin{aligned} 4NushL &= \frac{2nuNs(1 - h)}{1 + hs} \\ n^* &= \frac{4NushL(1 + hs)}{2Nus(1 - h)} \\ &= \frac{2hL(1 + hs)}{(1 - h)} \end{aligned}$$

- Recall that multiplicative epistasis is the overall fitness is the product of the fitness of independence loci
- $W_{1,2}$ is the overall fitness, assuming multiplicative epistasis, of the A_1A_2 heterozygote at equilibrium, where $n = n^*$, i.e. the expected number n^* of A_1A_2 heterozygotes is equal to the the expected number n^* of A_1A_1 homozygotes
- $W_{1,1}$ is the overall fitness, assuming multiplicative epistasis, of the A_1A_1 homozygote at equilibrium

- $W = W_{1,1} / W_{1,2}$ is the relative fitness advantage, assuming multiplicative epitasis,

$$W_{1,2} = (1 + hs)^{n^*}$$

$$W_{1,1} = (1 + s)^{n^*}$$

$$W = \frac{W_{1,1}}{W_{1,2}} = \left(\frac{1 + s}{1 + hs} \right)^{n^*}$$

- During the time when a parthenogenetic species reaches expected number n^* of A_1A_2 heterozygotes, a sexually reproducing species reaches expected number n^* of A_1A_1 heterozygotes, so the selective advantage W_s of sexual reproduction over that of parthenogenesis is

$$W_s = \frac{W_{1,1}}{W_{1,2}} = \left(\frac{1 + s}{1 + hs} \right)^{n^*}$$

- When $W_s > 2$ the advantage to the sexual population more than compensates for its intrinsic twofold reproductive deficit, compared with parthenogenesis.

When does sexual reproduction more advantageous than parthenogenesis?

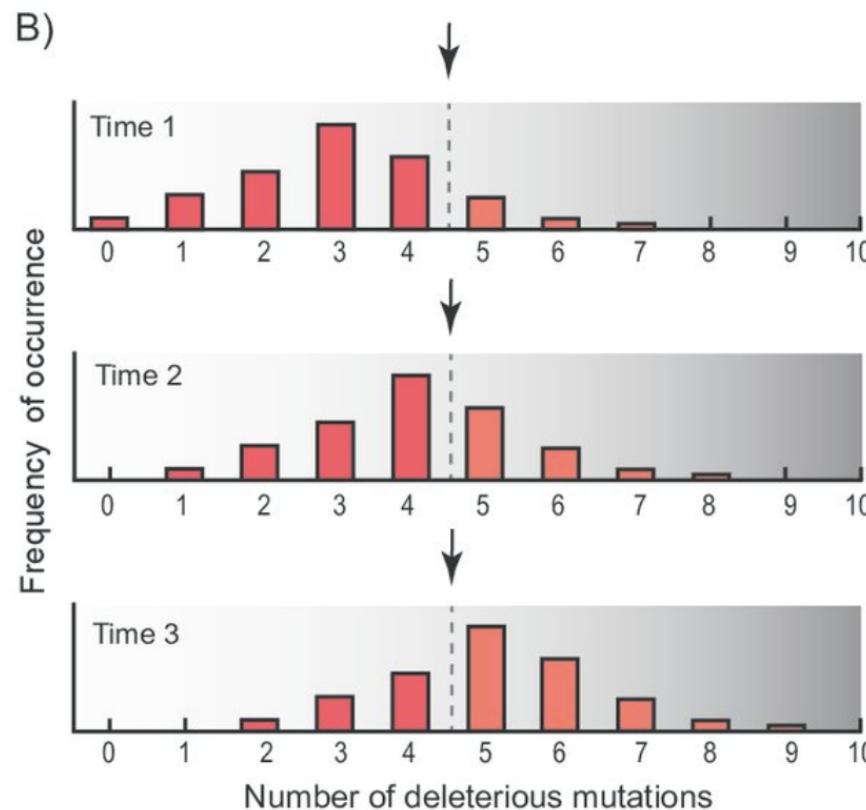
TABLE 1 Equilibrium values of W_s , the fitness of a sexual population relative to an asexual population

| $s \backslash L$ | 10 | 50 | 100 | 500 | 1,000 |
|------------------|-----|-------------------|-------------------|----------------------|----------------------|
| 0.001 | 1.0 | 1.1 | 1.1 | 1.6 | 2.7 |
| 0.005 | 1.1 | 1.3 | 1.6 | 1.2×10 | 1.5×10^2 |
| 0.01 | 1.1 | 1.6 | 2.7 | 1.5×10^2 | 2.1×10^4 |
| 0.05 | 1.6 | 1.2×10 | 1.4×10^2 | 5.3×10^{10} | 2.8×10^{21} |
| 0.1 | 2.7 | 1.3×10^2 | 1.7×10^4 | 1.6×10^{21} | 2.7×10^{42} |

Values below the heavy line exceed 2, indicating a net fitness advantage to the sexual population. Calculations assume $h=1/2$.

Image from "Genetic segregation and the maintenance of sexual reproduction", Nature 339 (1989) 300-301)

Sex removes deleterious alleles: Muller's ratchet



(Image from “Resolving genome evolution patterns in asexual plant”, Hojsgaard et al. (2014), in book **Next-Generation Sequencing in Plant Systematics**, Eds: E. Mörandl and M. Appelhans, Publisher: IAPT)

- Muller's ratchet: depending on number of gametes in population of size N , which have $i = 0, 1, 2, \dots$ many deleterious alleles, there is a gradual increase in the number of deleterious alleles in the parthenogenetic population over time

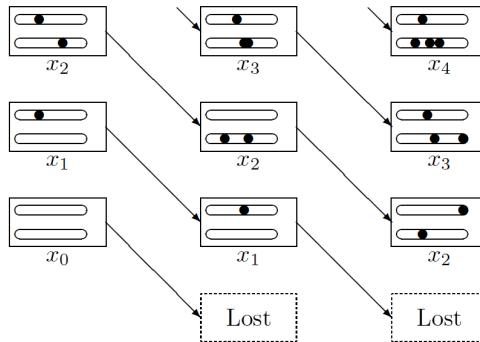


Figure 7.4: An illustration of two clicks of Muller's ratchet. The boxes represent the class of individuals with i deleterious mutations. The two chromosomes within each box represent typical individuals. The number of generations between each column of boxes depends on the efficacy of genetic drift.

- evolutionary model: mutation, selection, genetic drift
- mutation rate is very small, so will be assume to always occur in heterozygous form
- let x_i denote the proportion of the population, having size N , of individuals having exactly i mutations
- $x_0, x_1, \dots \in [0, 1]$
- Poisson distribution used to model the number of events, provided that events are rare and that the population is large.
- Poisson distribution is a finite approximation to the normal distribution, which however, tends to have a long right tail. If mean is μ , then Poisson probability of i events is

$$e^{-\mu} \cdot \frac{\mu^i}{i!}$$

- Mutations are rare events, and genome sizes are large, so it is reasonable to model the probability that there are i mutations in a gamete, where the average number of mutations in the population is μ by

$$x_i = \text{Prob}[i \text{ mutations in current population}] = \text{Poisson}(\mu, i)$$

$$= e^{-\mu} \cdot \frac{\mu^i}{i!}$$

- assume that h, s are constant, regardless of loci
- assume multiplicative epistasis, so individual having i mutations has fitness $(1 - hs)^i$
- assume offspring in next generation have a Poisson distributed number of new mutations (not counting the inherited mutations), with average U

$$\text{Prob}[j \text{ new mutations}] = \text{Poisson}(U, j) = e^{-U} \cdot \frac{U^j}{j!}$$

- let x'_k denote the proportion of individuals having k mutations, where $k = i + j$, and the individual inherited i mutations and has acquired j new mutations, where $0 \leq i, j \leq k$

$$\begin{aligned} \text{Prob}[i \text{ inherited mutations}, j \text{ new mutations}] &= \text{Prob}[i \text{ inherited mutations}] \cdot \text{Prob}[j \text{ new mutations}] \\ &= e^{-\mu} \cdot \frac{\mu^i}{i!} \cdot e^{-U} \cdot \frac{U^j}{j!} \\ &= e^{-(\mu+U)} \cdot \frac{\mu^i U^j}{i! j!} \end{aligned}$$

- Muller's ratchet:** depending on the proportion x_i of gametes in population of size N , that have $i = 0, 1, 2, \dots$ many deleterious alleles, there is a gradual increase in the number of deleterious alleles in the parthenogenetic population over time

