Module 4 - Principles of Inheritance

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Outline

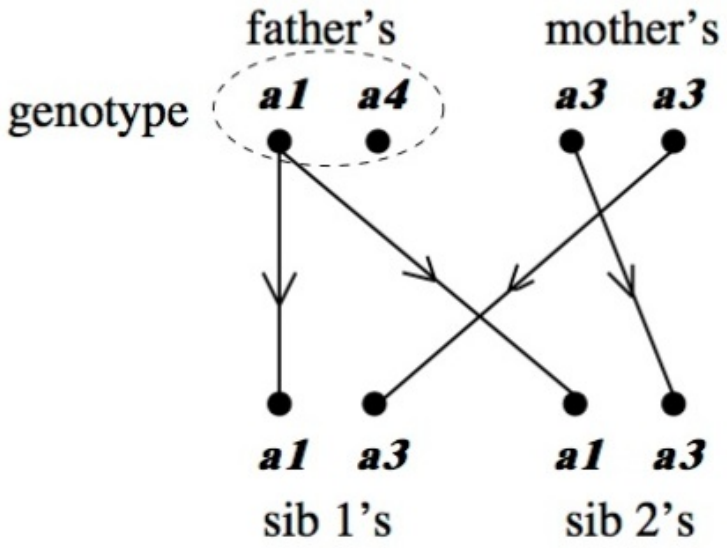
Chapter 2 - Principles of Inheritance: Mendel (First) Law and Genetic Models

- Mendel first law of segregation
- Related distribution of offsprings genotype conditional upon parental genotypes
- Mendel's garden pea study and testing Mendel first law
- Simple Mendelian genetic model and more general mode of inheritance - how phenotype Y depends on genotype G.
- Recessive, dominant, co-dominant, additive models etc.
- Penetrance function for binary trait

- Penetrance function for quantitative traits using normal distributions
- Graphic illustrations for Y and G dependence/association studies as the classic regression analyses of Y on G

Mendel's First Law of Segregation

- It is a model for **single locus inheritance**: *underlies the concept of Mendelian transmissions of alleles from one generation to the next.*
- *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 offspring genotype.*
- Note that transition between the two sibs are assumed to be independent of each other as well.



Segregation and Conditional Probability

- The simple Mendelian segregation model leads to the probabilities in **Table 2.1.** *Distribution of offspring genotype conditional upon parental genotypes*.

$$P(G \text{ of offspring} | G \text{ of the two parents}).$$

- Formally, let P_1 and P_2 be the random variables representing the genotypes of two parents, and X be the genotype of an offspring,

$$P(X = x | P_1 = g_1, P_2 = g_2),$$

where $x, g_1, g_2 \in \{aa, Aa, AA\}$.

- We can simply the table to consider **6 parental mating types** (for a biallelic marker)

6 parental mating type	Offspring Genotype (in probability)		
	AA	Aa	aa
AA x AA	1	0	0
AA x Aa	1/2	1/2	0
AA x aa	0	1	0
Aa x Aa	1/4	1/2	1/4
Aa x aa	0	1/2	1/2
aa x aa	0	0	1

Ordered vs. Unordered Genotypes

- Note that in the case when the genotypes of the two parents are unordered, then e.g. (AA, Aa) is the same as (Aa, AA).

- This usually does not make a difference in the **conditional probability**

$$P(X = Aa|\text{ Parents } = Aa, AA) = P(X = Aa|P_1 = Aa, P_2 = AA)$$

$$= P(X = Aa|P_1 = AA, P_2 = Aa) = \frac{1}{2}.$$

(In some analysis, we may want to distinguish the two parents: Genomic imprinting and parent-of-origin effects on complex traits: *Parent-of-origin effects occur when the phenotypic effect of an allele depends on whether it is inherited from the mother or the father.*)

- But, it can make a difference in the **marginal probability** (which we will see the importance of it later)

$$P(\text{Parents } = Aa, AA) = P(P_1 = Aa, P_2 = AA) + P(P_1 = AA, P_2 = Aa)$$

$(= 2p(1-p)p^2 + p^2 2p(1-p) = 4p^3(1-p),$ assuming random mating and HWE).

