Module 3 - Population Genetics

(Fundamentals of) Statistical Genetics

Lei Sun

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

Outline

Chapter 3: Some Basic Concepts from Population Genetics

(If needed) Review of Some Basic Concepts from Statistical Inference

- Genotype and allele frequency estimation
- Direct counting
- Review and use of likelihood, score function, Fisher Information and MLE
- Hardy Weinberg Equilibrium: from genotype to allele frequency estimation
- Allele frequency hypothesis testing
- Review and use of normal approximation, likelihood ratio test, score test
- Frequency estimation complexities: missing data and related individuals
- Testing HWE using the Pearson χ^2 test.
- The Sickle Cell Anemia example
- Additional complexities and opportunities

Population Allele/Genotype Frequencies

- For **a given population**, different genotypes and alleles may appear with different frequencies.

- How to estimate genotype and allele frequencies?
  - Direct counting and examples
  - What are the **assumptions**?
  - What are the **alternative methods**?
  - Do we need to estimate both?

Frequencies Estimation - Direct Counting

- Also see Textbook Box 3.2 Example.

$$\hat{p}_{AA} = n_{AA}/n, \quad \hat{p}_{Aa} = n_{Aa}/n, \quad \hat{p}_{aa} = n_{aa}/n.$$

Genotypes	AA	Aa/aA	aa	Total
Observed Counts	189	89	9	287
Genotype frequency	.66	.31	.03	1.0

$$\hat{p}_A = (2n_{AA} + n_{Aa})/2n, \quad \hat{p}_a = (2n_{aa} + n_{Aa})/2n$$

Alleles	A	a	Total
Observed Counts	$189 \cdot 2 + 89 = 467$	$9 \cdot 2 + 89 = 107$	574
Allele frequency	.81	.19	1.0

- This can be easily extended for markers with more than 2 alleles (Textbook Equation (3.2)), e.g.

$$\hat{p}_A = (2n_{AA} + n_{AB} + n_{AC} + \ldots)/2n.$$

Frequencies Estimation - Important Stat Gene Questions

- Is this a 'good' estimator?
  - What is a good estimator? (**unbiasedness and minimal variance**)
  - How do we show that this estimator is unbiased (and has minimal variance)?
  - Need to define random variables and work with statistical models (binomial and multinomial distributions)

- What are the assumptions?
  - Are the two alleles of a genotype randomly paired (**in HWE**)?
  - What if some of the subjects/individuals are related to each other, say siblings (**correlated data**)?
  - What if the subjects/individuals are not randomly selected from the population of interest (**biased sampling**)?

- What are the alternative methods other than direct counting?
  - Why do we need alternative methods? (to handle more complex data such as correlated data and missing data)
  - Maximum Likelihood Estimates (**MLE**)

Review of Likelihood, Score Function, MLE, Fisher Information I

- θ: the parameter (vector - could be more than 1 parameter) of interest.
- X: the data with probability density distribution P_θ.

- **Likelihood function** (as a function of θ):

$$L(\theta) = P_\theta(X) = P(X; \theta).$$

  The kernel: the part of the likelihood function involving the parameters.

- Log-likelihood function:

$$l(\theta) = log(L(\theta)).$$

  Why work on the Log version of a likelihood?

Review of Likelihood, Score Function, MLE, Fisher Information II

- **Score function**: the first derivative of the log-likelihood function w.r.t. the parameter(s).

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta_i}, i = 1, \ldots, k.$$

- $\hat{\theta}$ is **MLE** if $l(\theta) = log(L(\theta)) = log(P(X;\theta))$ is maximized.
- Obtain MLE by solving the score equation:

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta_i} = 0, i = 1, \ldots, k.$$

- Is this condition sufficient to claim MLE?
- Note that MLE may not exist; may not be unique; may not easy to compute (different algorithms); may not be unbiased.
- Justification of the method of MLE: large sample properties, e.g. consistency, efficiency, asymptotic distribution.