Kondrashov's ratchet

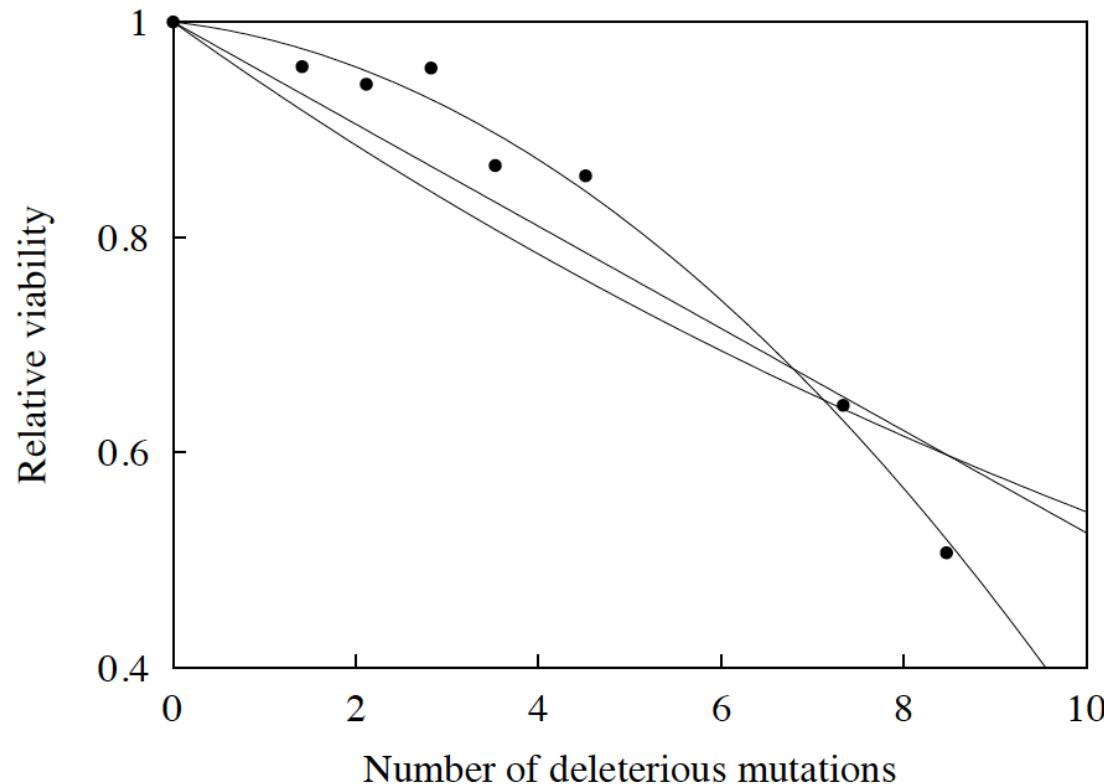


Figure 7.7: The relative viability as a function of the inferred number of homozygous deleterious mutations in *Drosophila melanogaster*. The upper concave curve is a quadratic synergistic model, the middle straight line is an additive model, and the lower convex curve corresponds to multiplicative epistasis. The data are from Mukai (1968).

- Curves above correspond to 3 models of epistasis for n deleterious mutations, where w_n is the fitness of n deleterious mutations, and s is the (constant) selection coefficient
- multiplicative epistasis

$$w_n = (1 - s)^n$$

$s = 0.5897$ estimated with least squares curve fitting

- additive epistasis

$$w_n = (1 - s)^n$$

$s = 0.05897$ estimated with least squares curve fitting

- quadratic synergistic epistasis

$$w_n = 1 - sn - an^2$$

$$s = 0.009813$$

$a = 0.00555$ s, a estimated with least squares curve fitting

- Kondrashov considered truncation selection

$$w_n = \begin{cases} 1 & \text{if } n \leq k \\ 0 & \text{if } n > k \end{cases}$$

- In each generation, the probability that a given offspring receives $X = k$ deleterious mutations follows a Poisson-distribution with mean U

$$\text{Prob}[X = k] = e^{-U} \cdot \frac{U^k}{k!}$$

- Starting from a population with no deleterious mutations, after a “burn-in” time during which individuals gradually accumulate up to k deleterious mutations with no effect on their fitness, the entire population has exactly k deleterious mutations.

- From that point on, any offspring with additional mutations will die, so the probability that a given individual will not receive any deleterious mutations (hence will survive) is

$$\begin{aligned} \text{Prob}[X = k] &= e^{-U} \cdot \frac{U^k}{k!} \\ \text{Prob}[X = 0] &= e^{-U} \end{aligned}$$

so it follows that average fitness \bar{w} and genetic load L satisfy

$$\begin{aligned} \bar{w} &= e^{-U} \cdot 1 + (1 - e^{-U}) \cdot 0 = e^{-U} \\ L &= \frac{w_{\max} - \bar{w}}{w_{\max}} & \frac{1 - \bar{w}}{1} = e^{-U} \end{aligned}$$

- In Chapter 3, we saw that genetic load of a sexually reproducing organism was also

$$L = e^{-U}$$

provided that multiplicative epistasis is assumed

- Note that for asexual reproduction, multiplicative epistasis is not assumed.
- If x_0 denotes the proportion of individuals having 0 mutations, and x'_0 denotes the proportion of individuals in the next generation having 0 mutations, then

$$x'_0 = \frac{x_0 w_0 e^{-U}}{\bar{w}}$$

- At equilibrium, $x_0 = x'_0$, and since $w_0 = 1$, we have

$$\begin{aligned} x_0 &= x'_0 = \frac{x_0 w_0 e^{-U}}{\bar{w}} \\ 1 &= \frac{1 \cdot e^{-U}}{\bar{w}} \\ \bar{w} &= e^{-U} \end{aligned}$$

- This last derivation makes no assumption about the type of epistasis, but only assumes that the number of deleterious mutations is Poisson distributed with mean U
- Let h^2 denote the hereditability of a trait, defined by

$$h^2 = \frac{V_A}{V_P} = \frac{V_A}{V_A + V_E}$$

where V_A denotes the additive variance, which is the sum $V_A = V_m + V_p$ of the variance due to maternal and paternal sources, where V_E denotes the variance due to the environment, and $V_P = V_A + V_E$ is the variance in the population (due to all causes).

- There is no environmental contribution to deleterious mutations, so $h^2 = 1$ in Kondrashov's threshold selection model.
- Chapter 6 defines the selection differential S to be the deviation of the average value of the (quantitative) trait of both parents from the population mean

$$R = h^2 S$$

where R is the (quantitative) phenotypic response to the selection differential S

- In each generation, the average number of mutations per individual is increased by U , and the average number of mutations per individual is decreased by R , so at equilibrium

$$U = R = h^2 S = 1 \cdot S = S$$

- From Chapter 6, the selection intensity $i(p)$

$$i(p) = \frac{S}{\sqrt{V_P}}$$

$$U = S$$

$$i(p) = \frac{U}{\sqrt{V_P}}$$

U = average number of deleterious mutations per individual (Poisson distributed)

$\sqrt{V_P}$ = standard deviation of total number of deleterious mutations per individual

- Kondrashov determined that for *D. melanogaster*, $U = 2$ and $\sqrt{V_P} = 10$, so in the case of (sexually reproducing) flies, selection intensity $i(p)$ satisfies

$$i(p) = \frac{2}{10} = 0.2$$

which corresponds in Figure 6.6 to about 10% of fly population dying, i.e. a genetic load of

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}} = \frac{1 - 0.9}{1} = 0.1$$

- If flies were instead to produce asexually, then genetic load

$$L = 1 - e^U = 1 - e^{-2} \approx (1 - 0.1353) = 0.8647$$

- Since the ratio of asexual to sexual genetic load is greater than 2,

$$\frac{L_{\text{asexual}}}{L_{\text{sexual}}} = \frac{0.8647}{0.1} = 8.657 > 2$$

and the ratio of sexual to asexual average fitness is greater than 2,

$$\frac{\bar{w}_{\text{sexual}}}{\bar{w}_{\text{asexual}}} = \frac{0.9}{0.1353} \approx 6.6519 > 2$$

it follows that the double cost of sexual reproduction is justified



Probability Appendix

Some probability distributions

- Expectation $E[X]$ of discrete random variable X with **probability mass function** $Prob[X = k]$ and values among $\dots - 3, -2, -1, 0, 1, 2, 3 \dots$

$$E[X] = \sum_k k \cdot Prob[X = k]$$

- Expectation $E[X]$ of continuous random variable X with **probability density function (PDF)** $p(x)$ and values in \mathbb{R}

$$E[X] = \int_{-\infty}^{\infty} x \cdot p(x) dx$$

- Variance $V[X]$ of random variable X with mean $\mu = E[X]$

$$V[X] = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2$$

- Theorem 1:** Expectation is additive

$$\begin{aligned} E[X + Y] &= E[X] + E[Y] \\ E[c \cdot X] &= c \cdot E[X] \end{aligned}$$