Testing Mendel's First Law - Mendel's Garden Pea Study

- Trait of interest: colour of unripe pods
  yellow (AA Aa), or green (aa)

- Design and data (also See Textbook Figure 2.1):

```
F1:      Aa       x    Aa
        (yellow)     (yellow)
                  |
                  V

F2:  AA(1/4)  Aa(1/2)    aa(1/4)  <- Mendel's first law
     ---------------     --------
         (yellow)         (green)

Obs.     705                224
```

Testing Mendel's First Law - Formulation of the Problem

- Assume the total number of peas is fixed $n = 705 + 224 = 929$.
- Let X denote the number of green peas.
  (X is the random variable, and $x = 224$ is the observed data/count.)
- What is a reasonable distribution for X?

$$X \sim Bino(n, \theta).$$

- **Translate Mendel's First Law into a null hypothesis**

$$H_0 : \theta = \frac{1}{4}.$$

Testing Mendel's First Law

- Apply the Pearson χ^2 test:

$$T = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_1$$

$$T_{obs} = \frac{(224 - 929 \cdot \frac{1}{4})^2}{929 \cdot \frac{1}{4}} + \frac{(705 - 929 \cdot \frac{3}{4})^2}{929 \cdot \frac{3}{4}} = 0.39$$

$$\Longrightarrow \text{p-value} = P(T \geq T_{obs} | \chi^2_1) = 0.53.$$

Note that E_is are counts Expected under the null hypothesis H_0.

- Alternative tests
  - Testing the binomial parameter using normal approximation.
  - Can you conduct a likelihood ratio test?
  - Other tests?

Mendel vs. Fisher

- Mendel's studies of many other traits gave similar results.
- Fisher (1936)'s commented that the observed counts in different classes **might fit too well** to the expected counts in all Mendel's studies, suggesting some selection or adjustment of the data?
  (Of course, Mendel's discovery is still brilliant and important!)

	Chi^2_1	p-value
form of seed	.26	.61
color of albumen	.01	.90
color of flower	.39	.53
form of pods	.06	.80
color of unripe pods	.39	.53
position of flowers	.35	.55
length of stem	.61	.44

sum of Chi^2_1 = 2.13 on 7 d.f. p-value = 0.95

- Related to the topic of multiple hypothesis testing, STA4515.

- There would also be **Mendel's Second Law** for multi-locus inheritance which turned out be only **partially true**; we delay this topic and materials in Textbook Section 2.3 until we talk about genetic map and linkage.

Simple (Mendelian) Genetic Models I

- Notes on notations
  - For a biallelic marker (A and a): a is the wild/normal allele, typically assumed to be more frequent than A.
  - For a genetic marker /locus that is **Disease Susceptibility Locus (DSL)** (for binary trait) or **Quantitative Trait Locus (QTL)**, we often use D and d; D would be the mutation allele or Disease Susceptibility Allele.

- **Autosomal Mendelian models**

```
                Genotype
              DD    Dd    dd
Dominant       1     1     0    <- P(affected|genotype)
Recessive      1     0     0

e.g. Huntington's disease (dominant)
     Cystic Fibrosis disease (recessive)
```

Simple (Mendelian) Genetic Models II

- Sex-linked Mendelian models

	Genotype			
	DD	Dd	dd	
X-linked Dominant	1	1	0	<- P(female affected|genotype)
X-linked Recessive	1	0	0	
	D*	d*		
X-linked Dominant	1	0		<- P(male affected|genotype)
X-linked Recessive	1	0		
	*D	*d		
Y-linked Dominant	1	0		<- P(male affected|genotype)
Y-linked Recessive	1	0		

Simple (Mendelian) Genetic Models III

e.g.1 color-blindness is recessive on X chromosome.

Case I		Case II	
father	mother	father	mother
(normal)	(color-blind)	(normal)	(normal;carrier)
d*	DD	d*	Dd
daughters: all normal (Dd)		daughters: all normal (Dd or dd)	
sons: all color-blind (D*)		sons: half normal (d*); half(D*)	

e.g.2 hairy ears is thought to be due to a gene on Y chromosome.

- There are other complications of analyzing X-chromosome, e.g. the issue of X-inactivation: one of the two alleles for a female might be inactivated; an area that has not been studies extensively in statistical genetics.

Penetrance and More Complex Models for Binary Traits

- **Penetrance function**
  - Let Y be the (binary) phenotype of interest,
    i.e. Y is a rv and $Y = 1$ denotes an individual is affected by the disease.
  - Let G be the genotype of the Disease Susceptibility Locus.

$$f_0 = P(Y = 1|G = dd), \quad f_1 = P(Y = 1|G = Dd), \quad f_2 = P(Y = 1|G = DD).$$

- Complete penetrance or simple Mendelian disease model (autosome):
  - Recessive model
    $P(Y = 1|G = dd) = 0, \quad P(Y = 1|G = Dd) = 0, \quad P(Y = 1|G = DD) = 1.$
  - Dominant model
    $P(Y = 1|G = dd) = 0, \quad P(Y = 1|G = Dd) = 1, \quad P(Y = 1|G = DD) = 1.$