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow N(0, I(\theta)^{-1}).$$

Review of Likelihood, Score Function, MLE, Fisher Information III

- **Fisher Information** $I(\theta)$:

$$I(\theta) = E_\theta(S(\theta)^2) = E_\theta\left((\frac{\partial log l(\theta)}{\partial \theta})^2 \right) = E_\theta\left(-\frac{\partial^2 log l(\theta)}{\partial \theta^2} \right)$$

(The amount of information about θ contained in data X.

- Remark 1: The calculation using $E_\theta(-\frac{\partial^2 log l(\theta)}{\partial \theta^2})$ is more attractive than using $E_\theta(\{\frac{\partial log l(\theta)}{\partial \theta}\}^2)$

- Remark 2: The more information about θ provided on average by data, the smaller we expect the variance of a "good" estimator to be; related to the possible lower bound for the variance of unbiased estimators.

Genotype Frequencies Estimation - MLE I

- What is our data and **random variable**?

$$X = (n_{AA}, n_{Aa}, n_{aa}).$$

- What is the distribution of our random variable X?
  - **Multinomial distribution**: each **independent** trial/individual has K possible outcomes/categories, and it belongs to the k_{th} category with probability p_k.
  - Let n_k be the number of total n individuals in category k.
  - The counts $(n_1, n_2, \ldots, n_K)$ have the multinomial distribution:

$$P(X = (n_1, n_2, \ldots, n_K)) = \binom{n}{n_1, n_2, \ldots, n_{K-1}} p_1^{n_1} p_2^{n_2}, \ldots, p_k^{n_k}$$

$$= \frac{n!}{n_1! n_2! \cdots n_K!} p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K}.$$

  - The marginal distribution of each n_k is $Bino(n, p_k)$, and

$$E(n_k) = np_k, \quad Var(n_k) = np_k(1 - p_k).$$

Genotype Frequencies Estimation - MLE II

- In our case (assuming independent samples):

$$P(X = (n_{AA}, n_{Aa}, n_{aa})) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} p_{AA}^{n_{AA}} p_{Aa}^{n_{Aa}} p_{aa}^{n_{aa}}$$

$$= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} p_{AA}^{n_{AA}} p_{Aa}^{n_{Aa}} (1 - p_{AA} - p_{Aa})^{n_{aa}}.$$

- The **two parameters** of interest are

$$p_{AA} = P(AA \text{ genotype}), \quad p_{Aa} = P(Aa \text{ genotype}).$$

- Note that we only have 2 unknown parameters:

$$p_{aa} = P(aa \text{ genotype}) = 1 - p_{AA} - p_{Aa}.$$

Genotype Frequencies Estimation - MLE III

- The likelihood for parameters $\theta = (p_{AA}, p_{Aa})$:

$$L(\theta) = P(X = (n_{AA}, n_{Aa}, n_{aa}); \theta) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} p_{AA}^{n_{AA}} p_{Aa}^{n_{Aa}} (1 - p_{AA} - p_{Aa})^{n_{aa}}.$$

- The log-likelihood:

$$l(\theta) = log(L(\theta))$$

$$= log(\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}) + n_{AA} log(p_{AA}) + n_{Aa} log(p_{Aa}) + n_{aa} log(1 - p_{AA} - p_{Aa}).$$

- The score function: take the (partial) first derivatives of the log-likelihood function with respect to the parameters of the interest:

$$S(\theta) = \frac{\partial l(\theta)}{\partial p_{AA}} = \frac{n_{AA}}{p_{AA}} - \frac{n_{aa}}{1 - p_{AA} - p_{Aa}},$$

$$S(\theta) = \frac{\partial l(\theta)}{\partial p_{Aa}} = \frac{n_{Aa}}{p_{Aa}} - \frac{n_{aa}}{1 - p_{AA} - p_{Aa}}.$$

