BST227
Introduction to Statistical Genetics

Lecture 3:
Introduction to population genetics

Housekeeping

- HW1 due on Wednesday
- TA office hours today at 5:20 - FXB G11

What have we studied

Background

- Structure of Human Genome
- DNA Variants and disease

Mendelian Inheritance

- Mendel's first law
- Mendel's second law
- Mode of inheritance
- Genetic models for mendelian and complex disease

Overview of Today's Material

Population Genetics Concepts:

Estimation and Inference About Allele Frequencies

Hardy Weinberg Equilibrium

Population Substructure

Measuring Genetic Contribution to Traits

Recurrence Risk Ratios

Heritability

Allele Frequencies

- Definition:
  Allele frequency = proportion of chromosomes in population carrying the allele of interest. (e.g. a disease allele)
- Allele frequencies are compared in association studies to detect disease genes
- Allele frequencies tell us about the probability of observed genotypes

From genotypes to allele frequencies

Box 4.1: Calculation of Estimated Allele Frequencies from a Sample of Size n Subjects.

Genotype counts from the sample: n_{AA} = number out of n with genotype AA
n_{Aa} = number out of n with genotype Aa
n_{aa} = number out of n with genotype aa

where $n_{AA} + n_{Aa} + n_{aa} = n$. The sample proportion of A alleles,

$$\bar{p} = (2n_{AA} + n_{Aa})/2n, \qquad (4.1)$$

estimates the population proportion of A alleles. With a two allele system, the proportion of a alleles is $\bar{q} = 1 - \bar{p}$, as can be verified by exchanging a with A in formula (4.1).

