# BioStats 710 MidTerm

03-08-2016

*(Chapter one through Chapter 4)*

Jeff Du


## Chapter One and Chapter Two

- trait: Also called phenotype, a genetically determined individual characteristic not observed at molecular level; A trait is a notable feature or quality in a person. Traits are passed from generation to generation.

- genetic marker: A genetic marker is a gene or DNA sequence with a known location on a chromosome that can be measured reliably.

- linkage analysis: Aimed at establishing linkage between genes and disease. study patterns of inheritance of genetic markers of disease (phenotype) to infer genetic distance between marker and disease genes.

- Mendelian disorder: Disease or disorders which are initiated by variants in a single gene. A genetic disease, showing a mendelian pattern of inheritance, and caused by a single mutation in the structure of DNA, which causes a single basic defect that has some pathological consequence or consequences.

- Penetrance: Penetrance in genetics is the proportion of individuals carrying a particular variant of a gene (allele or genotype) that also expresses an associated trait (phenotype)

- heterozygote: An individual having two different alleles of a particular gene or genes, 'A' and 'a' as a pair, and so giving rise to varying offspring.

- polymorphism: Genetic polymorphism is the occurrence in the same population of two or more alleles at one locus, each with appreciable frequency

- gene mapping: Gene mapping describes the methods used to identify the locus of a gene and the distances between genes. The essence of all genome mapping is to place a collection of molecular markers onto their respective positions on the genome.

- homozygous: Alternative forms of a given gene are called alleles, and they can be dominant or recessive. When an individual has two of the same allele, whether dominant or recessive, they are homozygous, 'AA' or 'aa'.

- recessive: The terms dominant and recessive describe the inheritance patterns of certain traits. For a recessive allele to produce a recessive phenotype, the individual must have two copies, one from each parent. An individual with

one dominant and one recessive allele for a gene will have the dominant phenotype. They are generally considered carriers of the recessive allele: the recessive allele is there, but the recessive phenotype is not.

- dominant: The terms dominant and recessive describe the inheritance patterns of certain traits. A dominant allele produces a dominant phenotype in individuals who have one copy (or two copies) of the allele, which can come from just one parent.

- allele: An allele is one of the possible forms of a gene. Most genes have two alleles, a dominant allele and a recessive allele. If an organism is heterozygous for that trait, or possesses one of each allele, then the dominant trait is expressed.

- complex trait: Complex traits are believed to result from variation within multiple genes and their interaction with behavioral and environmental factors. Complex traits do not follow readily predictable patterns of inheritance.

- full penetrance: "Complete" or Full penetrance means the gene or genes for a trait are expressed in all the population who have the genes. "Incomplete" penetrance means the genetic trait is expressed in only part of the population.

- proband: "Proband", "proposito" (male proband), or "proposita" (female proband) is a term used most often in medical genetics and other medical fields to denote a particular subject (person or animal) being studied or reported on. Family samples with at least one individual affected.

- phenotype: A phenotype is an individual's observable traits, such as height, eye color, and blood type. The genetic contribution to the phenotype is called the genotype. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors

- autosomes: An autosome is any chromosome that is not a sex-determining chromosome, so most chromosomes are autosomes.

- homologous chromosomes: One chromosome of each homologous pair comes from the mother (called a maternal chromosome) and one comes from the father (paternal chromsosome). Homologous chromosomes are similiar but not identical. Each carries the same genes in the same order, but the alleles for each trait may not be the same.

- centromere: The centromere is the part of a chromosome that links sister chromatids or a dyad. During mitosis, spindle fibers attach to the centromere via the kinetochore.

- gene: A gene is a locus (or region) of DNA that encodes a functional RNA or protein product, and is the molecular unit of heredity.A gene is a locus (or region) of DNA that encodes a functional RNA or protein product, and is the molecular unit of heredity.

- exon: An exon is any part of a gene that will become a part of the final

mature RNA produced by that gene after introns have been removed by RNA splicing. The term exon refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts.