- **Incomplete penetrance**

$$0 < P(Y = 1|G) < 1, e.g.$$

reduced penetrance: $0 < P(Y = 1|G = DD) < 1$

phenocopy: $0 < P(Y = 1|G = dd) < 1.$

More Complex Genetic Models - Example

- *APOE* gene and Alzheimer's Disease (AD) (Textbook Page 21-22).
  - 3 alleles: E2, E3, E4.
  - 6 (unordered) genotypes G: E2E2, E3E3, E4E4, E2E3, E2E4, E3E4 (Recall the general formula for calculating the number of genotypes.)
  - Incomplete penetrance: $0 < P(Y = 1|G) < 1$, In particular

$$P(Y = 1|G = E3E3) \approx 20\%$$

$$P(Y = 1|G = E2E4) \approx 50\%$$

$$P(Y = 1|G = E3E4) \approx 60\%$$

$$P(Y = 1|G = E4E4) \approx 90\%$$

  - The risk of AD increases as the number of the E4 allele increases.

- How do we (accurately) estimate the penetrance in practice?
  - Sample size consideration
  - Prospective vs. retrospective, sampling scheme.
  - **Effects of other genes and covariates such as age and sex.**

General Mode of Inheritance I

- **Recessive model**

$$f_1 = P(Y = 1|dD) = P(Y = 1|dd) = f_0.$$

(Simple Mendelian recessive disease further assumes $f_1 = f_0 = 0$ and $f_2 = 1$)

- **Dominant model**

$$f_1 = P(Y = 1|dD) = P(Y = 1|DD) = f_2.$$

- **Co-dominant model**

f_1 is somewhere between f_0 and f_2

Note that if we assume $Y = 1$ is not a good thing (at risk) and mutation D allele increases this risk, then we can say

$$f_0 < f_1 < f_2$$

(With binary trait, we can always switch the definition of $Y = 1$, then we would have $f_2 < f_1 < f_0$.)

General Mode of Inheritance II

- **Additive model** is a special case of co-dominant model:

  f_1 is the 'average' of f_0 and f_2

  - Linear scale:

$$f_1 = P(Y = 1|dD) = \frac{P(Y = 1|dd) + P(Y = 1|DD)}{2} = \frac{f_0 + f_2}{2}.$$

  - Log (or multiplicative) scale:

$$f_1 = P(Y = 1|dD) = \sqrt{P(Y = 1|dd) \times P(Y = 1|DD)} = \sqrt{f_0 \times f_2}.$$

    Note that

$$log(f_1) = log(\sqrt{f_0 \times f_2}) = \frac{log(f_0) + log(f_2)}{2}.$$

- **Heterozygote advantage model** (or disadvantage model)

  $f_1 <$ both f_0 and f_2 (or $>$ both)

Penetrance and Genetic Models for QuantitativeTraits I

- Keep the phenotype Y and genotype G notations, except now Y is now a quantitative (continuous) trait, e.g. blood pressure, height etc.
- For simplicity let's assume there is no environmental effect, and ust one G.
- But we still have to think about how do we code the G.
- If we **code** $G = 0, 1, 2$ to denote the number of D allele in the genotype, then

$$Y = \alpha + \beta G + e, \;\; e \sim N(0, \sigma^2).$$

- And this model implies that

$$(Y|G = dd) = (Y|G = 0) \sim N(\alpha, \sigma^2), \;\; E(Y|G = 0) = \mu_0 = \alpha,$$
$$(Y|G = dD) = (Y|G = 1) \sim N(\alpha + \beta, \sigma^2), \;\; E(Y|G = 1) = \mu_1 = \alpha + \beta,$$
$$(Y|G = DD) = (Y|G = 2) \sim N(\alpha + 2\beta, \sigma^2), \;\; E(Y|G = 2) = \mu_2 = \alpha + 2\beta.$$

- So by coding $G = 0, 1, 2$ we actually assumed an **additive model**

$$\mu_1 = \frac{\mu_0 + \mu_2}{2}.$$

Penetrance and Genetic Models for QuantitativeTraits II

- Without the additive restriction, we can use the following **genotypic model**:

$$Y = \alpha + \beta_1 I(G = dD) + \beta_2 I(G = DD) + e, \;\; e \sim N(0, \sigma^2),$$

where $I(.)$s are the indicator variables (defining dummy variables would lead to the same result). Then

$$\begin{aligned}
(Y|G = dd) &\sim N(\alpha, \sigma^2), \;\; \mu_{dd} = \alpha, \\
(Y|G = dD) &\sim N(\alpha + \beta_1, \sigma^2), \;\; \mu_{dD} = \alpha + \beta_1, \\
(Y|G = DD) &\sim N(\alpha + \beta_2, \sigma^2), \;\; \mu_{DD} = \alpha + \beta_2.
\end{aligned}$$