- Theorem 2:** Variance is additive, provided random variables are independent – otherwise must take into account the covariance between random variables (explained later)

$$\begin{aligned} V[X + Y] &= V[X] + V[Y] \\ V[c \cdot X] &= c^2 \cdot V[X] \end{aligned}$$

Bernouilli and binomial distributions

- Bernouilli random variable with success probability p , i.e. single coin flip with heads probability p

$$\text{Prob}[X = 1] = p$$

$$\text{Prob}[X = 0] = 1 - p$$

$$E[X] = p \cdot 1 + (1 - p) \cdot 0 = p$$

$$E[X^2] = p \cdot 1^2 + (1 - p) \cdot 0^2 = p$$

$$V[X] = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - E[X]^2 = p - p^2 = p(1 - p) = pq$$

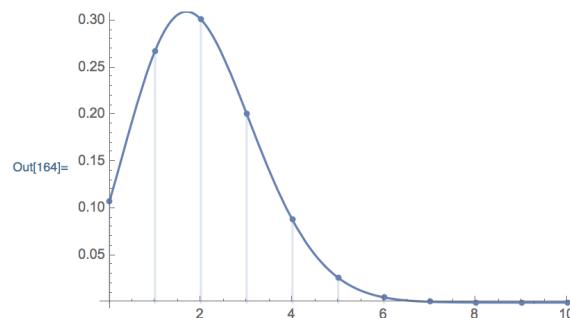
- Binomial random variable with parameters p, n , i.e. number of heads flipped in n coin flips with heads probability p

$$\text{Prob}[X = k] = \binom{n}{k} \cdot p^k \cdot q^{n-k}$$

$$E[X] = E[X + \dots + X] = E[X] + \dots + E[X] = nE[X] = np$$

$$V[X] = V[X + \dots + X] = V[X] + \dots + V[X] = nV[X] = npq$$

```
In[160]:= n = 10; p = 2/10;
binomDist = PDF[BinomialDistribution[n, p], k];
G1 = DiscretePlot[binomDist, {k, 0, 10}, Filling -> Axis];
G2 = Plot[binomDist, {k, 0, 10}];
Show[{G1, G2}]
```



Hypergeometric distribution

- hypergeometric probability $\text{Prob}[X = k]$ of drawing k red balls in a sample of n balls, drawn from an urn having R red balls, $N - R$ black balls, so that the urn contains a total of N balls.

$$\text{Prob}[X = k | n \text{ balls drawn without replacement}] = \frac{\binom{R}{r} \cdot \binom{N-R}{n-r}}{\binom{N}{n}}$$

- Note: the probability of getting k heads when flipping a coin n times, whose heads probability is p , is **equal** to the probability of drawing k red balls in a sample of n balls, when drawing n balls from a large urn containing R red balls and $N - R$ black balls (hence N balls altogether), where $p = \frac{R}{N}$, where after drawing each ball, the drawn ball is put back into the urn so that for the next draw, the probability of drawing a red ball is still p . This is known as **drawing with replacement**. Thus

$$\text{Prob}[X = k | n \text{ balls drawn with replacement}] = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

is the probability of k heads when flipping a coin n times, whose heads probability is p , and **equivalently** is as well the probability of drawing k red balls in a sample of n balls, when drawing balls from an urn having $p = \frac{R}{N}$ proportion of red balls, provided that balls are drawn **with replacement**.

- hypergeometric distribution is used for **sampling without replacement**, in contrast to the binomial distribution, which is used for **sampling with replacement**

Mean and variance in Wright-Fisher model

- Question: In the **Wright-Fisher model** for a diploid, monoecious, hermaphroditic species, population size is restricted to be a constant value N . Suppose that the proportion of A_1 alleles at a given locus in the current population is $p = \frac{k}{2N}$, while that for A_2 alleles is $q = \frac{2N-k}{2N}$. If we **sample** $2N$ gametes from the current population **with replacement**, then what is the **expected** proportion of A_1 alleles in the sampled population, and what is the **variance** in the proportion of A_1 alleles in the population?
- Answer: Let X be the Bernoulli random variable whose value is 1 if we have selected the allele A_1 from the pool of $2N$ alleles, proportion p of which are A_1 . Let Y be the random variable whose value is the proportion of A_1 alleles after sampling $2N$ alleles.

$$X = \begin{cases} 1 & \text{if } A_1 \text{ selected} \\ 0 & \text{if } A_2 \text{ selected} \end{cases}$$

$$\text{Prob}[X = 1] = p$$

$$\text{Prob}[X = 0] = q = 1 - p$$

$$Y = \frac{1}{2N} \cdot \sum_{i=1}^{2N} X_i$$

$$E[Y] = E\left[\frac{1}{2N} \cdot \sum_{i=1}^{2N} X_i\right] = \frac{1}{2N} \cdot \sum_{i=1}^{2N} E[X_i] = \frac{1}{2N} \cdot 2Np = p$$

$$V[Y] = V\left[\frac{1}{2N} \cdot \sum_{i=1}^{2N} X_i\right] = \frac{1}{(2N)^2} \cdot \sum_{i=1}^{2N} V[X_i] = \frac{1}{(2N)^2} \cdot 2Npq = \frac{pq}{2N}$$

- It follows that variance is **larger** for a **small** population than for a larger population – this is the cause of **genetic drift**

Moran model

- The Wright-Fisher model may seem contrived, since it assumes **no overlap** between generations – since all members of the current generation die, after having selected the surviving offspring for the next generation
- A somewhat more realistic model is the **Moran model**, proposed in

Random processes in genetics

P.A.P. Moran

Mathematical Proceedings of the Cambridge Philosophical Society

Volume 54, Issue 1 January 1958 , pp. 60-71

Given a current population of n individuals, one applies n successive steps to create the next generation, where a single step consists of selecting 2 individuals, the first is destined to be copied (birth of new individual) while the second is destined to be removed (die) – note that the **same** individual can be selected both to be copied and to be removed, in which case the population remains unchanged

- The **Moran model** is a type of **birth and death process**, introduced originally for n haploid individuals
- In the diploid case, one applies $2n$ steps as described above – note this does not adequately model mating and production of diploid offspring, but is nevertheless used in this fashion for diploid populations

Mean and variance in Moran model

- Question: In the **Moran model** for a diploid, monoecious, hermaphroditic species, population size is restricted to be a constant value N . Suppose that the proportion of A_1 alleles at a given locus in the current population is $p = \frac{k}{2N}$, while that for A_2 alleles is $q = \frac{2N-k}{2N}$. In the next generation, what is the **expectation** and the **variance** in the proportion of A_1 alleles in the population?
- Answer: Assume the initial proportion of A_1 alleles is

$$p = \frac{k}{2N}$$

- Let X be random variable whose value is the the proportion of A_1 alleles after **one step** in the Moran birth-and-death process – i.e. after selecting one allele for birth and one allele for death among the $2N$ alleles. Since there were initially k many A_1 alleles, X can only take on the values initial value

$$X \in \left\{ \frac{k-1}{2N}, \frac{k}{2N}, \frac{k+1}{2N} \right\}$$

$$\text{Prob}[X = \frac{k}{2N}] = p^2 + q^2 \quad (\text{prob of selecting } A_1 \text{ [resp } A_2 \text{] allele for both birth and death})$$

$$\text{Prob}[X = \frac{k+1}{2N}] = pq \quad (\text{prob of selecting } A_1 \text{ allele for birth, } A_2 \text{ allele for death})$$

$$\text{Prob}[X = \frac{k-1}{2N}] = qp \quad (\text{prob of selecting } A_2 \text{ allele for birth, } A_1 \text{ allele for death})$$

so the expected value of X is

$$\begin{aligned} E[X] &= (p^2 + q^2) \cdot \frac{k}{2N} + pq \cdot \frac{k-1}{2N} + pq \cdot \frac{k+1}{2N} \\ &= \frac{k(p^2 + q^2) + 2kpq}{2N} = \frac{k}{2N}(p+q)^2 = \frac{k}{2N} = p \end{aligned}$$

- it follows that after $2N$ such steps the expected proportion of A_1 alleles is still its initial value $p = \frac{k}{2N}$