Inter-generational allele transmission

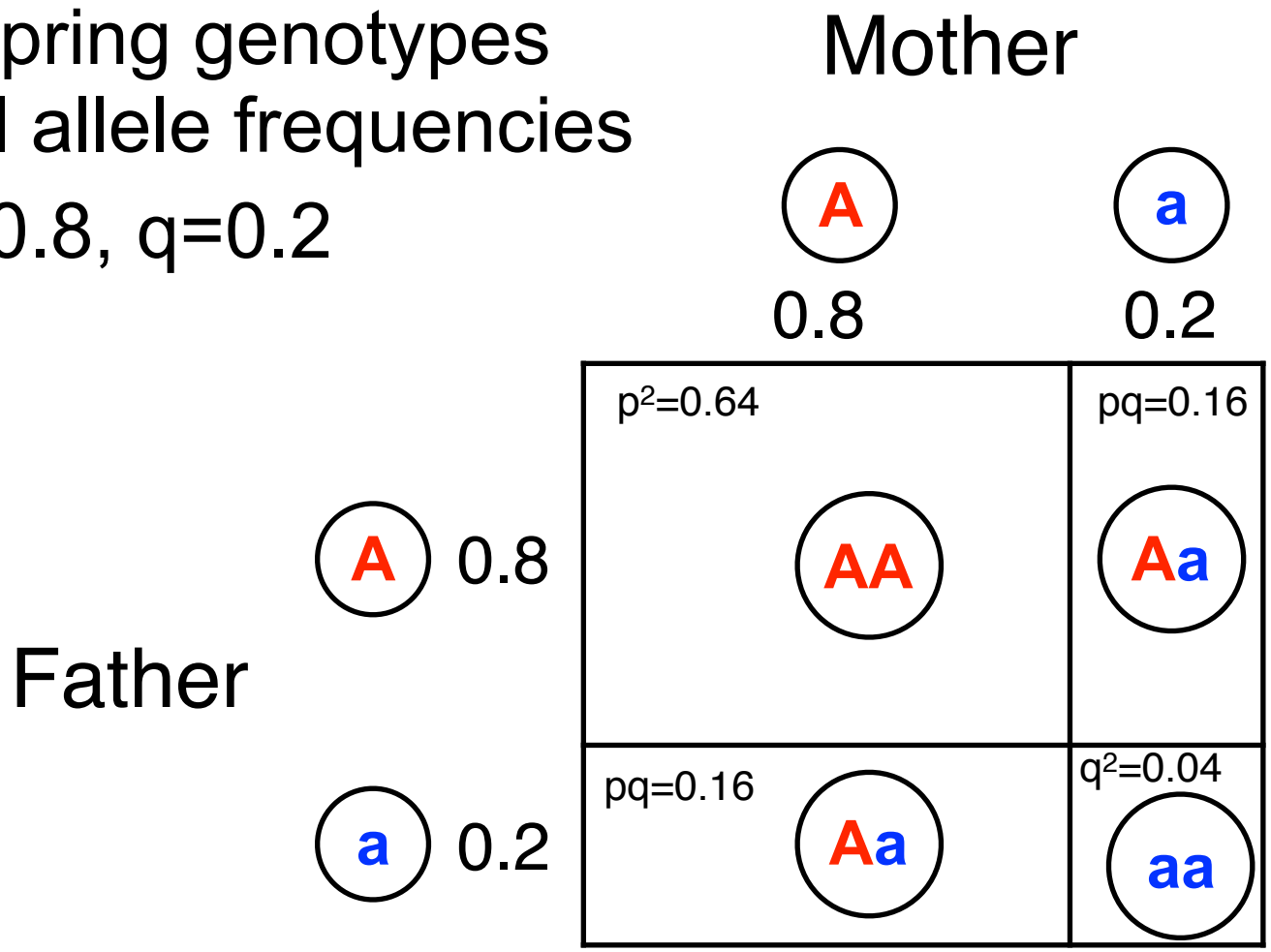
- Assume random mating

Parents: AA Aa Aa aa

Gametes: A A A a A a a a

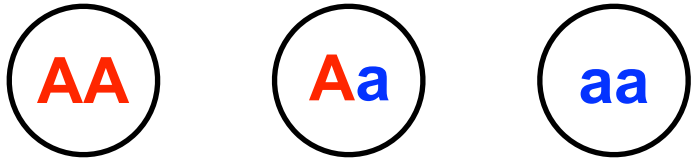
Inter-generational allele transmission

- Estimate offspring genotypes from parental allele frequencies
- Example: p=0.8, q=0.2



Inter-generational allele transmission

- What are the allele frequencies (p*) in the next generation?



p^2	2pq	q^2
0.64	0.32	0.04

p_star = (2 p2 + 2pq) / 2 = p2 + pq = p (p+q) = p

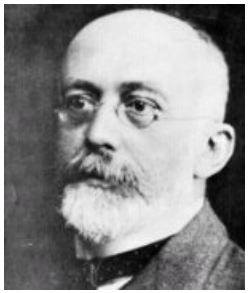
- Allele frequencies remain constant

Hardy Weinberg Equilibrium (HWE)

Theorem: Allele frequencies in a population remain constant if no evolutionary forces exist.

Requirements for Hardy-Weinberg equilibrium:

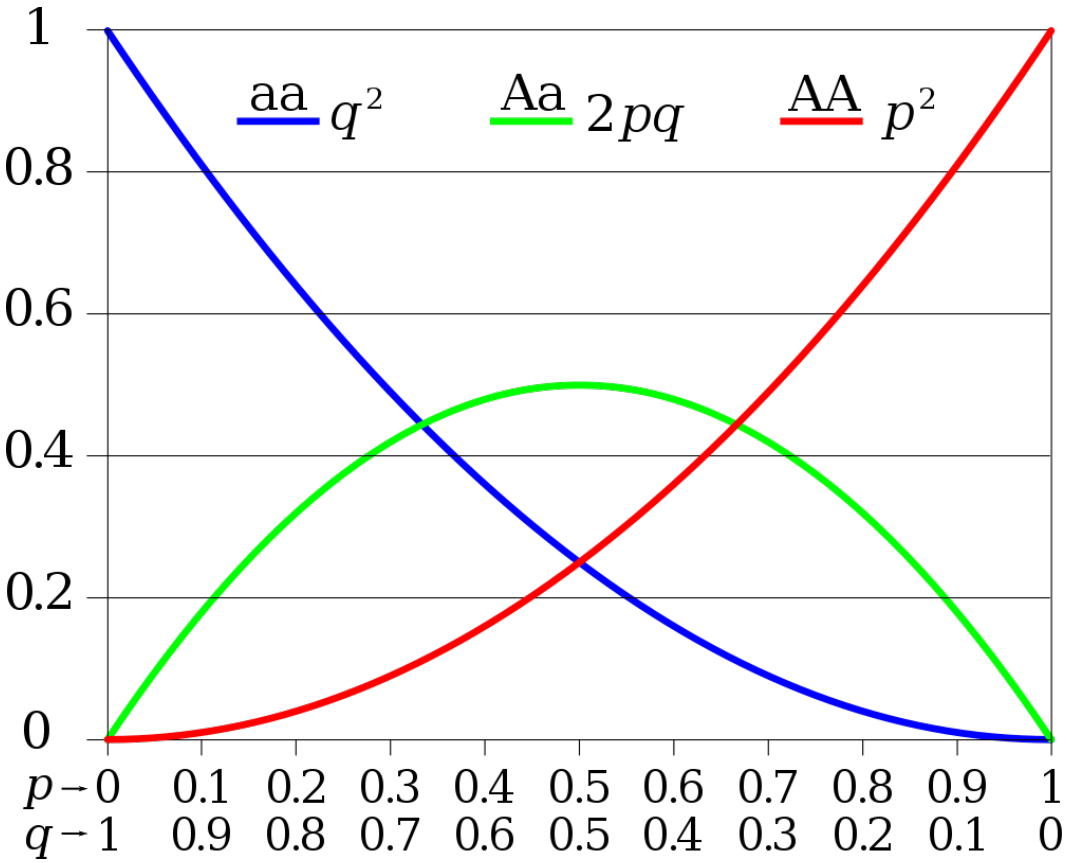
- Large population
- Random mating
- No mutation
- No migration
- No selection



Departures from HW equilibrium provide a mechanism to study evolution

Hardy Weinberg Equilibrium (HWE)

Rule: If you know allele frequency, use HWE to calculate genotype probabilities.



Johnuniq
https://commons.wikimedia.org/wiki/File:Hardy-Weinberg.svg

Inter-generational **genotype** transmission

- Parental population in HWE
- Assume random mating
- p = 0.5

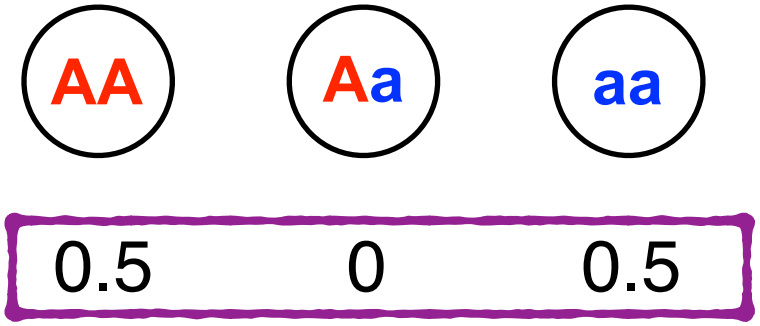
AA	Aa	aa
p^2	2pq	q^2
0.25	0.5	0.25

- What are the genotype frequencies in the next generation?
  Unchanged: **In HWE**

p^2	2pq	q^2
0.25	0.5	0.25

Inter-generational **genotype** transmission

- Parental population **not in** HWE
- Assume random mating
- p = 0.5



- What are the genotype frequencies in the next generation?

p^2	2pq	q^2
0.25	0.5	0.25

In HWE

Implications of HWE

Suppose population is in HWE, then it will remain in HWE after a round of random mating.

Suppose population is not in HWE, then it will get in HWE after one round of random mating.

The allele frequency does not change from one generation to the next.

When is HWE is useful?