- intron: Introns are noncoding sections of an RNA transcript, or the DNA encoding it, that are spliced out before the RNA molecule is translated into a protein. The sections of DNA (or RNA) that code for proteins are called exons.

- mutation: In biology, a mutation is a permanent alteration of the nucleotide sequence of the genome of an organism, virus, or extrachromosomal DNA or other genetic elements.

- splice site: the specific site at which splicing takes place during the processing of precursor messenger RNA into mature messenger RNA.

- SNP: A single nucleotide polymorphism, often abbreviated to SNP (pronounced snip; plural snips), is a variation in a single nucleotide that occurs at a specific position in the genome, where each variation is present to some appreciable degree within a population (e.g. $> 1\%$).

- SNV: A single nucleotide variation is just a variation in a single nucleotide without any limitations of frequency.

- somatic mutation: Somatic mutation, genetic alteration acquired by a cell that can be passed to the progeny of the mutated cell in the course of cell division. Somatic mutations differ from germ line mutations, which are inherited genetic alterations that occur in the germ cells (i.e., sperm and eggs).

- germ-line mutation: A germline mutation is any detectable and heritable variation in the lineage of germ cells. Mutations in these cells are transmitted to offspring, while, on the other hand, those in somatic cells are not.

- de novo variant: An alteration in a gene that is present for the first time in one family member as a result of a mutation in a germ cell (egg or sperm) of one of the parents or in the fertilized egg itself

- ascertainment: In epidemiologic and genetic research, the method by which a person, pedigree, or cluster is brought to the attention of an investigator; ascertainment has a bearing on the interpretation of segregation ratios, concordance rates, linkage analysis, and other probability features.

- germline mutation: A germline mutation is any detectable and heritable variation in the lineage of germ cells. Mutations in these cells are transmitted to offspring, while, on the other hand, those in somatic cells are not.

- VNTRs: A variable number tandem repeat (or VNTR) is a location in a genome where a short nucleotide sequence is organized as a tandem repeat. These can be found on many chromosomes, and often show variations in length between individuals.

- indel: An insertion/deletion polymorphism, commonly abbreviated indel, is a type of genetic variation in which a specific nucleotide sequence is present (insertion) or absent (deletion). While not as common as SNPs, indels are widely spread across the genome.

- structural variants: Structural variation (SV) is generally defined as a region of DNA approximately 1 kb and larger in size and can include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs)

- codon: a sequence of three nucleotides that together form a unit of genetic code in a DNA or RNA molecule.

- synonymous/nonsynonymous variant: synonymous variant is the evolutionary substitution of one base for another in an exon of a gene coding for a protein, such that the produced amino acid sequence is not modified. A nonsynonymous variant is a nucleotide mutation that alters the amino acid sequence of a protein.

- missense mutation: In genetics, a missense mutation is a point mutation in which a single nucleotide change results in a codon that codes for a different amino acid. It is a type of nonsynonymous substitution.

- nonsense mutation: In genetics, a nonsense mutation is a point mutation in a sequence of DNA that results in a premature stop codon, or a nonsense codon in the transcribed mRNA, and in a truncated, incomplete, and usually nonfunctional protein product.

- silent mutation: Silent mutations are mutations in DNA that do not significantly alter the phenotype of the organism in which they occur. Silent mutations can occur in non-coding regions (outside of genes or within introns), or they may occur within exons.

- Medelian transmission: Mendelian inheritance is inheritance of biological features that follows the laws proposed by Gregor Johann Mendel.

- genetic model: Some analyses of genetic data; ideas on how genes work in individuals to affect phenotypes; Genetic Models specifically relate genotype to phenotype. From the TEXT: A genetic model specifies a probability distribution for the trait, conditional on the underlying genotype at the hypothesized disease locus.

- aggregation analyses: Aggregation Analysis (for dichotomous traits): By estimating the correlation or similarity of a phenotype among family members, one can assess whether a phenotype aggregates in families.

- chromatids: A chromatid, is one copy of a newly copied chromosome which is still joined to the other copy by a single centromere. Before replication, one chromosome is composed of one DNA molecule.