- for the variance, proceed as follows

$$V[X] = E[X^2] - E[X]^2$$

$$\begin{aligned}E[X^2] &= (p^2 + q^2) \cdot \frac{k^2}{2N} + pq \cdot \frac{(k-1)^2}{(2N)^2} + pq \cdot \frac{(k+1)^2}{(2N)^2} \\&= \frac{k^2(p^2 + q^2) + pq(k^2 + 1 + k^2 + 1)}{(2N)^2} = \frac{k^2(p^2 + 2pq + q^2) + pq(1 + 1)}{(2N)^2} = \frac{k^2(p+q)^2 + 2pq}{(2N)^2} = \frac{k^2 + 2pq}{(2N)^2}\end{aligned}$$

$$V[X] = \frac{k^2 + 2pq}{(2N)^2} - p^2 = \frac{k^2 + 2pq}{(2N)^2} - \frac{k^2}{(2N)^2} = \frac{2pq}{(2N)^2} = \frac{pq}{2N^2}$$

- by additivity of variance (due to independence – is this true?), after $2N$ such steps the variance is

$$2N \cdot \frac{pq}{2N^2} = \frac{pq}{N}$$

- the previous computation of $E[X]$ and $V[X]$ is correct in the essentials – to be more rigorous, the previous computation actually shows that

X_t = proportion of A_1 alleles at time $t = 0, 1, 2, \dots$

$$X_0 = \frac{k}{2N} \quad (\text{initial proportion of } A_1 \text{ alleles})$$

$$E[(X_{t+1} - X_t) \mid X_t] = \frac{0}{2N} \cdot (X_t^2 + (1 - X_t)^2) + \frac{1}{2N} \cdot (X_t(1 - X_t)) + \frac{-1}{2N} \cdot ((1 - X_t)X_t) = 0$$

$$E[(X_{t+1} - X_t)^2 \mid X_t] = \frac{0^2}{(2N)^2} \cdot (X_t^2 + (1 - X_t)^2) + \frac{1^2}{(2N)^2} \cdot (X_t(1 - X_t)) + \frac{(-1)^2}{(2N)^2} \cdot ((1 - X_t)X_t) = \frac{2X_t(1 - X_t)}{(2N)^2}$$

$$E[X_{t+1}] = E[X_t] = \dots = X_0 = \frac{k}{2N}$$

$$V[(X_{t+1} - X_t)] = E[(X_{t+1} - X_t)^2] - E[(X_{t+1} - X_t)]^2 = \frac{2X_t(1 - X_t)}{(2N)^2}$$

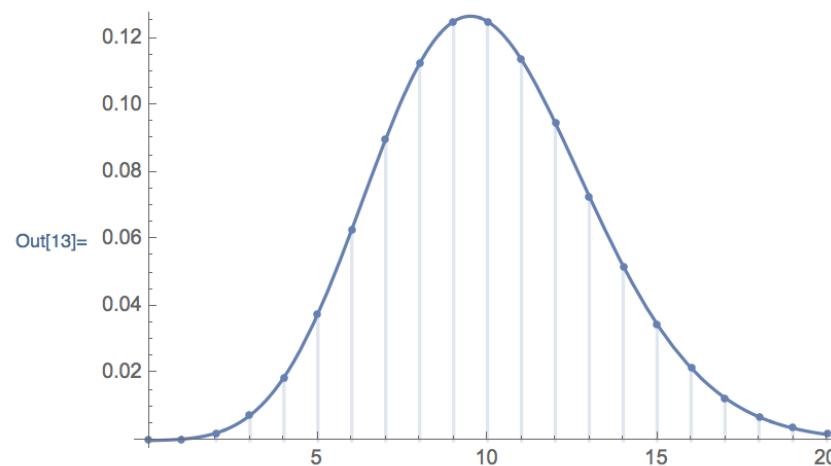
$$\begin{aligned} V[X_{t+1} - X_0] &= V[(X_{t+1} - X_t) + (X_t - X_{t-1}) + \dots + (X_1 - X_0)] \\ &\approx 2N \cdot \frac{2X_0(1 - X_0)}{(2N)^2} = \frac{pq}{N} \end{aligned}$$

- Note that the variance in the Moran model is **twice** that of the Wright-Fisher model, but in both models, the variance of small populations is much larger than that of large populations – i.e. variance is higher in bonobos than in humans!
- Moran showed that the rate of **decrease of heterozygosity** is **twice** as fast in the **Moran** model than in the **Fisher-Wright** model – decrease of heterozygosity is discussed in Chapter 2 of Gillespie.

Poisson distribution

```
In[9]:= mu = 10;
poissonDist = PDF[PoissonDistribution[mu], k]
G1 = DiscretePlot[poissonDist, {k, 0, 20}, Filling -> Axis];
G2 = Plot[poissonDist, {k, 0, 20}];
Show[G1, G2]
```

$$\text{Out}[10]= \begin{cases} \frac{10^k}{e^{10} k!} & k \geq 0 \\ 0 & \text{True} \end{cases}$$



- Poisson distribution is **discrete** distribution defined for all non-negative integers k

$$\text{Prob}[X = k] = e^{-\mu} \cdot \frac{\mu^k}{k!}$$

- Image shows Mathematica commands to plot Poisson distribution for $\mu = 10$ in the bounded range $[0, 20]$.

■ **Poisson** random variable with mean μ

$$\text{Prob}[X = k] = e^{-\mu} \cdot \frac{\mu^k}{k!}$$

$$E[X] = \mu$$

$$V[X] = \mu$$

Indeed,

$$E[X] = \sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} = \sum_{k=1}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} = \sum_{k=1}^{\infty} \mu \frac{\mu^{k-1}}{(k-1)!} e^{-\mu} = \mu \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} e^{-\mu} = \mu$$

Note that the previous line shows that

$$\sum_{k=0}^{\infty} k \frac{\mu^k}{k!} = \mu \cdot e^{\mu}$$

which will be used in the next calculation. The second moment is

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\infty} k^2 \frac{\mu^k}{k!} e^{-\mu} = \mu e^{-\mu} \frac{d}{d\mu} \left(\sum_{k=0}^{\infty} \frac{k \mu^k}{k!} \right) \\ &= \mu e^{-\mu} \frac{d}{d\mu} (\mu e^{\mu}) \\ &= \mu e^{-\mu} (e^{\mu} + \mu e^{\mu}) = \mu + \mu^2. \end{aligned}$$

Thus

$$V[X] = E[X^2] - E[X]^2 = (\mu + \mu^2) - \mu^2 = \mu$$

Geometric distribution

- **geometric** random variable, i.e. number of times necessary to flip a coin with heads probability p until obtain the first heads

$$\text{Prob}[X = k] = (1 - p)^{k-1} \cdot p$$

$$E[X] = \frac{1}{p}$$

$$V[X] = \frac{q}{p^2}$$

Indeed

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} kq^{k-1}p = p \sum_{k=0}^{\infty} \frac{d}{dq}(q^k) \\ &= p \frac{d}{dq} \left(\sum_{k=0}^{\infty} q^k \right) = p \frac{d}{dq} \left(\frac{1}{1-q} \right) \\ &= \frac{p}{(1-q)^2} = \frac{1}{p} \end{aligned}$$

- A similar computation yields the variance

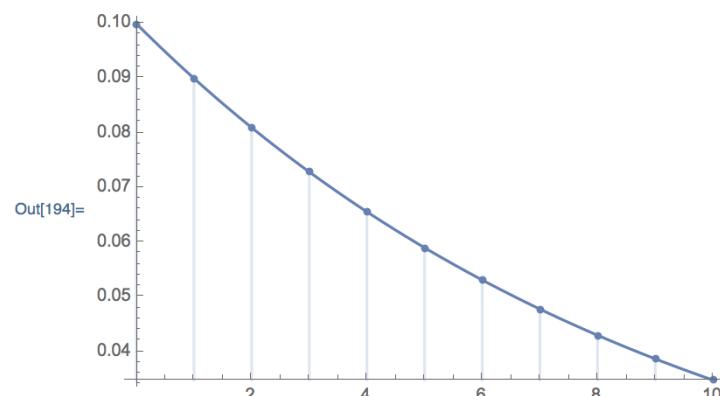
$$\begin{aligned}
 E[X^2] &= \sum_{k=0}^{\infty} k^2 q^{k-1} p = p \sum_{k=0}^{\infty} \frac{d}{dq} (kq^k) \\
 &= p \frac{d}{dq} \left(\sum_{k=0}^{\infty} kq^k \right) = p \frac{d}{dq} \left(\frac{q}{(1-q)^2} \right) \\
 &= p \frac{1+q}{(1-q)^3} = \frac{1+q}{p^2} \\
 V[X] &= E[X^2] - E[X]^2 = \frac{1+q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2} = \frac{1-p}{p^2} = \frac{1}{p^2} - \frac{1}{p} \approx \frac{1}{p^2} \quad \text{if } p \approx 0
 \end{aligned}$$

```

n = 10; p = 1/10;
geomDist = PDF[GeometricDistribution[p], k]
G1 = DiscretePlot[geomDist, {k, 0, 10}, Filling → Axis];
G2 = Plot[geomDist, {k, 0, 10}];
Show[G1, G2]