Genotype Frequencies Estimation - MLE IV

- The MLE: $S(\theta) = 0 \Longrightarrow$

$$\hat{p}_{AA} = \frac{n_{AA}}{n},$$
$$\hat{p}_{Aa} = \frac{n_{Aa}}{n}.$$

- So in this case the MLE approach gives us the exact same estimator as the direct counting method. However, direct counting may not be possible in other cases; more on this later.
- Note that if we plug in a particular observed value for n_{AA}, e.g.$n_{AA} = 189$,

$$\hat{p}_{AA} = \frac{n_{AA}}{n} = \frac{189}{287} = 0.66 \text{ is a MLE point } \textbf{estimate}$$

- Otherwise,

$$\hat{p}_{AA} = \frac{n_{AA}}{n} \text{ is a MLE } \textbf{estimator}\text{, which is a random variable}$$

Genotype Frequencies Estimation - MLE V

- Now we can show that $\hat{p}_{AA} = \frac{n_{AA}}{n}$ and $\hat{p}_{Aa} = \frac{n_{Aa}}{n}$ are unbiased estimator for p_{AA} and p_{Aa}.
- We have assumed that the observed data $X = (n_{AA}, n_{Aa}, n_{aa})$ come from a multinomial distribution with parameters p_{AA} and p_{Aa}. Although we do not know the underlying true parameter value, we know $E(n_{AA}) = np_{AA}$ and $Var(n_{AA}) = np_{AA}(1 - p_{AA})$, therefore,

  $$E(\hat{p}_{AA}) = E(\frac{n_{AA}}{n}) = \frac{E(n_{AA})}{n} = \frac{np_{AA}}{n} = p_{AA}.$$

  And has variance

  $$Var(\hat{p}_{AA}) = Var(\frac{n_{AA}}{n}) = \frac{Var(n_{AA})}{n^2} = \frac{np_{AA}(1 - p_{AA})}{n^2} = \frac{p_{AA}(1 - p_{AA})}{n}.$$

- Is this the minimal variance that we can achieve?

From Genotype to Allele Frequencies

How do we estimate the allele frequencies using the likelihood framework? Do we need to estimate both genotype and allele frequencies?

- A microsatellite marker with 15 alleles has
  - $15 + \binom{15}{2} = 15 + 105 = 120$ genotypes.
  - May not have enough data; loss of statistical efficiency.

- A SNP with alleles 1 and 2, and observe

Genotypes	AA	Aa/aA	aa	Total
Observed Counts	239	11	0	250

  - Direct estimate of frequency of genotype aa: $0/250 = 0$.
  - However, such estimate may not be accurate; allele a does appear in the data.

- Use allele frequency to estimate genotype frequency.
  - 14 parameters (15 alleles) are much easier to estimate than 119 parameters (120 unordered genotypes).
  - HWE permits the calculation of genotype frequency from allele frequency.

Hardy Weinberg Equilibrium (HWE)

- Independent discovery of Godfrey Hardy (English mathematician) and Wihelm Weinberg (German physician) in 1908.

- A population is in HWE, if two alleles in a genotype are **independent** draws from the same distribution, or

  if genotype frequency in the population depends only on the allele frequency.

- E.g. for a biallelic marker (also see Textbook Equation (3.3)),

$$freq(AA) = freq(A)^2, \quad freq(Aa) = 2freq(A)freq(a), \quad freq(aa) = freq(a)^2.$$

$$p_{AA} = p_A^2, \quad p_{Aa} = 2p_A p_a = 2p_A(1 - p_A), \quad p_{aa} = p_a^2 = (1 - p_A)^2.$$

- To simply the statistical theory, many statistical genetics methods rely on the assumption of HWE.

Application of HWE - Allele Frequency Estimation I

Provide some theoretical justification for the direct counting method for the allele frequency estimation (Also see Textbook Box 3.3)

$$\hat{p} = p_A = (2n_{AA} + n_{Aa})/2n, \quad p_a = 1 - p_A = (2n_{aa} + n_{Aa})/2n$$

- Let p be the population allele frequency of A.
- What is the likelihood under the null hypothesis of HWE? (What are the assumptions needed for iid?)

$$L(\theta) = P(n_{AA}, n_{Aa}, n_{aa}) \sim (p^2)^{n_{AA}}(2p(1-p))^{n_{Aa}}((1-p)^2)^{n_{aa}}$$