The failure of HWE can reveal a lot about sample features:

- Selection of subjects related to genotype
- Population Substructure
- Genotyping errors

How to Detect Failure of HWE: Testing for HWE in a Sample

- Estimate allele frequencies
- Compute Expected genotype frequencies assuming HWE holds
- Use Pearson Chi-Square test

Hardy-Weinberg Equilibrium (HWE)

- Test for HWE based on Pearson chi-square test:

	Genotype			
	AA	Aa	aa	
Observed	n_{AA}	n_{Aa}	n_{aa}	n
Expected	np^2	$2np(1-p)$	$n(1-p)^2$	n

- Estimate p as $(2n_{AA} + n_{Aa}) / 2n$
- The Chi Square Test has 1 degree of freedom. (Why?) can put only one prameter p

Population Substructure: Stratification / Admixture / Inbreeding

- Population stratification: distinct subgroups within a population.

- Population admixture: mating among individuals of different genetic origin over multiple generations. Usually occult.

- Inbreeding: mating between ‘close’ relatives

Stratification

Hereditas 102: 219–223 (*1985*)

Polymorphism of serum albumin in dog breeds and its relation to weight and leg length

K. CHRISTENSEN[1], J. ARNBJERG[2] and E. ANDRESEN[1]

The Royal Veterinary and Agricultural University, Department of Animal Genetics[1], and Small Animal Clinic[2], Copenhagen, Denmark

Leg length

Fig. 2. Dog No. 6 and 7 from Table 2, at an age of 7 weeks, representing the two leg types.

Albumin genotypes

FS SS FS FF

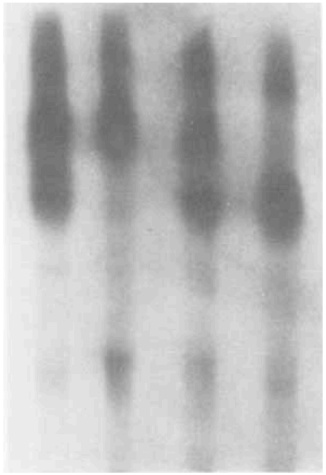


Fig. 1. Gel electrophoresis of serum albumin for 4 dogs showing the three different phenotypes F, S and FS.

Stratification

Breed	Genotypes SS	SF	FF	Total	Frequency of S
Basset Hound	0	2	30	32	0.03
Beagle	3	14	52	69	0.14
Dachshund	2	8	26	36	0.17
Collie	2	21	18	41	0.33
Cocker Spaniel	7	24	20	51	0.37
Labrador Retriever	8	10	10	28	0.46
German Shepherd	36	47	23	106	0.56
Terrier, Tibetanian	10	11	3	24	0.64
Newfoundland	35	33	3	61	0.68
Poodle	39	36	6	81	0.70
Boxer	54	14	1	69	0.88
Golden Retriever	53	3	1	57	0.96
Basenji	44	0	0	44	1.00
Other pure breeds	94	57	38	189	0.65
Mongrels	22	41	24	87	0.49
Total	399	321	255	975	
Overall Gene Frequency					$p_s = 0.574$
Genotypic Frequencies	0.409	0.329	0.262		

Table 4.1: Distribution of albumin types among selected dog breeds and mongrels.
Source: **Hereditas** 102 (1985) p 219-223 Christensen et al.

Dog Breeds and the Albumin Alleles

Test for HWE within breed:

Breed	Genotypes				Frequency of S
	SS	SF	FF	Total	
German Shepherd	36	47	23	106	0.56

	Genotype			
	SS	SF	FF	
Observed	36	47	23	106
Expected	33	52	21	106
	p-value = 0.79			

In HWE

Dog Breeds and the Albumin Alleles

Test for HWE using entire population:

	Genotype			
	SS	SF	FF	
Observed	399	321	255	975
Expected	321	477	177	975
	Highly significant: p-value < 1e-10			

New Topic: How do we measure extent to which a trait is genetic?

Two primary measures:

Recurrence Risk Ratios (dichotomous traits)

Heritability (quantitative traits)

Recurrence Risk Ratio

Definitions:

Proband: Subject selected into sample because of disease status.