```

Out[191]=
$$\begin{cases} 9^k \times 10^{-1-k} & k \geq 0 \\ 0 & \text{True} \end{cases}$$

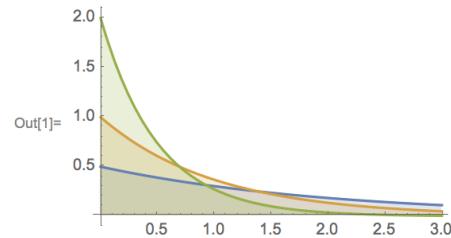


Exponential distribution

Exponential distribution is **continuous** analogue of geometric distribution

Probability density function:

```
In[1]:= Plot[Table[PDF[ExponentialDistribution[\lambda], x], {\lambda, {1/2, 1, 2}}] // Evaluate, {x, 0, 3}, Filling -> Axis, PlotRange -> All]
```

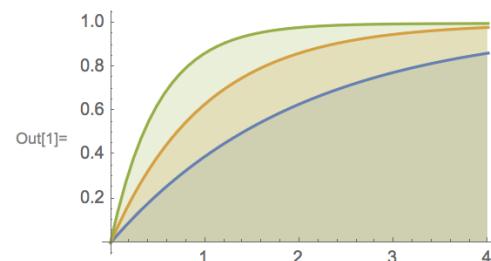


```
In[2]:= PDF[ExponentialDistribution[\lambda], x]
```

$$\text{Out}[2]= \begin{cases} e^{-x\lambda} \lambda & x \geq 0 \\ 0 & \text{True} \end{cases}$$

Cumulative distribution function:

```
In[1]:= Plot[Table[CDF[ExponentialDistribution[\lambda], x], {\lambda, {1/2, 1, 2}}] // Evaluate, {x, 0, 4}, Filling -> Axis]
```



```
In[2]:= CDF[ExponentialDistribution[\lambda], x]
```

$$\text{Out}[2]= \begin{cases} 1 - e^{-x\lambda} \lambda & x \geq 0 \\ 0 & \text{True} \end{cases}$$

- **exponential distribution:** random variable X is exponentially distributed with parameter $\lambda > 0$ if its **probability density function** is

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and the **cumulative distribution function** (CDF) is

$$\begin{aligned} F(x) &= \int_{-\infty}^x p(t) dt = \int_{-\infty}^x \lambda e^{-\lambda t} dt \\ &= \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

- By integration by parts

$$\begin{aligned} E[X] &= \mu = \frac{1}{\lambda} \\ V[X] &= \frac{1}{\lambda^2} = \mu^2 \end{aligned}$$

- Note that the **exponential** distribution is the **continuous** analogue of the **geometric** distribution and that mean and standard deviation of the exponential distribution with parameter λ is equal to λ . In particular, for a geometric and exponential distribution having the same mean μ [$\mu = 1/p$ in geometric case, $\mu = 1/\lambda$ in exponential case], the variance of the geometric distribution is q/p^2 , which for $p \approx 0$, is close to $\mu^2 = 1/p^2$, while that of the exponential distribution is $\mu^2 = 1/\lambda^2$.

Time between successive mutations modeled by exponential distribution

- λ is nucleotide mutation rate per site per generation, with typical value of $\lambda = 10^{-9}$ for eukaryotes
- expected number of mutations occurring at a given site in time (generations) t is λt
- for $\Delta t \approx 0$, we can take $\lambda \Delta t$ to be the probability that one mutation occurs at fixed site (really one or more, but since λ and Δt are small, no more than one mutation can occur within Δt time)
- let $P_0(t)$ denote the probability that **no** mutation occurs at a given site for time t
- probability that no mutations occur in time $t + \Delta t$ equals probability $P_0(t)$ that no mutations occur in time t times probability that no mutations occur in time Δt

$$P_0(t + \Delta t) = P_0(t) \cdot (1 - \lambda \Delta t)$$
$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = \frac{P_0(t) \cdot (1 - \lambda \Delta t) - P_0(t)}{\Delta t} = \frac{-\lambda \Delta t P_0(t)}{\Delta t} = -\lambda P_0(t)$$
$$\frac{dP_0(t)}{dt} = \lim_{t \rightarrow \infty} \frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t)$$

$$P_0(t)^{-1} dP_0(t) = -\lambda dt$$
$$\int P_0(t)^{-1} dP_0(t) = - \int \lambda dt$$
$$\ln P_0(t) = -\lambda t + c$$
$$P_0(t) = C_0 \cdot e^{-\lambda t} \quad \text{where } C_0 = e^c$$

- clearly $P_0(0) = 1$, since it is certain (with probability 1) that no mutations have occurred in time 0
- it follows that $C_0 = 1$ and so

$$P_0(t) = e^{-\lambda t}$$

- Since

$$P_0(t) = e^{-\lambda t}$$

it follows that the probability that a mutation occurs within time t is $1 - P_0(t)$ hence

$$(1 - P_0(t)) = 1 - e^{-\lambda t}$$

- $1 - e^{-\lambda t}$ is the cumulative density function CDF of the **exponential distribution** with parameter λ , hence with expectation $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$
- It follows that if mutations occur randomly (i.e. do not occur in bursts at particular times), then the expected time between successive mutations occurring in a genome is exponentially distributed with mean $\frac{1}{\lambda}$
- replacing time by genome length, this argument shows that if particular motifs (occurrences of AUG, TATA-boxes, etc.) are randomly located in the genome, then the distance between two successive motifs is exponentially distributed
- since the **geometric** distribution is a discrete analogue of the exponential distribution, the relative frequency histogram of distances (number of nucleotides) between two successive occurrences of a given motif should be approximately geometrically distributed
- a similar argument shows that the **number of mutations** over time t for a given (constant) mutation rate μ is modeled by the **Poisson process**

$$\text{Prob}[X = k] = e^{-\mu t} \cdot \frac{(\mu t)^k}{k!}$$

- in the context of the coalescent, **waiting times** T_k for a coalescence of 2 individuals (alleles) at level k is obtained by **sampling** from the **exponential distribution** with parameter

$$N \cdot \frac{k(k-1)}{2} \quad (\text{haploid case})$$

$$2N \cdot \frac{k(k-1)}{2} \quad (\text{diploid case})$$

where N is population size (more generally N_e is effective population size).

χ^2 distribution with $df = m$

- random variable X has a χ^2 distribution with m degrees of freedom (df) if X is the sum of m random variables Y , where each $Y = Z^2$ and Z is a **standard normal variate**

$$X = Y + \dots + Y$$

$$Y = Z^2$$

$$X = Z^2 + \dots + Z^2$$

- recall of course that for any random variable X , it is usually the case that $X + X \neq 2X$