- Putting different constraints on β_1 and β_2 would lead to different models, e.g.
  - Recessive: $\beta_1 = 0$
  - Dominant: $\beta_1 = \beta_2$
  - Additive: $\beta_1 = \beta_2/2$

Penetrance and Genetic Models for QuantitativeTraits III

- More generally we can just say the density function of $Y|G$, $f(y|G)$, is the density function of $N(\mu_G, \sigma^2)$, i.e.

$$(Y|G) \sim N(\mu_G, \sigma^2),$$

$$E(Y|G) = \mu_G, \;\; Var(Y|G) = \sigma^2.$$

- A side note on the terminology
  - **Genetic model** refers to the entire density (or penetrance) function, $f(y|G)$ (typically normal for continuous traits but does not have to be).
  - **Mode of Inheritance** specifies the relationship between the parameters e.g.
    Quantitative Additive Model: $\mu_1 = \frac{\mu_0+\mu_2}{2}$.
    Binary Additive Model: $p_1 = \frac{p_0+p_2}{2}$.

Penetrance and Genetic Models for QuantitativeTraits IV

- We would discuss this later in more details, but at this moment, we can see that **genetic association studies** between Y and G boils down to this type of **regression analyses** and testing $H_0 : \beta = 0$!

- For binary trait, we can also formulate the question as a regression one, albeit logistic regression; we will discuss this in details (e.g. why the log? what's the interpretation of β in this case?) later in association analysis.

$$log\left(\frac{P(Y=1|G)}{1-P(Y=1|G)}\right) = \alpha + \beta G.$$

QuantitativeTraits - Graphic Example I

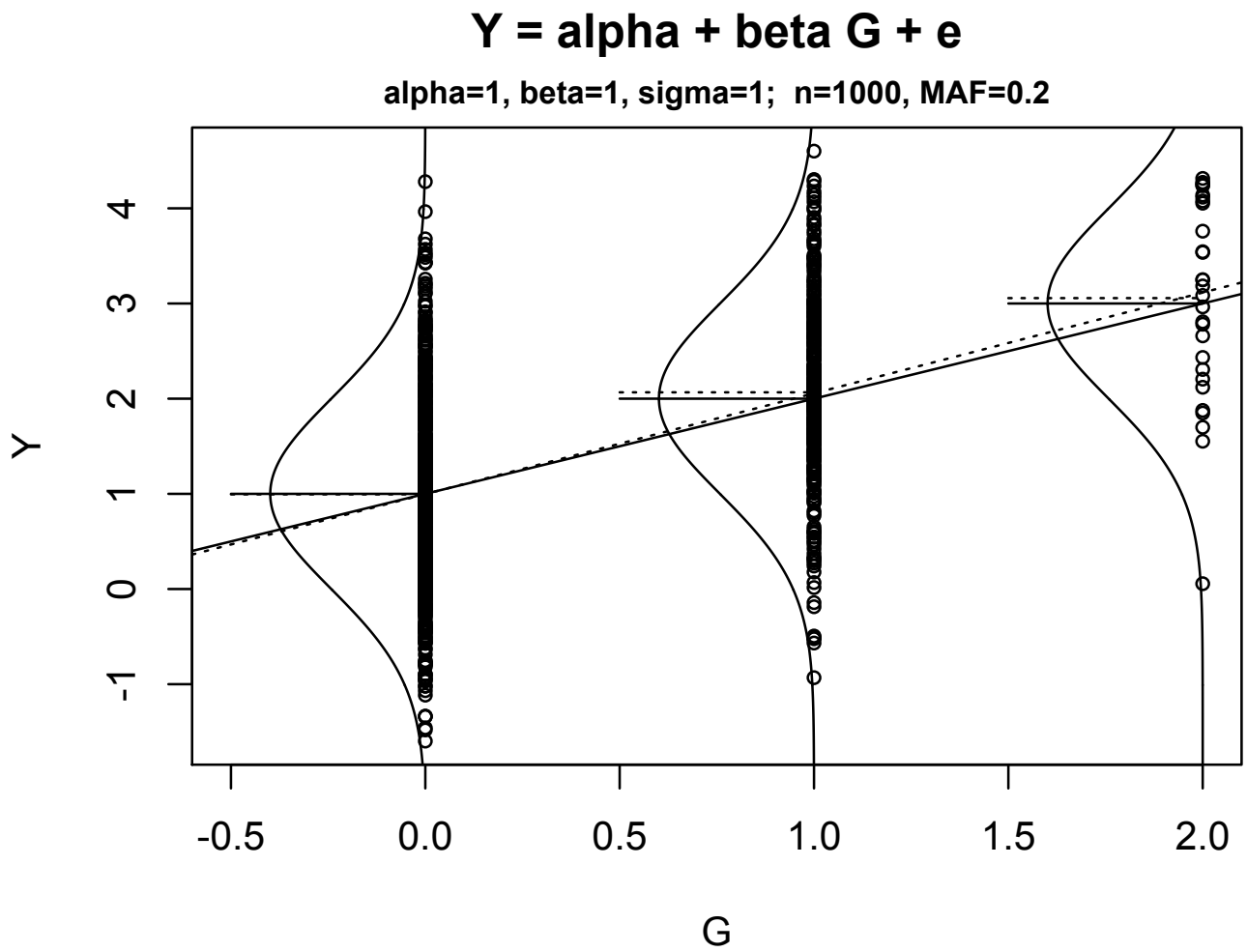
- Back to the simple additive normal model,