- chiasma: A chiasma (plural: chiasmata), in genetics, is thought to be the point where two homologous non-sister chromatids exchange genetic material during chromosomal crossover during meiosis (sister chromatids also form chiasmata between each other (also known as a chi structure), but because their genetic material is identical, it does not cause any change in the resulting daughter cells).

- meiosis: Meiosis is a process where a single cell divides twice to produce four cells containing half the original amount of genetic information. These cells are our sex cells sperm in males, eggs in females.

- mitosis: Mitosis is a part of the cell cycle in which chromosomes in a cell nucleus are separated into two identical sets of chromosomes, and each set ends up in its own nucleus.

- diploid: of a cell or nucleus) containing two complete sets of chromosomes, one from each parent

- haploid: Haploid cells are a result of the process of meiosis. Haploid cells have half the number of chromosomes (n) as diploid - i.e. a haploid cell contains only one complete set of chromosomes.

- gametes: a mature haploid male or female germ cell that is able to unite with another of the opposite sex in sexual reproduction to form a zygote.

- zygote: Zygote is a eukaryotic cell formed by a fertilization event between two gametes. The zygote's genome is a combination of the DNA in each gamete, and contains all of the genetic information necessary to form a new individual.

- tetrad: A group of four chromatids formed from each of a pair of homologous chromosomes that split longitudinally during the prophase of meiosis.

- interference: the crossovers in adjacent chromosome regions independent or does a crossover in one region affect the likelihood of there being a crossover in an adjacent region. It turns out that often they are not independent: the interaction is called interference.

- obligatory crossover: The obligate crossover refers to the fact that, in most species, it is rare to find chromosomes that do not undergo crossing-over.

- recombination: the rearrangement of genetic material, especially by crossing over in chromosomes or by the artificial joining of segments of DNA from different organisms.

- population genetics: Population genetics is the study of the distribution and change in frequency of alleles within populations, and as such it sits firmly within the field of evolutionary biology.

- genetic epidemiology: the study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such

genetic factors with environmental factors. Genetic epidemiology seeks to derive a statistical and quantitive analysis of how genetics work in large groups.

• mode of inheritance: The term Mode of Inheritance refers to exactly how parameters of the distribution of Y depend on the number of disease alleles. we generally use the mode of inheritance to indicate how the parameters of the penetrance function depend on the number of disease alleles.

• Mather's formula: The relationship between $\Theta$ and the distribution of crossovers is given by Mathers law: $\Theta = (1 - P_0)/2$ where $P_0$ is the probability of zero crossovers.

• Heterozygote advantage: A heterozygote advantage (heterozygous advantage) describes the case in which the heterozygote genotype has a higher relative fitness than either the homozygote dominant or homozygote recessive genotype. The specific case of heterozygote advantage due to a single locus is known as overdominance.

• recombination fraction: Recombination fraction (also called recombination frequency) between two loci is defined as the ratio of the number of recombined gametes to the total number of gametes produced. Recombination fraction $\Theta$ is given by P(recombination occurs between two loci).

• Questions: • 2. Name two key choices that Mendel made that allowed him to establish his laws of inheritance

Answer: One of the keys to success for Mendel was his selection of pea plants, self-fertilize. The other, Mendel concentrated in one or few characters at a time.

• 3. Explain what Mendel's 1st law of segregation is all about?

Answer: Mendels First Law (Segregation): One allele of each parent is randomly and independently selected, with probability 1/2, for transmission to the offspring; the alleles unite randomly to form the offsprings genotype.

• 4. Explain Mendel's second law of independent assortment and why it is often false?

Answer: Mendels Second Law (Independent Assortment): The alleles underlying two or more different traits are transmitted to offspring independently of each other; the transmission of each trait separately follows the first law of segregation.

Some genes are located on the same chromosome, thus the lack of independent transmission of different genes is, in fact, fortuitous, as it provides the basis for mapping disease genes by linkage analysis.

