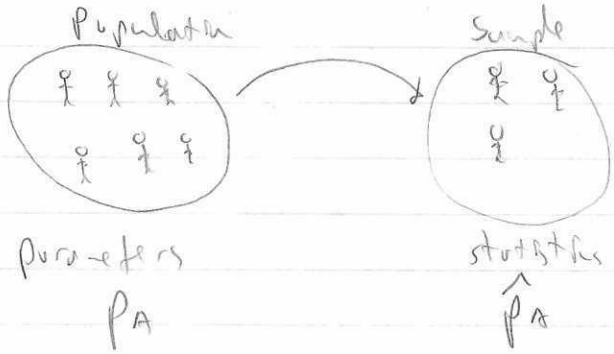
GDA §2 Notes



"estimator": A statistic that estimates a parameter
"estimate": specific numerical value of an estimator

- Goal of chapter:
  1. Framework for creating estimators
  2. Quantifying uncertainty in estimators (confidence intervals, hypothesis tests)

- "Consistent" $\hat{\theta} \xrightarrow[n\to\infty]{} \theta$
- "unbiased": $E[\hat{\theta}] = \theta$
- Estimators have "sampling distributions" so would be different if given a different sample of n individuals.
- Multinomial distribution:

1.) Each member of a population falls into one of k categories
Eg: aa aA AA

2.) $P_r(\text{category } i) = Q_i$

Eg: $P_r(aa) = Q_1$, $P_r(aA) = Q_2$, $P_r(AA) = Q_3$

3.) We sample n individuals which are independent.

If n_i = # individuals observed in sample from category i, then

$$P_r(n_1, n_2, \ldots, n_k) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} Q_i^{n_i}$$

So for $k=3$,

$$\frac{n!}{n_1!\, n_2!\, n_3!} Q_1^{n_1} Q_2^{n_2} Q_3^{n_3}$$

- Derivation:

If we sample in a sequence and see category sequence

$(1, 1, 2, 1, 3, 3, 2, 1, 3, 1, 1)$

Then Probability of this sequence is

$$Q_1 Q_1 Q_2 Q_1 Q_3 Q_3 Q_2 Q_1 Q_3 Q_1 Q_1$$
$$= Q_1^6 Q_2^2 Q_3^3$$

But the order does not matter! we want $P_r(n_1 = 6, n_2 = 2, n_3 = 3)$, not P_r(sequence)

sequences of size n w/ category sizes n_1, n_2, n_3 :

$$\frac{n!}{n_1!\, n_2!\, n_3!}$$

$n!$ ← # ways to order n individuals

Top 1 2 3 4 5 6

$Q_1\ Q_1\ Q_2\ Q_1$ ←

1 2 1 3

$n_1!\, n_2!\, n_3!$ ← # ways to order the n_i values

for $k=2$, This is the Binomial distribution

$$P_r(a, n-a) = \frac{n!}{a!(n-a)!} Q^a (1-Q)^{n-a}$$

- Property of multinomial

If $n_1, n_2, \ldots, n_k \sim \text{Multinomial}(Q_1, Q_2, \ldots, Q_k)$

and $n_{12} = n_1 + n_2$, then

$$n_{12}, n_3, \cdots, n_k \sim \text{Multinomial}(Q_1 + Q_2, Q_3, \ldots, Q_k)$$

- Eg $n_1, n_2 + n_3 + \cdots n_k \sim \text{Bin}(Q_1)$

I So each count is marginally binomially distributed

- $E[n_i] = \sum_{r=0}^{n} r \, P_r(n_i = r)$

$$= \sum_{r=0}^{n} r \frac{n!}{r!(n-r)!} Q_i^r (1-Q_i)^{n-r}$$

$$= n Q_i$$

(algebra) ← sample size × prob in category i

$$\text{Var}(n_i) = n Q_i (1-Q_i)$$

- So $E\left(\frac{n_1}{n}\right) = \frac{1}{n} E[n_1] = \frac{1}{n} n q_1 = q_1$