$$Y = \alpha + \beta G + e.$$

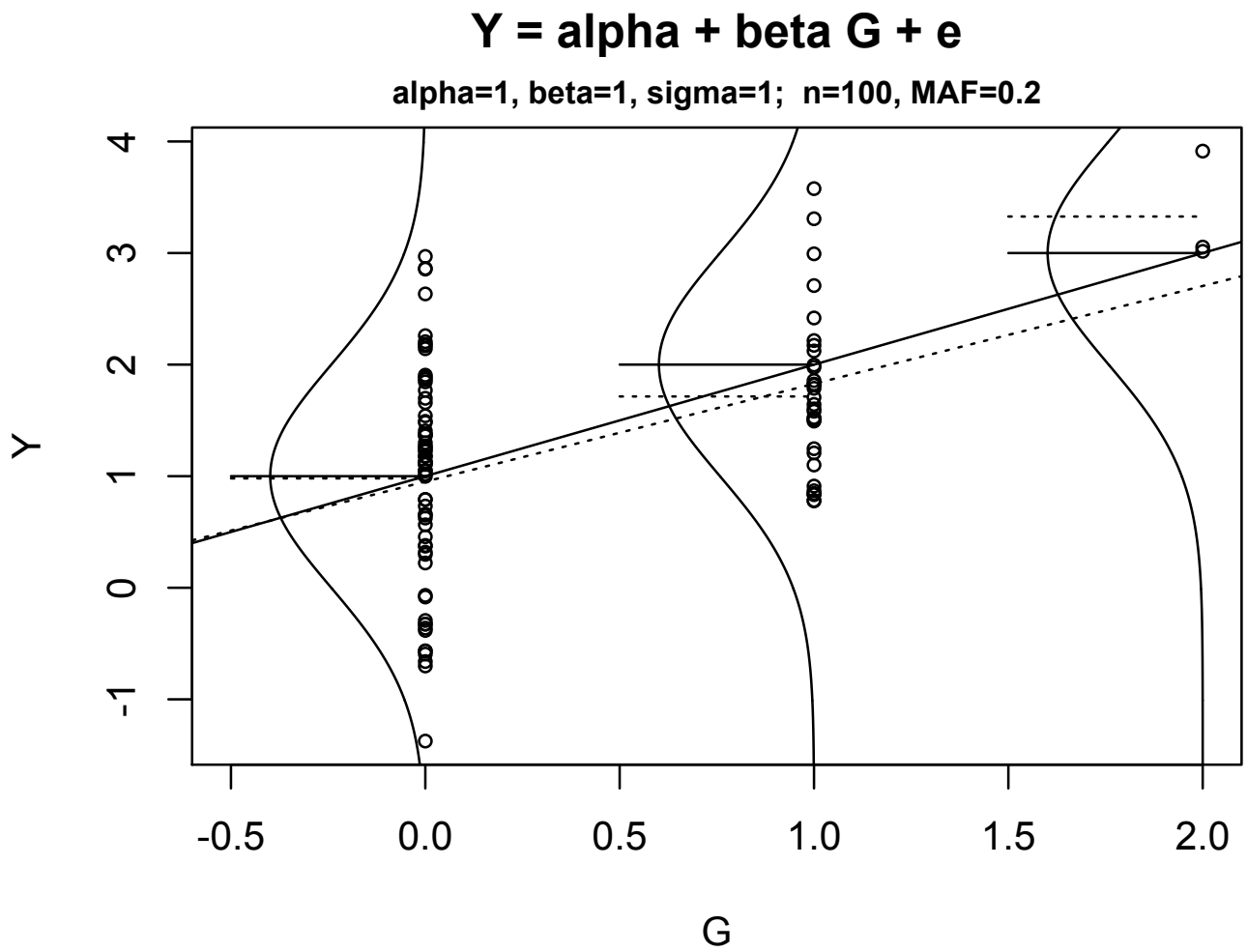
- To obtain some graphic illustrations, let's assume
  - $\alpha = 1, \beta = 1, \sigma = 1$.
  - allele frequency of D is $p = 0.2$
  - genotype frequency of G follow HWE.