$$E[Z] = 0$$

$$V[Z] = 1$$

$$E[Z^2] = 1$$

$$V[Z^2] = 2$$

$$E[\chi_m^2] = E[Z^2 + \dots + Z^2] = m$$

$$V[\chi_m^2] = V[Z^2 + \dots + Z^2] = 2m$$

- sample 10 values from the standard normal distribution

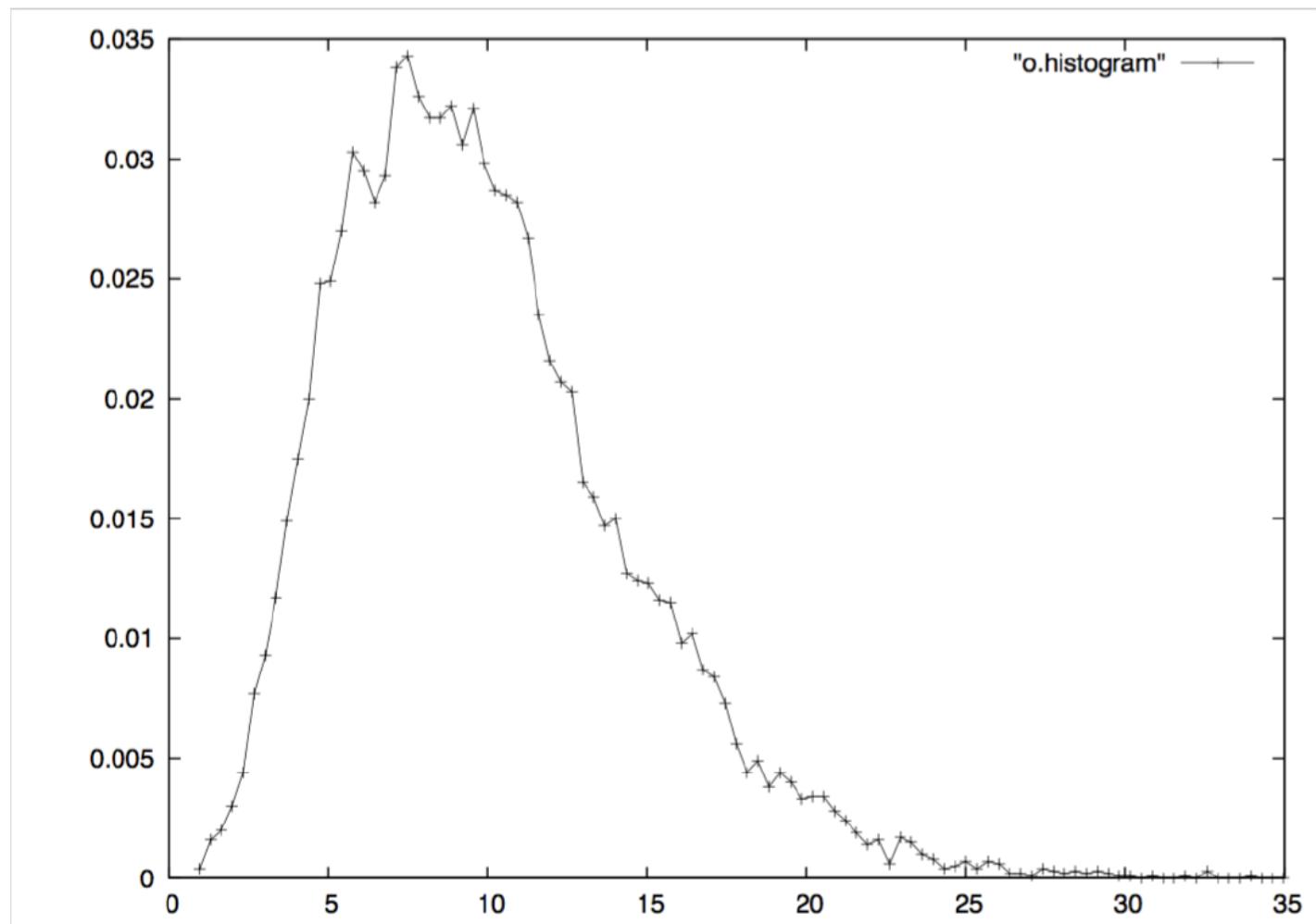
-1.14600012941
-1.77820027647
0.303347491845
-0.691491153282
-0.95336219333
-0.536636874138
1.16979658858
0.716272610588
-0.290723832861
0.208472562766

- compute the squares of these numbers

1.31331629661
3.16199622324
0.0920197008088
0.478160015067
0.908899471671
0.287979134684
1.36842405866
0.513046452678
0.0845203469934
0.0434608094261

- compute the sum of these 10 numbers: 8.25182250984
- repeat this operation 10,000 times and create the corresponding **relative frequency histogram** shown on the next slide

Approximation of χ^2 with $df = 10$ by 10,000 sums, each of 10 squared normal variates



Covariance

• Variance of X :

$$\begin{aligned}V[X] &= E[(X - \mu)^2] \\&= E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 \\&= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2\end{aligned}$$

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_x)(Y - \mu_y)] \\&= E[XY - \mu_x Y - \mu_y X + \mu_x \mu_y] = E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y \\&= E[XY] - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y = E[XY] - \mu_x \mu_y\end{aligned}$$

The covariance is linear in each coordinate, in that

$$\begin{aligned}\text{Cov}(X, Y + Z) &= \text{Cov}(X, Y) + \text{Cov}(X, Z) \\ \text{Cov}(X + Y, Z) &= \text{Cov}(X, Z) + \text{Cov}(Y, Z) \\ \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) &= \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)\end{aligned}$$

This is shown by induction from a simple calculation. For instance, to show linearity in the second argument, we have the following, where $U = Y + Z$.

$$\begin{aligned}\text{Cov}(X, Y + Z) &= \text{Cov}(X, U) = E[XU] - \mu_x \mu_u \\E[XU] &= E[X(Y + Z)] = E[XY] + E[XZ] \\ \mu_x \mu_u &= \mu_x (\mu_y + \mu_z) = \mu_x \mu_y + \mu_x \mu_z \\ \text{Cov}(X, Y + Z) &= E[XY] + E[XZ] - \mu_x \mu_y - \mu_x \mu_z = (E[XY] - \mu_x \mu_y) + (E[XZ] - \mu_x \mu_z) \\&= \text{Cov}(X, Y) + \text{Cov}(X, Z)\end{aligned}$$

Independence and correlation

- random variables X, Y are **independent**, if for all real numbers a, b ,

$$P(\{X \leq a\}, \{Y \leq b\}) = P(\{X \leq a\}) \cdot P(\{Y \leq b\})$$

- random variables X, Y are said to be **uncorrelated** if

$$E[XY] = E[X] \cdot E[Y]$$

- independent random variables are uncorrelated, but the converse is not always true

independent r.v. \Rightarrow uncorrelated r.v.

uncorrelated r.v. $\not\Rightarrow$ independent r.v.

- if X, Y are random variables, then **correlation** between X, Y is defined by

$$E[(X - \mu_x)(Y - \mu_y)]$$

Pearson linear correlation coefficient

- variance $V[X] = E[(X - \mu_x)^2]$ is correlation of X with itself
- (Pearson linear) population correlation coefficient, $\rho_{X,Y}$, between random variables X, Y is defined by

$$\begin{aligned}\rho_{X,Y} &= \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \\ &= \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \\ &= \frac{E[XY] - \mu_x \mu_y}{\sigma_x \sigma_y}\end{aligned}$$

- recalling that $E[X]$ and $E[Y]$ are **population means**, for finite samples x_1, \dots, x_n and y_1, \dots, y_n (equal size samples), we have

$$\mu_x = E[X] \approx \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\mu_y = E[Y] \approx \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

- (Pearson linear) sample correlation coefficient $r_{x,y}$** defined by

$$r_{x,y} = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right) \cdot \left(\frac{\sum_{i=1}^n y_i}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \cdot \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}}$$

$$= \frac{n \left(\sum_{i=1}^n x_i y_i \right) - \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{\sqrt{\left[n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right]} \cdot \sqrt{\left[n \cdot \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 \right]}}$$

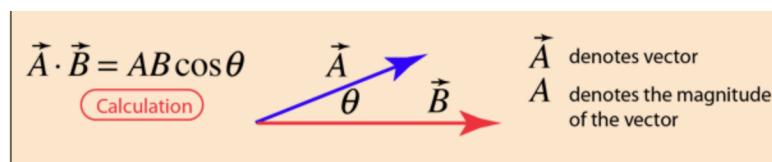
- **(population) correlation coefficient** $\rho_{x,y}$ measures the extent to which the n -coordinate vectors $\vec{x} = (x_1, \dots, x_n)$ and $\vec{y} = (y_1, \dots, y_n)$ are **co-linear**
- Why? Since correlation coefficient equals the **inner product** (also known as dot product) of the sequence of Z-scores corresponding to n -dimensional vectors $\vec{x} = (x_1, \dots, x_n)$ and $\vec{y} = (y_1, \dots, y_n)$

$$\begin{aligned}\rho_{x,y} &= \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_x \sigma_y} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y} \\ &= \left(\frac{1}{n}\right) \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x}\right) \cdot \left(\frac{y_i - \bar{y}}{\sigma_y}\right)\end{aligned}$$

- for 2-dimensional vectors $\vec{x} = (x_1, x_2)$ and $\vec{y} = (y_1, y_2)$, the **inner product** (called **scalar** product in physics, and sometimes called **dot** product) is defined by

$$\begin{aligned}\vec{x} \cdot \vec{y} &= ||\vec{x}|| \cdot ||\vec{y}|| \cdot \cos(\theta) \\ &= x_1 y_1 + x_2 y_2\end{aligned}$$

- note that 2-dimensional vectors \vec{x} and \vec{y} are **co-linear** exactly when $\cos(\theta) = 1$, i.e. when $\theta = 0$.