$\Rightarrow \hat{q}_1 = \frac{n_1}{n}$ is unbiased

- $\text{Var}\left(\frac{n_1}{n}\right) = \frac{1}{n^2} \text{Var}(n_1) = \frac{1}{n^2} n q_1 (1-q_1) = \frac{q_1(1-q_1)}{n}$

$\Rightarrow$ variance of $\hat{q}_1$ decreases as n gets larger (more accurately estimate q_1)

- Note: for any random variable X,

$E[X^2] = E(X)^2 + \text{Var}(X)$

So $E(\hat{q}_1^2) = q_1^2 + \frac{1}{n} q_1(1-q_1)$

- Multinomial Covariance

$E(n_i n_j) \overset{\text{math}}{=} n(n-1) q_i q_j$

$\text{Cov}(n_i, n_j) = E\left((n_i - E(n_i))(n_j - E(n_j))\right)$

$= E(n_i n_j) - E[n_i] E[n_j]$

$= -n q_i q_j$

$\Rightarrow \text{Cov}(\hat{q}_i, \hat{q}_j) = \text{Cov}\left(\frac{1}{n} n_i, \frac{1}{n} n_j\right) = \frac{1}{n^2} \text{Cov}(n_i, n_j) = -\frac{q_i q_j}{n}$

$$\boxed{\text{Cov}(\hat{q}_i, \hat{q}_j) = -\frac{q_i q_j}{n}}$$

← negative b/c "more i means less opportunity for j"

$$Cor(n_i, n_j) = \frac{Cov(n_i, n_j)}{\sqrt{Var(n_i)\, Var(n_j)}} = \frac{Q_i Q_j}{\sqrt{Q_i(1-Q_i)\, Q_j(1-Q_j)}} = Cor(\hat{Q}_i, \hat{Q}_j)$$

- If alleles $A_1, A_2, \ldots, A_k$, then

$$n_u = 2n_{uu} + \sum_{v \neq u} n_{uv}$$

$2n_{uu}$: double count, homozygotes; $\sum_{v \neq u} n_{uv}$: single count, heterozygotes

E.g. for 3 alleles

$$n_1 = 2n_{11} + n_{12} + n_{13}$$
$$n_2 = 2n_{22} + n_{12} + n_{23}$$
$$n_3 = 2n_{33} + n_{13} + n_{23}$$

- $E[n_u] = E\left[2n_{uu} + \sum_{v \neq u} n_{uv}\right]$

$$= 2E[n_{uu}] + \sum_{v \neq u} E[n_{uv}]$$

$$= 2P_{uu} + \sum_{v \neq u} P_{uv}$$ ← genotype frequencies

$$= 2np_u$$ ← allele frequency

$$p_u = P_{uu} + \frac{1}{2}\sum_{v \neq u} P_{uv}$$

$$\hat{p}_u = \frac{2n_{uu} + \sum_{v \neq u} n_{uv}}{2n}$$ ← want var and variance of this

↳ can't get it directly from multinomial

$$\hat{p}_u = \frac{n_u}{2n}$$

so $E[\hat{p}_u] = p_u$

- $\text{Var}(n_u) \overset{\text{approx}}{=} 2n\left(p_u + \underline{p}_{uu} - 2p_u^2\right)$
- $\text{Var}(\hat{p}_u) = \frac{1}{2n}\left(p_u + \underline{p}_{uu} - 2p_u^2\right)$

↳ can estimate this using $\hat{p}_u$ and $\hat{\underline{p}}_{uu}$

- 95% CI: $\hat{p}_u \pm 2\sqrt{\frac{1}{2n}\left(\hat{p}_u + \hat{\underline{p}}_{uu} - 2\hat{p}_u^2\right)}$

$$= \hat{p}_u \pm 2\,\widehat{\text{var}}(\hat{p}_u)$$

- This was complicated b/c allele numbers are not multinomial generally
- When they are multinomial, this is called Hardy-Weinberg equilibrium and is because

$$\underline{p}_{uu} = p_u^2$$

$$\underline{p}_{uv} = 2p_u p_v \quad \text{for } u \neq v$$

- Indicator variable: 1 if in category and 0 o.w.