- The following graphs are somewhat more intuitive in the context of regression than Figure 2.3 of the Textbook which simply shows that centres of the distribution (μ_G) differ between G.
- The R codes used to make the graphs.

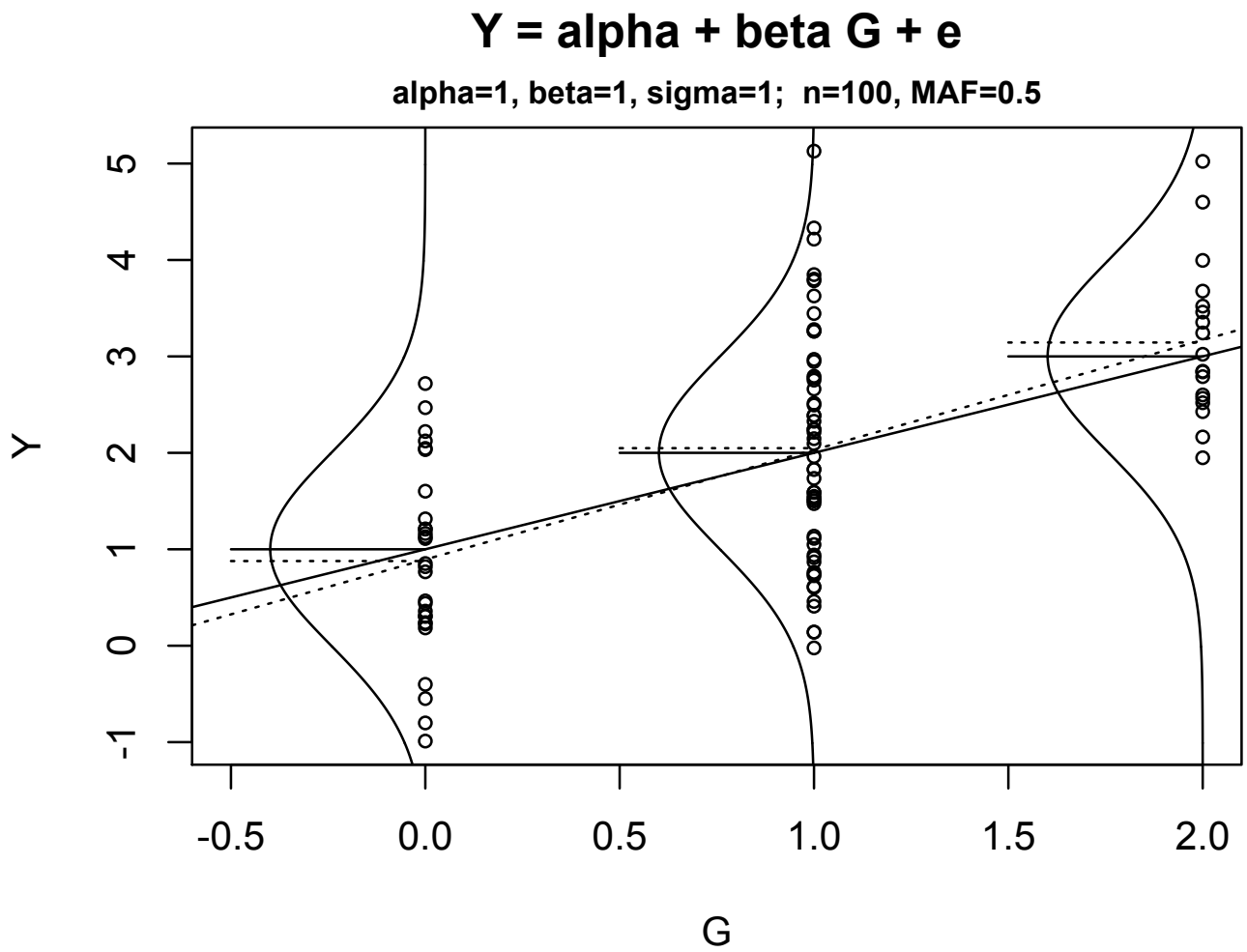
QuantitativeTraits - Graphic Example II



QuantitativeTraits - Graphic Example III