Application of HWE - Allele Frequency Estimation II

- Alternatively, following the Textbook define a random variable X_i: the number of copies of A for individual i.
- What is the distribution of X_i under HWE?
  - $p(X_i = 2) = p(X_i = AA) = p^2$
  - $p(X_i = 1) = p(X_i = Aa) = 2p(1-p)$
  - $p(X_i = 0) = p(X_i = aa) = (1-p)^2$

$$X_i \sim Bino(2, p), \quad \forall i.$$

- Define a new random variable, $X = \sum_{i=1}^{n} X_i$.
- Because X_is are iid $Bino(2, p)$ then

$$X \sim Bino(2n, p).$$

- Now we can use statistical theory to write down the likelihood, obtain MLE, prove unbiasedness and provide standard error, and perform hypothesis test.

Allele Frequency Estimation

Related to Binomial likelihood and inference (estimation)

- Let $X \sim Bino(2n, p)$, then we know that if we observed $X = x = 2n_{AA} + n_{Aa}$,

$$P(X = x) = \binom{2n}{x} p^x (1-p)^{2n-x},$$

- We also know that

$$E(X) = 2np, \quad Var(X) = 2np(1-p).$$

- The direct counting leads to the following estimate

$$\hat{p} = \frac{x}{2n} = \frac{2n_{AA} + n_{Aa}}{2n}.$$

- We can actually show that $\hat{p} = \frac{x}{2n}$ is MLE.

Allele Frequency Estimation - MLE I

- The likelihood (for parameter $\theta = p$):

$$L(p) = P(X = x; p) = \binom{2n}{x} p^x (1-p)^{2n-x}.$$

- The log-likelihood:

$$l(p) = log(\binom{2n}{x}) + xlog(p) + (2n-x)log(1-p).$$

- The score function:

$$S(p) = \frac{\partial l(p)}{\partial p} = \frac{x}{p} - \frac{2n-x}{1-p} = \frac{x-2np}{p(1-p)}.$$

- The MLE:

$$S(p) = 0 \Longrightarrow \hat{p} = \frac{x}{2n}.$$

Allele Frequency Estimation - MLE II

- Regardless, we can show that $\hat{p} = \frac{X}{2n}$ is an unbiased estimator ($\hat{p} = \frac{x}{2n}$ is an estimate assuming the observed value of $X = x$.)

$$E(\hat{p}) = E(\frac{X}{2n}) = \frac{E(X)}{2n} = \frac{2np}{2n} = p.$$

And has variance

$$Var(\hat{p}) = Var(\frac{X}{2n}) = \frac{Var(X)}{(2n)^2} = \frac{2np(1-p)}{(2n)^2} = \frac{p(1-p)}{2n}.$$

- To show that this is a minimal variance unbiased estimator (MVUE) estimator, we need more math stat knowledge.
- What is the Fisher's information in this case?

$$I(p) = E\left(-\frac{\partial^2 l(p)}{\partial p^2}\right) = E\left(\frac{X}{p^2} + \frac{2n-X}{(1-p)^2}\right) = \frac{2n}{p} + \frac{2n}{1-p} = \frac{2n}{p(1-p)}.$$

Allele Frequency Testing

Related to Binomial likelihood and inference (hypothesis testing)

- There are several ways to test the proportion
  - **Normal approximation**
  - **Likelihood ratio test**
  - **Score test**
  - **Wald test**
  - Exact

- The use of these various tests is an overkill for a simple problem like this, but it helps us grasp the essence and build our inference skill needed for more complex problems. We will combine our intuition with some math stat.

Allele Frequency Testing - Normal Approximation I

- We are interested the null hypothesis: $H_0 : p = p_0$.
- Our estimator is a random variable: $\hat{p} = \frac{X}{2n}$.

- For a given data (e.g. $2n = 574, X = x = 467$), how do we determine if the specific estimate $\hat{p} = \frac{x}{2n} = \frac{467}{574} = 0.8136$ is statistically different from say $p_0 = 0.75$?
- Is $0.8136 - 0.75 = 0.0636$ a 'big difference'?
- This depends on the **distribution** (or the background) of our estimator.