P(disease) = K

Relative of type R (parent, sib, etc)

Recurrence risk ratio defined for dichotomous disease trait as

$$\lambda_R = \frac{P(\text{relative of type R diseased} \mid \text{proband diseased})}{P(\text{disease})}$$

If the disease has a genetic basis, what should λ_R be?

How should λ_R vary with R?

If disease is NOT genetic, what should λ_R be?

Risk Ratio	λ_O	λ_S	λ_M	λ_D	λ_H	λ_N	λ_G	λ_C
Observed	10.0	8.6	52.1	14.2	3.5	3.1	3.3	1.8

Definitions of subscripts: O = offspring; ; S = sibling; M = MZ twins; D = DZ twins; H = half-sibs; N = niece/nephew; G = grandchild; C = first cousins.

Table 4.1: Observed recurrence risk ratios from a sample of families with schizophrenia. Source: Risch (1990a).

How do we use λ_R?

- Justifies doing a genetic study of the disease
- λ_R is the basis for power calculations for many types of linkage analysis
- Compare estimated λ_R to different genetic models
- We will look at how λ_R is calculated in simple Mendelian models

Notation

Disease Phenotype: Y (Y=1 is affected; Y=0 is unaffected)

Genotype at Disease Locus: X=0,1,2 (dd,Dd,DD)

Penetrance functions: $f_x = P(Y = 1 \mid X = x)$

R: Denotes a relative of the proband

p: Frequency of D allele

p(X) frequency of genotypes, p(DD, Dd or dd genotype)

<u>Hardy Weinberg Equilibrium (HWE):</u>

$p(dd) = (1-p)^2$ $p(dD) = 2p(1-p)$ $p(DD) = p^2$

What does λ_R depend on?

Reminder: $\lambda_R = \frac{\text{P(relative of type R diseased | proband diseased)}}{\text{P(disease)}}$

For Simple Mendelian Models:

P(disease) depends only on genotype at a single locus, no other factors influence disease

Denominator:

$$K = P(\text{disease}) = f_0 \cdot (1-p)^2 + f_1 \cdot 2p(1-p) + f_2 \cdot p^2$$

Assumes penetrance functions, allele frequency, HWE

What does λ_R depend on?

What about the numerator ?

P(relative of type R diseased | proband diseased)
= P(both diseased)/K

$\lambda_R = P(\text{both diseased})/K^2$

What does P(both diseased) depend on?

Calculating λ_R

Depends on degree of relationship R, penetrance functions and Mendel's Laws

Example: Consider the sibling recurrence risk ratio and a recessive Mendelian model:

Show that $\lambda_S = [(1+p)/2p]^2$

Step 1: Calculate K

Step 2: Calculate p(both sibs have disease)

Step 3: Calculate λ_S

Calculating λ_S

Denominator: K^2 = ?

Numerator:

Parents	DD (p^2)	Dd $(2pq)$	dd (q^2)
DD (p^2)	1	1/2	0
Dd $(2pq)$	1/2	1/4	0
dd (q^2)	0	0	0

Values in table represent probability of an affected child

$$P(Y_1 = 1, Y_2 = 1) = p^2 \cdot p^2 \cdot 1^2 + p^2 \cdot 2pq \cdot (1/2)^2 + 2pq \cdot p^2 \cdot (1/2)^2 + 2pq \cdot 2pq \cdot (1/4)^2$$

Recurrence Risk Ratio

Recurrence risk to relatives of type R:

How to calculate?

1) Assume a specific genetic model (e.g. single gene, dominant)
2) Assume a frequency for the disease allele p
3) Assume 3 penetrance functions: f_0, f_1, f_2
4) Simple to compute K=P(disease in population)
5) Assume random mating and HWE to get all possible genotypes for common ancestors
6) Use Mendel's Laws to get offspring genotypes phenotypes and to compute P(both relatives affected)
7) Easiest when use Parent-Offspring or Sibling for R, and deterministic Mendelian models

Heritability

- Originally defined for continuous traits; can be adapted to dichotomous disease traits
- Heritability is defined as percent of total trait variance ‘explained’ by genes
- Requires a very specific genetic model explaining how genes affect outcome
- Can be estimated using relative data or case/control GWAS data