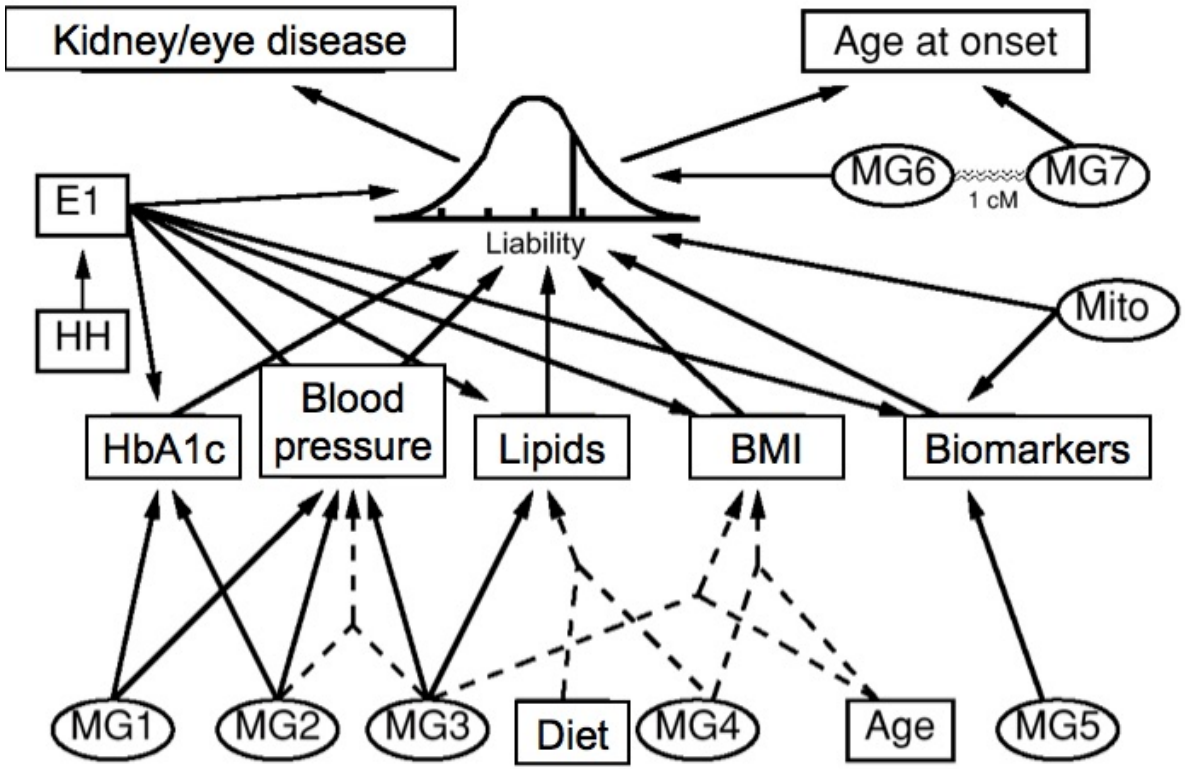


QuantitativeTraits - Graphic Example IV



QuantitativeTraits - More Complex Models I

- For complex traits, underlying models are much more complex even we assume a normal model: multiple *G*s and *E*s, and GxG and GxE interactions.
- e.g. Genetics of diabetic complications and their risk factors (Dr. Andrew Paterson, HSC).