- For simplicity we assume that we will do a one-sided test.
- We want to calculate the probability of obtaining estimates that are as extreme as or even more extreme than what we have now:

$$P(\frac{X}{2n} \geq 0.8136).$$

Allele Frequency Testing - Normal Approximation II

- If we assume 'large sample', then under then null H_0,

$$\hat{p} = \frac{X}{2n} \approx N(p_0, p_0(1-p_0)/2n).$$

- Then

$$P(\hat{p} \geq 0.8136) = P(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/2n}} \geq \frac{0.8136 - p_0}{\sqrt{p_0(1-p_0)/2n}})$$

$$= P(Z \geq \frac{0.8136 - p_0}{\sqrt{p_0(1-p_0)/2n}}).$$

- Note that $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/2n}} = \frac{\sqrt{2n}(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)}}$ is the Z value in the Textbook, and the above the probability is the (one-sided) p-value.

$$P(Z \geq \frac{0.8136 - p_0}{\sqrt{p_0(1-p_0)/2n}}) = P(Z > 3.52) = 0.0002.$$

Allele Frequency Testing - Normal Approximation III

- What's the other side?

$$P(\hat{p} \leq p_0 - (0.8136 - p_0)) = P(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/2n}} \leq \frac{-(0.8136 - p_0)}{\sqrt{p_0(1-p_0)/2n}})$$

$$= P(Z \leq \frac{-(0.8136 - p_0)}{\sqrt{p_0(1-p_0)/2n}}) = P(Z \geq \frac{0.8136 - p_0}{\sqrt{p_0(1-p_0)/2n}}).$$

- One can formally show that $\frac{(\hat{p}-p_0)}{\sqrt{p_0(1-p_0)/2n}}$ is a **Score test**.

- **Wald test** will take the form $\frac{(\hat{p}-p_0)}{\sqrt{\hat{p}(1-\hat{p})/2n}}$.

- **Likelihood test** yet has another form.

Allele Frequency Testing - Likelihood Ratio Test I

- Wish to test a null hypothesis H_0 which is specified by imposing r restrictions on $\theta = (\theta_1, ..., \theta_k)$, say $h_j(\theta) = 0$, $j = 1, ..., r$.
- $\hat{\theta}$: unrestricted MLE, and
  $\tilde{\theta}$: restricted MLE under the null.
- Subject to regularities and with 'large sample'

$$T = 2log \frac{L(\hat{\theta})}{L(\tilde{\theta})} = 2(l(\hat{\theta}) - l(\tilde{\theta})) \sim \chi^2_r.$$

- Under the simplest case of one parameter θ and $H_0 : \theta = \theta_0$, then

$$T = 2log \frac{L(\hat{\theta})}{L(\theta_0)} = 2(l(\hat{\theta}) - l(\theta_0)) \sim \chi^2_1.$$

Allele Frequency Testing - Likelihood Ratio Test II

- Back to the binomial example for the allele frequency testing, $\theta = p$.
- The (kernel) Log-likelihood function:

$$l(p) \sim xlog(p) + (2n - x)log(1 - p).$$

- The null hypothesis: $H_0 : p = p_0$.
- Restricted MLE under the null: $\tilde{p} = p_0 = 0.75$.
- Unrestricted MLE: $\hat{p} = \frac{x}{2n}$.

Allele Frequency Testing - Likelihood Ratio Test III

- The Log-likelihood ratio test statistic: if we treat x as a random variable whose value is generated from the underlying $Bino(2n, p)$, then

$$T = 2(l(\hat{p}) - l(\tilde{p}))$$
$$= 2(x log \frac{x}{2np_0} + (2n - x) log \frac{2n - x}{2n(1 - p_0)}) \sim \chi_1^2.$$

- The observed test statistic, that is if we plug in the data that we actually observed for x, $x = 467$, to the above expression, then we have

$$T_{obs} = 2(x log \frac{x}{2np_0} + (2n - x) log \frac{2n - x}{2n(1 - p_0)}) = 13.2.$$

- Obtain p-value by using the $T \sim \chi_1^2$ distribution:

$$P(T \geq T_{obs}) = P(T \geq 13.2) = 0.00028.$$

Allele Frequency Testing - Likelihood Ratio Test IV

- Note that the above the result can be expressed as

$$T_{obs} = 2 \sum \text{observed} \times log \frac{\text{observed}}{\text{expected}}.$$

(This formula also holds for tests about multinomial and Poisson parameters.)