- Hardy-Weinberg Equilibrium: The HardyWeinberg principle, also known as the HardyWeinberg equilibrium, model, theorem, or law, states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.

$$P(GenotypeFrequency) \begin{cases} P(AA - genotype) = p^2 \\ P(Aa - genotype) = 2pq \\ p(aa - genotype) = q^2 \end{cases}$$

- loss of heterozygosity: Loss of heterozygosity (LOH) is a common genetic event in cancer development, and is known to be involved in the somatic loss of wild-type alleles in many inherited cancer syndromes.

- population structure: population structure, population geneticists mean that, instead of a single, simple population, populations are subdivided in some way.

- population stratification: Population stratification is perhaps the simplest form of population substructure,as it coincides with the intuitive notion that individuals in a population can be subdivided into mutually exclusive strata; within each strata the allele frequency is the same for all individuals, but it varies between strata.

- admixture: Population admixture refers to a situation where individuals in a population have a mixture of different genetic ancestries due to the mixing of two or more populations at a previous point in time.

- selection:
In HWE, selection (meaning that certain alleles do not confer a selective advantage or disadvantage in reproduction).

- inbreeding: Population inbreeding occurs when there is a preference for mating among relatives in a population or because geographic isolation of sub-groups restricts mating choices.

- drift: drift occurring in small populations, and mutations,as well as understanding genetic differences in populations

- $F_{st}$ :

- inbreeding coefficient: The inbreeding coefficient, denoted by F, is the probability that a random individual in the population inherits two copies of the same allele from a common ancestor. In large, randomly mating populations

the chances that any two mating parents have a common ancestor allele is low, hence F is negligible and often considered to be zero.

- ancestry informative marker (AIM): For those with the highest degree of Indian ancestry, the allele frequency is almost zero, whereas it is almost 70% for those with no great grandparents with Indian Heritage. Such a marker with strong differences among population subtypes is called an Ancestry Informative Marker (AIM).

3. Given a biallelic locus with alleles a and A, express the probability that a randomly chosen allele from the population will be an A in terms of the population genotype frequencies

4. Give the allele counting estimator $\hat{p}$ for the population allele frequency p

5. Assuming HWE, give a variance estimator for $\hat{p}$

• identity by descent: Gene identity by descent (IBD) is a fundamental concept that underlies genetically mediated similarities among relatives. Gene IBD is traced through ancestral meioses and is defined relative to founders of a pedigree, or to some time point or mutational origin in the coalescent of a set of extant genes in a population.

Two alleles are IBD if one is a physical copy of the other, or if they are both physical copies of the same ancestral allele.

• aggregation analysis: Aggregation Analysis (for dichotomous traits): By estimating the correlation or similarity of a phenotype among family members, one can assess whether a phenotype aggregates in families. While a positive result of an aggregation analysis confirms the plausibility of a disease gene, it cannot rule out common environmental effects within families as the origin for the observed correlations.

• recurrence risk ratio: The strength of the genetic aggregation among relatives is generally measured by the recurrence risk ratio. It is defined as a probability ratio which compares the probability of a study subject being affected given that a relative is affected to the general risk in the population. The general risk in the population is commonly referred to as the population prevalence of the disease, and denoted by K. For the relative of an affected individual, the recurrence risk ratio is thus:

$\lambda_R = \frac{P(Y_2=1|Y_1=1)}{K}$

• prevalence: The total number of cases of a given disease in a specified population at a designated time. It is differentiated from incidence, which refers to the number of new cases in the population at a given time.

• attributable fraction: The attributable fraction assesses the genetic effect relative to the disease prevalence and is defined by

$AF = \frac{K-f_0}{K} = 1 - \frac{P(Y=1|no-risk-alleles)}{P(Y=1)}$

• narrow-sense heritability: the narrow sense heritability ($h^2$) is defined as the proportion of the phenotypic variance that is explained by just the additive genetic effects,

$h^2 = V_A/Var(Y) = \frac{V_A}{Var(Y)}$

The advantage of the narrow sense heritability is that it can be directly estimated from the phenotypic data on relatives.

• broad-sense heritability: the broad sense heritability of a trait is defined as the proportion of the overall phenotypic variation in the trait that is attributable to genetic comonents, e.g.,

$Var(G)/Var(Y) = \frac{Var(G)}{Var(Y)}$.