(Image credit: hyperphysics.phy-astr.gsu.edu)

- population correlation coefficient ρ satisfies $-1 \leq \rho \leq 1$
- sample correlation coefficient r satisfies $-1 \leq r \leq 1$
- hypothesis testing** for correlation

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

compute correlation coefficient r and **test statistic**

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

using T-distribution with $n - 2$ **degrees of freedom**

- regression line** for a **scatter plot** of the points of the points (x_i, y_i) , for $i = 1, \dots, n$ satisfies the property that $\sum_{i=1}^n |y_i - \hat{y}_i|^2$ is a **minimum** – known as **least squares linear fit**

$$y = mx + b$$

of the points (x_i, y_i) to a line, where $\hat{y}_i = mx_i + b$

- slope** of line is m , and the y -intercept) is b

Regression

- given arbitrary equation for line $y = mx + b$, consider the **residuals**, $(y_1 - \hat{y}_1), \dots, (y_n - \hat{y}_n)$, where $\hat{y}_i = m \cdot x_i + b$. In order to determine optimal m, b to minimize the sum of squares of the residuals

$$f(m, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - mx_i - b)^2$$

compute the first (partial) derivatives

$$\frac{\partial f(m, b)}{\partial m} = \sum_{i=1}^n 2(y_i - mx_i - b) \cdot (-x_i)$$

and

$$\frac{\partial f(m, b)}{\partial b} = \sum_{i=1}^n 2(y_i - mx_i - b) \cdot (-1)$$

and set equal to 0. By calculus, we obtain that the slope of the least squares fit is

$$m = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{n \cdot (\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} = \frac{Cov(X, Y)}{\sigma_x^2} = \frac{Cov(X, Y)}{Var(X)}$$

- least squares (or regression) line has the property that

$$\bar{y} = m \cdot \bar{x} + b$$

so to compute b , one computes the average \bar{y} of y -values and the average \bar{x} of x -values finds

$$b = \bar{y} - m \cdot \bar{x}.$$

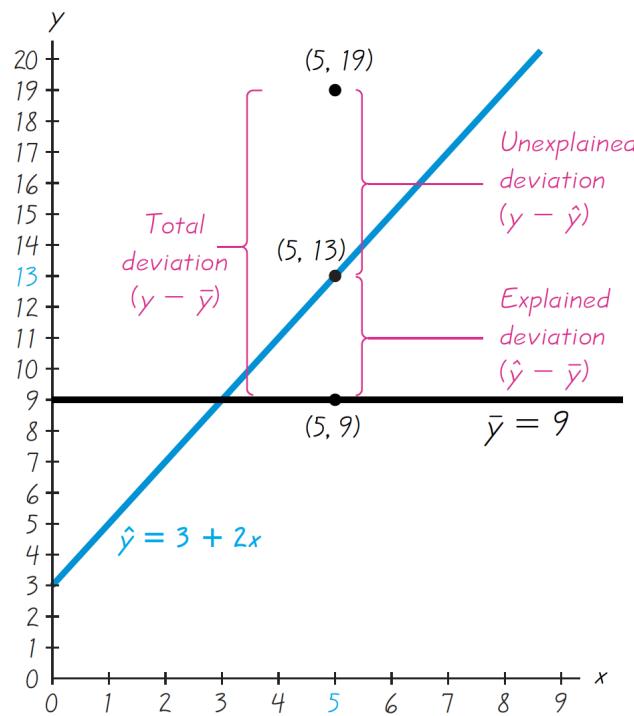
- standard error of estimate s_e is defined by

$$s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

where the predicted value $\hat{y}_i = mx_i + b$ for m, b computed by regression (least squares)

- do not confuse **standard error of estimate** with **standard error (of the mean)**

$$s_i = \frac{\sigma}{\sqrt{n}}$$



(Image credit: Biostatistics by Triola and Triola)

Beta distribution

- The beta distribution with shape parameters α, β is designated by $Beta(\alpha, \beta)$, and has probability density function

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} \cdot x^{\alpha-1} \cdot (1 - x)^{\beta-1}$$

where $0 \leq x \leq 1$

- recall that for positive integers n , $\Gamma(n) = (n - 1)!$
- recall that the binomial probability mass function for the probability of k heads in n coin flips with heads probability p is given by

$$Prob[X = k] = \binom{n}{k} p^k \cdot (1 - p)^{n-k} = \frac{n}{k \cdot (n - k)} \cdot p^k \cdot (1 - p)^{n-k}$$

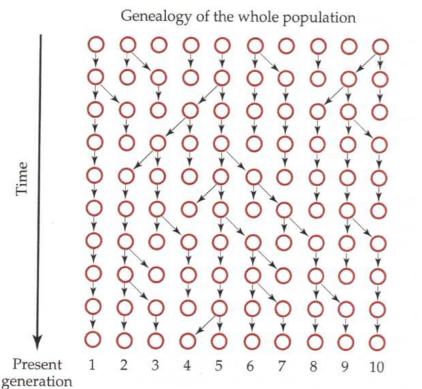
so by approximate superficial similarity you might find it easier to recall the density function for the beta distribution



$$\begin{aligned}\mu &= E[X] = \frac{\alpha}{\alpha + \beta} \\ \sigma^2 &= V[X] = \frac{1}{4(2\beta + 1)}\end{aligned}$$

Warning: typos on page 199 of book!

Coalescent



```

A = present population {1,...,n} with indicated alleles
t = 0 #time backwards from present time
k = n #number of samples at time t
T = {} #T is dictionary of times T[n],T[n-1],...,T[2]
while k>1:
    randomly choose 2 of the k to coalesce
    mean = 2/(k*(k-1)) #mean coalescent time for k samples
    x = random floating point number in (0,1)
    T[k] = -mean*log(x) #sample coalescent time from exp dist
    k = k-1 #number of samples is decreased
return dictionary T
#Note: time for k-th coalescence is 2N*T[k] for diploid population size N

```

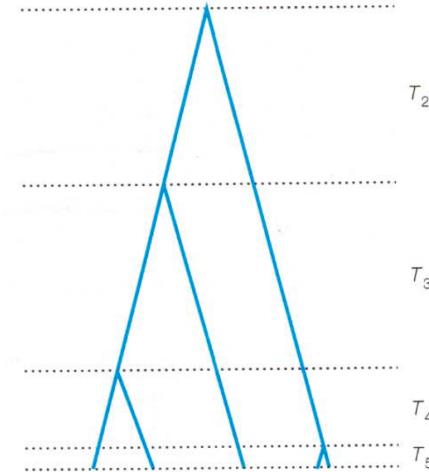


FIGURE 6.2 Example of a coalescent genealogy for a sample size of $n = 5$. The times between coalescent events, T_i , are equivalent to the amount of time that there are i lineages in the genealogy.

- the **coalescent** is an approximate **backward** simulation of the (forward) Wright-Fisher process, which samples **potential** genealogies for the ancestors of a given current population of individuals (alleles)
- a **coalescence** occurs when two individuals (alleles) from generation k have the same parent in generation $k - 1$, an event which occurs with probability $\frac{1}{N}$ (haploid case) and $\frac{1}{2N}$ (diploid case), where N is the (constant) population size – later N will be replaced by **effective population size** N_e
- the coalescent has many applications, since mutations can be “thrown” at coalescent-generated genealogies, recombination can be accommodated, etc.
- to understand the simple algorithm behind the coalescent, we need first to revisit the **exponential distribution**, which is the continuous analogue of the **geometric distribution**

Time between successive mutations modeled by exponential distribution

- λ is nucleotide mutation rate per site per generation, with typical value of $\lambda = 10^{-9}$ for eukaryotes
- expected number of mutations occurring at a given site in time (generations) t is λt
- for $\Delta t \approx 0$, we can take $\lambda \Delta t$ to be the probability that one mutation occurs at fixed site (really one or more, but since λ and Δt are small, no more than one mutation can occur within Δt time)
- let $P_0(t)$ denote the probability that **no** mutation occurs at a given site for time t
- probability that no mutations occur in time $t + \Delta t$ equals probability $P_0(t)$ that no mutations occur in time t times probability that no mutations occur in time Δt

$$P_0(t + \Delta t) = P_0(t) \cdot (1 - \lambda \Delta t)$$
$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = \frac{P_0(t) \cdot (1 - \lambda \Delta t) - P_0(t)}{\Delta t} = \frac{-\lambda \Delta t P_0(t)}{\Delta t} = -\lambda P_0(t)$$
$$\frac{dP_0(t)}{dt} = \lim_{t \rightarrow \infty} \frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t)$$

$$P_0(t)^{-1} dP_0(t) = -\lambda dt$$
$$\int P_0(t)^{-1} dP_0(t) = - \int \lambda dt$$
$$\ln P_0(t) = -\lambda t + c$$
$$P_0(t) = C_0 \cdot e^{-\lambda t} \quad \text{where } C_0 = e^c$$