Allele Frequency Testing - Score Test I

- $H_0 : \theta = \theta_0$.
- $S(\theta)$: the score function.
- Score test of H_0 has test statistic:

$$T_{Score} = S(\theta_0)' I(\theta_0)^{-1} S(\theta_0).$$

- Asymptotically χ^2 distributed with d.f. same as the likelihood ratio test.
- For single parameter:

$$T_{Score} = \frac{S(\theta_0)^2}{I(\theta_0)}$$

- Rationale: $S(\hat{\theta}) = 0$, so if $\hat{\theta}$ is far from θ_0, then $S(\theta_0) >> 0$.
- We also want to standardize it by its variance, and we note that $E(S(\theta)) = 0$, and $Var(S(\theta)) = E(S(\theta)^2) - (E(S(\theta)))^2 = E(S(\theta)^2) = I(\theta)$.
- Can you prove and **understand** the above properties using Binomial distribution?

Allele Frequency Testing - Score Test II

- Back to the binomial example, $\theta = p$.

$$S(p) = \frac{x - 2np}{p(1-p)}.$$

$$I(p) = \frac{2n}{p} + \frac{2n}{1-p} = \frac{2n}{p(1-p)}.$$

$$T_{Score} = \frac{S(p_0)^2}{I(p_0)} = \left(\frac{\frac{x}{2n} - p_0}{\sqrt{p_0(1-p_0)/2n}} \right)^2 = \left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/2n}} \right)^2 = Z^2$$

- Note the importance of assess statistical significance of test statistics in the context of the **underlying distribution**!
  - To measure the ‘distance’ between H_0 and the observed data
  - Use the normal approximation, the test statistic is $Z_{obs} = 3.52$.
  - Use the score test, the test statistic is $T_{obs} = 12.38$.
  - However, we have the same statistical evidence as measured by p-value: $P(T \geq T_{obs}) = 2P(Z \geq Z_{obs}) = 0.0004$.

Allele Frequency Testing - Wald Test I

- $H_0 : \theta = \theta_0$.
- $\hat{\theta}$: the MLE.
- Wald test of H_0 has test statistic:

$$T_{Wald} = (\hat{\theta} - \theta_0)'[Cov(\hat{\theta})]^{-1}(\hat{\theta} - \theta_0).$$

- Asymptotically χ^2 distributed with d.f. equal the rank of $Cov(\hat{\theta})$.
- $Cov(\hat{\theta})$ is often approximately by the inverse of the **observed Fisher information** matrix:

$$I(\theta)_{ij} = E[-\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}].$$

- For a single parameter:

$$T_{Wald} = \frac{(\hat{\theta} - \theta_0)^2}{1/I(\hat{\theta})}.$$

Allele Frequency Testing - Wald Test II

- Rationale is quite intuitive: comparing MLE of the parameter value supported by the data with the null hypothesized value, and standardized by the (estimated asymptotic) variance of the MLE.

- Back to the binomial example, $\theta = p$.

$$\hat{p} = \frac{x}{2n}.$$

$$I(p) = \frac{2n}{p} + \frac{2n}{1-p} = \frac{2n}{p(1-p)}.$$

$$T = \frac{(\hat{p}-p_0)^2}{1/I(\hat{p})} = \frac{(\hat{p}-p_0)^2}{\hat{p}(1-\hat{p})/2n} = \left(\frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})/2n}}\right)^2$$

- Comparing with the Score test and the LRT?