• genotypic values (in the context of the quantitative trait model):

The genotypic value is the phenotype exhibited by a given genotype averaged across environments. A related concept is the breeding value, which is the portion of the genotypic value that determines the performance of the offspring.

Genotypic value is property of the genotype and therefore is a concept that describes the value of genes to the individual, whereas breeding value describes the value of genes to progeny and therefore helps us understand how a trait is inherited and transmitted from parents to offspring . Remember that only additive genetic effects can be passed on to progeny. Non-additive genetic effects and environmental effects cannot be inherited by offspring.

G = A + D + I

where G is the genotypic value of all loci considered together, A is sum of all additive effects (i.e., breeding values) for separate loci, D is the sum of all dominance deviations (i.e., interaction between alleles at a locus or so-called intralocus interactions) and I is the interaction of alleles among loci (also referred to as the interaction deviation or epistatic deviation).

• breeding values: An individuals breeding value can also be described as the sum of the average effect of the individuals alleles. For example, if an A1 allele is worth +5 and an A2 allele is worth -2, then an A1A2 heterozygote has a breeding value of 3. Breeding values are expressed as a deviation of the population mean (with the population mean dependent on genotypic values and frequencies)

• kinship coefficient:

$\Phi_{ij}$ The kinship coefficient $\Phi_{ij}$ between two individuals i and j is the probability that an allele selected randomly from i and an allele selected randomly from the same autosomal locus of j are IBD.

2. Why is narrow-sense heritability used more often than broad-sense?

3. Relate the phenotypic covariance between two relatives in terms of genetic components of variance and measures of relatedness. De
ne all quantities. What assumptions does this relationship depend on?

4. Explain how heritability is relative to environment

5. Assume that you have left foot length of a collection of fathers and sons. Describe how you would go about estimating narrow-sense heritability

6. What is the goal of segregation analysis?

The idea of segregation analysis is to test and to compare different statistical models formally, using phenotypic data on related individuals, with the goal of identifying the genetic model that describes the data best.

7. Text, section 4.5, problems 1-5, 11 and 12.
Answer Problem 2 (a) of chapter 4

$Cov(Y_1, Y_2)$
$= E((Y_1 - \mu_1) * (Y_2 - \mu_2))$
$= E(Y_1 * Y_2 - Y_1 * \mu_2 - Y_2 * \mu_1 + \mu_1 * \mu_2)$
$= E(Y_1 * Y_2) - \mu_2 * E(Y_1) - \mu_1 * E(Y_2) + \mu_1 * \mu_2$
    Here:    $E(Y_1) = \mu_1, E(Y_2) = \mu_2$
$= E(Y_1 * Y_2) - \mu_1 * \mu_2$
$here \mu_1 and \mu_2\ Binomail(0, p)$

$E(Y_1 * Y_2) = \sum_{y_1} \sum_{y_2} y_1 * y_2 * P(Y_1, Y_2)$
$= P(Y_1 = 1, Y_2 = 1) - \sum_{y_1}^{max} y_1 * P(Y_1) * \sum_{y_2} y_2 * P(Y_2)$
$Only\quad when\quad y_1 = 1\quad or\quad y_2 = 1\quad could\quad count$

$= P(Y_1 = 1, Y_2 = 1) - P(Y_1)P(Y_2)$
$= P(Y_1 = Y_2 = 1) - K^2$

Answer Problem 2 (b) of chapter 4
$\lambda_R = \frac{P(Y_2=1|Y_1=1)}{K}$
Here we plug in 2 (a)
$P(Y_1 = 1|Y_2 = 1) * P(Y_2 = 1) = P(Y_1 = 1, Y_2 = 1)$

$\lambda_R = \frac{P(Y_2=1,Y_1=1)}{P(Y_1=1)*K}$
$\lambda_R = \frac{P(Y_2=1,Y_1=1)}{K^2}$
$\lambda_R = \frac{Cov(Y_1,Y_2)+K^2}{K^2}$