- clearly $P_0(0) = 1$, since it is certain (with probability 1) that no mutations have occurred in time 0
- it follows that $C_0 = 1$ and so

$$P_0(t) = e^{-\lambda t}$$

- Since

$$P_0(t) = e^{-\lambda t}$$

it follows that the probability that a mutation occurs within time t is $1 - P_0(t)$ hence

$$(1 - P_0(t)) = 1 - e^{-\lambda t}$$

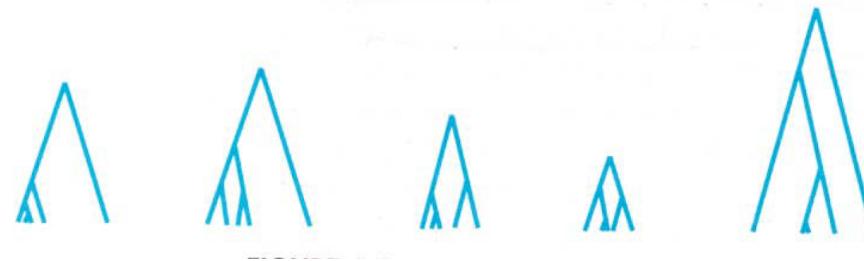
- $1 - e^{-\lambda t}$ is the cumulative density function CDF of the **exponential distribution** with parameter λ , hence with expectation $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$
- It follows that if mutations occur randomly (i.e. do not occur in bursts at particular times), then the expected time between successive mutations occurring in a genome is exponentially distributed with mean $\frac{1}{\lambda}$
- replacing time by genome length, this argument shows that if particular motifs (occurrences of AUG, TATA-boxes, etc.) are randomly located in the genome, then the distance between two successive motifs is exponentially distributed
- since the **geometric** distribution is a discrete analogue of the exponential distribution, the relative frequency histogram of distances (number of nucleotides) between two successive occurrences of a given motif should be approximately geometrically distributed
- a similar argument shows that the **number of mutations** over time t for a given (constant) mutation rate μ is modeled by the **Poisson process**

$$\text{Prob}[X = k] = e^{-\mu t} \cdot \frac{(\mu t)^k}{k!}$$

- in the context of the coalescent, **waiting times** T_k for a coalescence of 2 individuals (alleles) at level k is obtained by **sampling** from the **exponential distribution** with parameter

$$N \cdot \frac{k(k-1)}{2} \quad (\text{haploid case})$$
$$2N \cdot \frac{k(k-1)}{2} \quad (\text{diploid case})$$

Sampling waiting times from exponential distribution



- For computer simulations, an exponentially distributed random variable for a given mean μ can be computed as follows. Let X be a uniformly distributed continuous random variable with $0 \leq X \leq 1$. (In Python, this can be done by calling function `random()`.) Note that $Pr[X > x] = 1 - x$. Then

$$\begin{aligned} Pr[\text{there is an arrival in time } t] &= 1 - e^{-\frac{t}{\mu}} \\ &= Pr[X > e^{-\frac{t}{\mu}}] = Pr\left[\ln X > -\frac{t}{\mu}\right] = Pr[-\mu \ln X < t] \end{aligned}$$

Thus repeatedly evaluating $-\mu \ln X$ for uniformly distributed random real numbers $0 \leq X \leq 1$ yields a sequence of sample interarrival times with mean μ .

- the stochastic variation in sampled times T_k , for $k = n, \dots, 2$ explains the difference in heights of sampled genealogies shown at the top of this slide

Some useful approximations

- The following is the **geometric series** that you know from high school.

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + \dots$$

It follows that for small $x \approx 0$,

$$\frac{1}{1-x} \approx 1 + x$$

- Integrate the geometric series term by term to get the following identity.

$$\begin{aligned}\ln(1-x) &= - \int_0^x \frac{dt}{1-t} \quad (\text{recall that } \ln x = \int_0^x \frac{dt}{t}) \\ &= \sum_{n=1}^{\infty} \frac{x^n}{n} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots \quad \ln(1-x) \\ &\qquad\qquad\qquad = - \sum_{n=1}^{\infty} \frac{x^n}{n} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots\end{aligned}$$

Replace x by $-x$ in previous equation to get the following identity.

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

Alternatively, one obtains the same result by performing a Taylor expansion of $f(x) = \ln(1-x)$ about 0 (i.e. a MacLaurin expansion), where one notes that $f(0) = \ln 1 = 0$. It follows that for small $x \approx 0$,

$$\ln(1+x) \approx x$$

- Differentiate the geometric series term by term to get the following.

$$\frac{1}{(1-x)^2} = \sum_{n=1}^{\infty} nx^{n-1} = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + \dots$$

It follows that for small $x \approx 0$,

$$\frac{1}{(1-x)^2} \approx 1 + 2x$$

- Multiply previous identity by x to get

$$\frac{x}{(1-x)^2} = \sum_{n=1}^{\infty} nx^n = x + 2x^2 + 3x^3 + 4x^4 + 5x^5 + \dots$$

It follows that for small $x \approx 0$,

$$\frac{x}{(1-x)^2} \approx x + 2x^2 \approx x$$

- Differentiate the geometric series twice to get the following

$$\frac{2}{(1-x)^3} = \sum_{n=1}^{\infty} n(n-1)x^{n-2} = 2 + (3 \cdot 2)x + (4 \cdot 3)x^2 + (5 \cdot 4)x^3 + \dots$$

It follows that for small $x \approx 0$,

$$\frac{2}{(1-x)^3} \approx 2 + 6x$$

Approximations involving the exponential function

- By the **binomial theorem**

$$(a+b)^n = \sum_{i=0}^n \binom{n}{i} a^i b^{n-i}$$

it follows that

$$\begin{aligned}(1-x)^n &= (-x+1)^n = \sum_{i=0}^n \binom{n}{i} \cdot (-x)^i \cdot 1^{n-i} \\&= 1 - nx + \frac{n(n-1)}{2}x^2 - \frac{n(n-1)(n-2)}{3 \cdot 2}x^3 + \cdots + \frac{n(n-1)}{n}x^n \\&\approx 1 - nx + \frac{(nx)^2}{2!} - \frac{(nx)^3}{3!} + \frac{(nx)^4}{4!} - \cdots = \sum_{i=0}^{\infty} \frac{(-nx)^i}{i!} = e^{-nx} \quad (\text{when } n \text{ is large})\end{aligned}$$

and hence for large n ,

$$(1 - \frac{1}{n})^n \approx e^{-1} \approx \frac{1}{2.7183} \approx 0.3679$$

The value of e^{-1} comes up surprisingly often – for instance, e^{-1} is (approximately) the probability that **no one** gets back their own hat in a room of n people, for large n , where each person chooses a hat randomly from the collection of all hats.

Arithmetic, geometric, harmonic mean

- **arithmetic mean (AM):** Given x_1, \dots, x_n , arithmetic mean is

$$AM = \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right)$$

- **geometric mean (GM):** Given x_1, \dots, x_n , geometric mean is

$$GM = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

- **harmonic mean (HM):** Given x_1, \dots, x_n , harmonic mean is

$$HM = \frac{1}{\frac{1}{n} \cdot \left(\sum_{i=1}^n \frac{1}{x_i} \right)} = \frac{n}{\sum_{i=1}^n x_i^{-1}}$$

■ **Theorem 1** (Arithmetic-geometric mean inequality $GM \leq AM$). *Given non-negative numbers x_1, \dots, x_n , the geometric mean $\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$ is less than or equal to the arithmetic mean $\frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right)$, where equality holds if and only if $x_1 = x_2 = \dots = x_n$.*

Since $AM = \ln(GM)$, the inequality portion of the statement of the theorem is obvious.

■ **Theorem 2** (Geometric-harmonic mean inequality $HM \leq GM$). *Given non-negative numbers x_1, \dots, x_n , the harmonic mean $\frac{1}{\frac{1}{n} \cdot \left(\sum_{i=1}^n \frac{1}{x_i} \right)}$ is less than or equal to the geometric mean $\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$.*

Weighted arithmetic, geometric, harmonic mean

- **weighted arithmetic mean (AM):** Given x_1, \dots, x_n , and probabilities p_1, \dots, p_n that sum to 1, the weighted arithmetic mean, or expectation, is

$$AM = \sum_{i=1}^n p_i \cdot x_i$$

- **weighted geometric mean (GM):** Given x_1, \dots, x_n and probabilities p_1, \dots, p_n that sum to 1, the geometric mean is

$$GM = \prod_{i=1}^n x_i^{p_i}$$

- **weighted harmonic mean (HM):** Given x_1, \dots, x_n and probabilities p_1, \dots, p_n that sum to 1, the harmonic mean is

$$HM = \frac{1}{\sum_{i=1}^n \frac{p_i}{x_i}}$$