Frequency Estimation Complexities - Missing Data I

The ABO-blood type example

- The ABO-gene or ABO-locus is on chromosome 9
- It has 3 alleles (antigens) (A, B, O)
- And it determines 4 blood type (A, B, AB, O):

genotype	phenotype
AA AO	A
BB BO	B
AB	AB
OO	O

A, B are dominant to O.

O is recessive to A, B.

A, B are co-dominant.

- 6 possible genotypes, but only 4 phenotypically distinguishable.

Frequency Estimation Complexities - Missing Data II

- E.g. in a large random sample obtained from Berlin (Bernstein 1925, Sham's book page 44):
  - $n_A = 9123$ blood type A
  - $n_B = 2987$ blood type B
  - $n_{AB} = 1269$ blood type AB
  - $n_O = 7725$ blood type O

- How to estimate the allele frequencies of alleles A, B and O
  (or even just to estimate the genotype frequencies of genotypes AA, AO, etc)?

- Problem with the direct counting method: missing data w.r.t. the count of each of the 6 genotypes: $n_{AA}, n_{AO}, n_{BB}, n_{BB}, n_{AB}, n_{OO}$!
  - $n_A = 9123 = n_{AA} + n_{AO}$: Among 9123 blood type A individuals, some have genotype AA and the others have genotype AO.
  - $n_B = 2987 = n_{BB} + n_{BO}$: Among 2987 blood type B individuals, some have genotype BB and the others have genotype BO.

Frequency Estimation Complexities - Missing Data III

- So we need a formal statistical approach, because there are no closed formed solutions for the score function, requiring learning of
  - The likelihood approach and
  - Related numerical algorithms (essentially gradient-decent algorithms).
  - Separate lecture notes: missing data and the Newton-Raphson and EM algorithms.

Frequency Estimation Complexities - Related Individuals

- Boehnke (1991). Allele frequency estimation from data on relatives. *American Journal of Human Genetics* 48:22-25.
- Broman (2001). Estimation of allele frequencies with data on sibships. *Genetic Epidemiology* 20:307-315.
- McPeek et al. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* 60(2):359-367.
- Lin Zhang, Sun (working paper). On 'Reverse' Regression for Robust Genetic Association Studies and Allele Frequency Estimation with Related Individuals.

Additional Important **Statistical Considerations**

- Most of the genetic association and linkage methods assume that allele frequencies are precisely known.
- Some of methods may not be robust to the misspecification of allele frequency - what is the effect on type I error and power?
- Some general stat research questions: How to improve the robustness of methods? How to incorporate the uncertainty of allele frequency estimation directly into down-stream methods?

Testing HWE - Why

- HWE provides simplification in our inference, however, several assumptions needed for a population to achieve HWE:
  - infinite population size, discrete generations,
  - random mating,
  - no selection, no migration, no mutation, and
  - equal initial genotype frequencies in females and males.
- The assumptions are never really satisfied in practice, but HWE may still be a good approximation.
- But, not all statistical methods are **robust** to the HWE assumption; it is important to test if the assumption HWE holds in practice.

Population Substructure and Other Causes of HWE Failure

- *We use the term population substructure loosely to refer to features of a population which result in variation of expected allele frequencies across individuals in a population.*
  - Population stratification: different populations (e.g. European, Asian, African) may have systematically different allele frequency distributions or features.
  - Population admixture: mixing of two more more populations due to migration.
  - Population inbreeding: mating among relatives.
- See Textbook Chapters 3.2 for some examples.

- See Textbook Chapters 3.3.2 for a general discussion of how population substructure can cause HWE failure.
- We defer more detailed discussions on population substructure until later when we discuss association analysis.

Testing HWE

Main Idea (also see Textbook Box 3.4 and Table 3.3)

- Compare the **Observed** genotype counts/frequencies with the **Expected** values calculated under the assumption of HWE.

- Use Pearson χ^2 test with test statistic:

$$T = \sum_{i_{th} \text{ category}} \frac{(O_i - E_i)^2}{E_i}.$$

- T follows the χ^2 distribution with the d.f. $=$ the number of categories $-$ the number of parameters used to obtain the expected counts E_is.
- Note that E_is are calculated/estimated under the null hypothesis H_0.