Eg. $X_{ij} = \begin{cases} 1 & \text{if allele } j \text{ in individual } i \text{ is } A \\ 0 & \text{o.w.} \end{cases}$

Then

$$\hat{p}_A = \frac{1}{2n} \sum_{i=1}^{\hat{n}} \sum_{j=1}^{2} X_{ij}$$

$$= n_A$$

- For all indicator variables: $E[X_{ij}] = P(X_{ij} = 1)$

$$\text{Var}(X_{ij}) = P(X_{ij}=1)(1 - P(X_{ij}=1))$$

- Useful sometimes b/c multinomial counts are always sums of indicator variables

$$E[\hat{p}_A] = \frac{1}{2n} \sum_{i=1}^{\hat{n}} \sum_{j=1}^{2} E[X_{ij}] = \frac{1}{2n} \sum_{i=1}^{\hat{n}} \sum_{j=1}^{2} p_A = p_A$$

much simpler than

$$E[n_{AA}] = \sum_{r=0}^{\hat{n}} r \frac{n!}{r!(n-r)!} p_{AA}^{r} (1 - p_{AA})^{n-r}$$

$E[n_{Aa}]$ = similar

$n_A = 2n_{AA} + n_{Aa}$

$\hat{p} = \frac{1}{2n} E[n_A]$

- Skip total Variance of allele frequencies
- the δ-method:

Suppose $\quad \mathrm{var}\left[\sqrt{n}(\theta_0 - \hat{\theta}_n)\right] \xrightarrow[n\to\infty]{} \sigma^2$

Let $g()$ be a non-zero twice-diff. s.t. g' exists

then $\quad \mathrm{var}\left(\sqrt{n}(g(\theta) - g(\hat{\theta}_n))\right) \xrightarrow[n\to\infty]{} \sigma^2 [g'(\theta)]^2$

since $g(x) = g(\theta) + g'(\theta)[x-\theta] + \text{remainder}$] taylor series

$\Rightarrow \quad g(\hat{\theta}) = g(\theta) + g'(\theta)(\hat{\theta} - \theta)$

$$\approx \mathrm{var}\left(\sqrt{n}\left(g(\theta) - g(\theta) - g'(\theta)(\hat{\theta} - \theta)\right)\right)$$

$$= \mathrm{var}\left(\sqrt{n}\, g'(\theta)(\hat{\theta} - \theta)\right)$$

$$= [g'(\theta)]^2 \mathrm{var}\left(\sqrt{n}(\hat{\theta} - \theta)\right) \quad //$$

Multivariate analogue — now a vector in p dimensions

Suppose $\mathrm{cov}\left(\sqrt{n}\left(\theta - \hat{\theta}_n\right)\right) \xrightarrow[n\to\infty]{} \Sigma$ ($p \times p$)

$g(\theta): \mathbb{R}^p \to \mathbb{R}$

Then $\mathrm{cov}\left(\sqrt{n}\left(g(\theta) - g'(\theta)\right)\right) \xrightarrow[n\to\infty]{} \left[\nabla g(\theta)^T\right] \Sigma \left[\nabla g(\theta)\right]$

$$\nabla g(\theta) = \begin{pmatrix} \frac{dg}{d\theta_1} \\ \frac{dg}{d\theta_2} \\ \vdots \\ \frac{dg}{d\theta_n} \end{pmatrix} = \text{gradient of } g$$

- the formula in the book uses

$$\theta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} \quad \text{and} \quad g(\theta) = T(\mu)$$

and they write it out in matrix notation.

MLE Notes from GDA II

- "Maximum Likelihood Estimation"
  - General technique to obtain estimators of parameters.