MG=major gene, E=environmental factor, HH=household effect

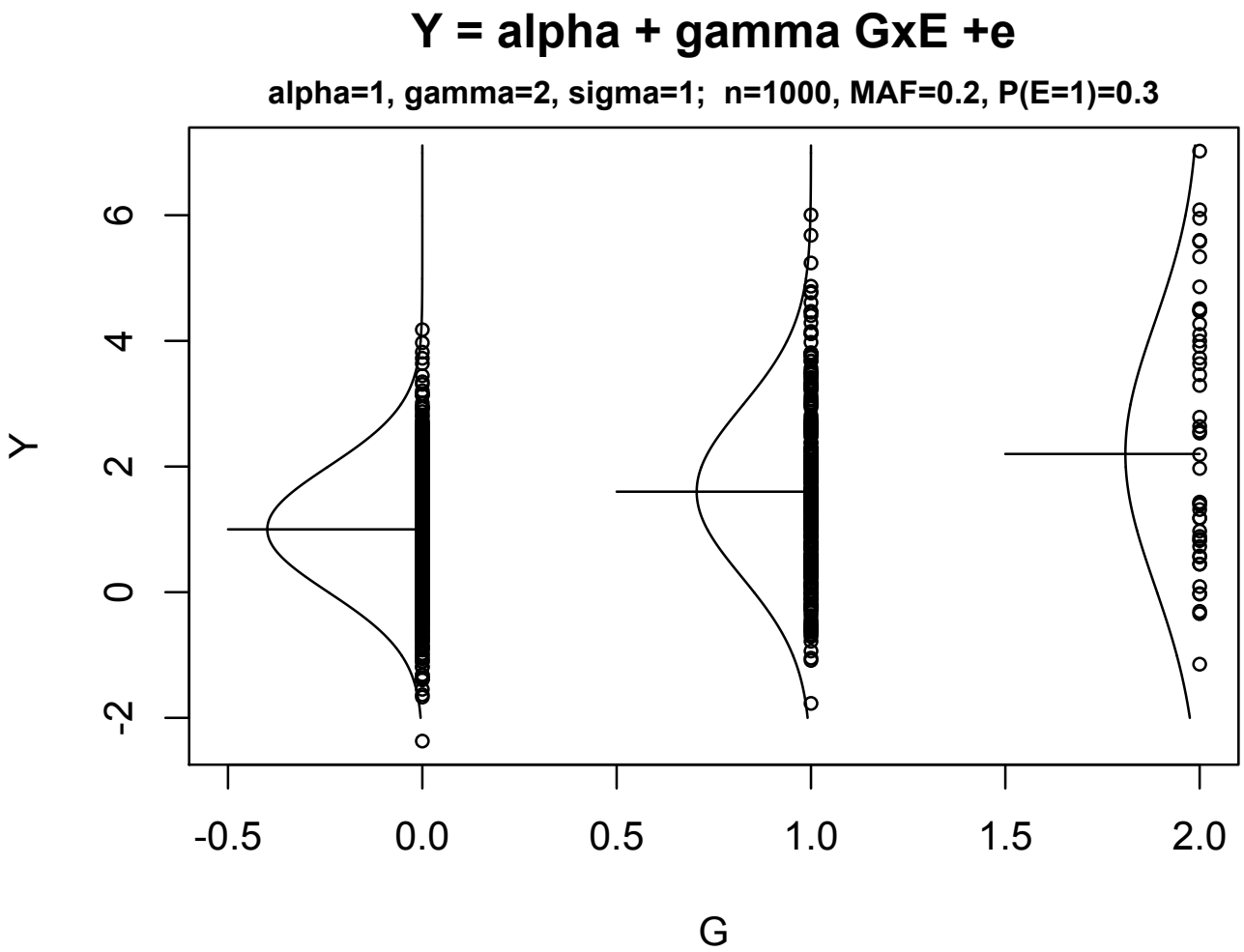
QuantitativeTraits - More Complex Models II

- Consider the simplest interaction model,

$$Y = \alpha + \gamma GxE + e.$$

  - $\alpha = 1, \gamma = 2, \sigma = 1$ ($e \sim N(0,1)$)
  - G is the same as before: allele frequency of D is $p = 0.2$, and genotype frequency of G follow HWE.
  - E is a binary covariate with $P(E = 1) = 0.3$.

QuantitativeTraits - More Complex Models III



QuantitativeTraits - More Complex Models IV

- For model: $Y = \alpha + \beta G + e$,

$$E(Y|G) = \alpha + \beta G, \;\; Var(Y|G) = \sigma^2.$$

- For model: $Y = \alpha + \gamma GxE + e$,

$$E(Y|G) = \alpha + \gamma GxE(E), \;\; Var(Y|G) = \gamma^2 G^2 Var(E) + \sigma^2.$$

- Thus, variance of Y in each genotype group G varies. This provides one example for Exercise 8 of Chapter 2.

- More advanced research question: How would one utilize the variance differences between genotypes to increase the power of association studies?

  A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways.

  A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty

Exercises

- Chapter 2 Exercise 1.
- Chapter 2 Exercise 5.
- Chapter 2 Exercise 6.
- Chapter 2 Exercise 7.