Testing HWE - Sickle Cell Anemia Example I

- *Sickle cell anemia is a Mendelian disorder that affects red blood cells and is associated with severe morbidity, including pain, hemolytic anemia and infections; without proper medical management, the death rate is high.*
- Caused by the hemoglobin, beta (HBB) gene on chromosome 11 (D mutation allele, and d normal allele)
- Highest frequency in Africa and descendants.
- Genotype data from infants and adults at the HBB gene in Tanzania (data from Allison 1956, courtesy of Dr. Andrew Paterson).

Testing HWE - Sickle Cell Anemia Example II

- Infants

Genotypes	dd	dD/Dd	DD	Total
Observed counts	189	89	9	287
Expected counts	188.3	88.3	10.4	287
Expected freq.	$0.81 \cdot 0.81$	$2 \cdot 0.81 \cdot 0.19$	$0.19 \cdot 0.19$	1.0

Alleles	d	D	Total
Observed counts	$189 \cdot 2 + 89 = 467$	$9 \cdot 2 + 89 = 107$	574
Estimated freq.	0.81	0.19	1.0

$T = 0.19$ with 1 df (why not 2?), p-value ≈ 0.66

Testing HWE - Sickle Cell Anemia Example III

- Adults

Genotypes	dd	dD/Dd	DD	Total
Observed counts	400	249	5	654
Expected counts	418.5	209.3	26.2	654
Expected freq.	$0.8 \cdot 0.8$	$2 \cdot 0.8 \cdot 0.2$	$0.2 \cdot 0.2$	1.0

Alleles	d	D	Total
Observed counts	$400 \cdot 2+249=1049$	$5 \cdot 2+249 =259$	1308
Estimated freq.	0.8	0.2	1.0

$T = 25.5$, 1 d.f. (why not 2?), p-value ≈ 0

Testing HWE - Sickle Cell Anemia Example IV

- Notes on the degree of freedom (d.f.).
  - 3 genotypes, 2 d.f. (assuming total is fixed).
  - 1 d.f. was first used to estimate the allele frequency.
  - 1 d.f. is left to test HWE.

  - If allele frequency was available *externally*, then 2 d.f. for test of HWE.
  - The internal allele frequency estimated from the current data tends to fit the current data better (in terms the 'distance' $O - E$ for each genotype count), as compared to using the external allele frequency estimated from other data. So we expect less variation (fewer d.f.) for the former and more variation (more d.f.) for the latter.
  - This is a form of selective inference or over-fitting!

  - In general, more d.f. implies more variation, thus one needs more evidence (here a bigger T) to reject the null hypothesis.

- Any alternative tests that we can use to test the HWE assumption?

Testing HWE - Sickle Cell Anemia Example V

- Why HWE was rejected in the adult population but not in the infant population?
  - The HBB gene has two alleles: D mutation allele, and d normal allele.
  - Sickle cell anemia is a recessive disease caused by the HBB gene: individuals with genotype DD are affected.
  - Some of the affected infants may not survive to adulthood, and such selection results in a departure of the genotype frequency among adults from the HWE expectation, particularly for the DD group.

Possible Complexities and Opportunities I

- **The open research questions**
  - Can you do better in each case?
  - Any other cases that have not been considered?
  - Unified approach that consider all cases jointly?

  - Lin Zhang, Sun (working paper). A generalized robust allele-based genetic association test.

Exercises

- Chapter 3 Exercise 1.
- Chapter 3 Exercise 2.
- Chapter 3 Exercise 3.
- Chapter 3 Exercise 5.
- Chapter 3 Exercise 6.
- Chapter 3 Exercise 7.
- Chapter 3 Exercise 9.

What's Next

- A closer look at the ABO-blood example: missing data and the Newton-Raphson and EM algorithms.
- Chapter 2 - Principles of Inheritance: Mendel's Laws and Genetic Models
- Likelihood for Pedigree Data, expanding Chapter 4.1 Preliminaries.