- Suppose $f(x_i|\theta)$ is density of x_i given parameter θ and we want to estimate θ.

- $x_1, \ldots, x_n$ are iid

- To get $f(x_1, \ldots, x_n|\phi) = \prod_{i=1}^{n} f(x_i|\phi) = \mathcal{L}(\phi)$ ← called likelihood

- MLE: maximize this over ϕ to get $\hat{\phi}$

- Intuition: find parameter that makes data most likely

- In HWK, recall

$$(n_1, n_2, \ldots, n_k) \sim \text{Multinomial}(n, \alpha_1, \alpha_2, \ldots, \alpha_k)$$

Let $\alpha_i = \alpha_i(\underbrace{\phi_1, \phi_2, \ldots, \phi_s}_{\text{unknown parameters}})$

$$f(n_1, \ldots, n_k | \phi_1, \ldots, \phi_s) = \frac{n!}{n_1! \cdots n_k!} \prod_{i=1}^{k} \alpha_i(\phi_1, \ldots, \phi_s)^{n_i} = \mathcal{L}(\phi_1, \ldots, \phi_s)$$

- Example: If aa aA AA
  θ_1 θ_2 θ_3

Under HWE

$\theta_1 = (1-p_A)^2$

$\theta_2 = 2p_A(1-p_A)$

And $\theta_1 = p_A$, no θ_2, θ_3

- Typically work w/ log-likelihood
- "Score function" – derivative of log-likelihood

$$S_j := \frac{\partial L}{\partial \theta_j}$$

- Back to example

$$\log\left(\frac{n!}{n_1! n_2! n_3!}\right) + 2n_1 \log_2(1-p_A) + n_2 \log_2 p_A(1-p_A) + 2n_3 \log_2$$

$$\frac{\partial}{\partial p_A}\Big\{ = (2n_1 + n_2)\log_2(1-p_A) + (2n_3 + n_2)\log_2(p_A) + c$$

$$-\frac{2n_1 + n_2}{1-p_A} + \frac{2n_3 + n_2}{p_A} \stackrel{\text{set}}{=} 0$$

$$\Rightarrow p_A(2n_1 + n_2) = (1-p_A)(2n_3 + n_2)$$

$$= p_A(2n_1 + n_2 + 2n_3 + n_2) = 2n_3 + n_2$$

$$= p_A(2n) = 2n_3 + n_2$$

$$\Rightarrow p_A = \frac{2n_3 + n_2}{2n} = \hat{p}_A$$

- Properties of MLE

(1) If $\hat{\phi}$ is the MLE of ϕ and $\theta = g(\phi)$, then $g(\hat{\phi})$ is the MLE of $g(\phi)$.

(2) MLE's are consistent

$$\hat{\phi} \xrightarrow[n \to \infty]{} \phi$$

(3) Standard error for MLE is easy to calculate for large n

$$\hat{\phi} \sim N\left(\phi, \left[\frac{\partial^2 \log L(\phi)}{\partial \phi^2}\Big|_{\phi=\hat{\phi}}\right]^{-1}\right) \quad \text{if one dimension}$$

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \vdots \\ \hat{\phi}_s \end{pmatrix} \qquad I(\phi) = \begin{pmatrix} \frac{\partial^2 \log L(\phi)}{\partial \phi_1^2} & \frac{\partial^2 \log L(\phi)}{\partial \phi_1 \partial \phi_2} & \cdots & \frac{\partial^2 \log L(\phi)}{\partial \phi_1 \partial \phi_s} \\ \frac{\partial^2 \log L(\phi)}{\partial \phi_1 \partial \phi_2} & \frac{\partial^2 \log L(\phi)}{\partial \phi_2^2} & & \\ \vdots & & \ddots & \\ \frac{\partial^2 \log L(\phi)}{\partial \phi_1 \partial \phi_s} & & & \frac{\partial^2 \log L(\phi)}{\partial \phi_s^2} \end{pmatrix}$$

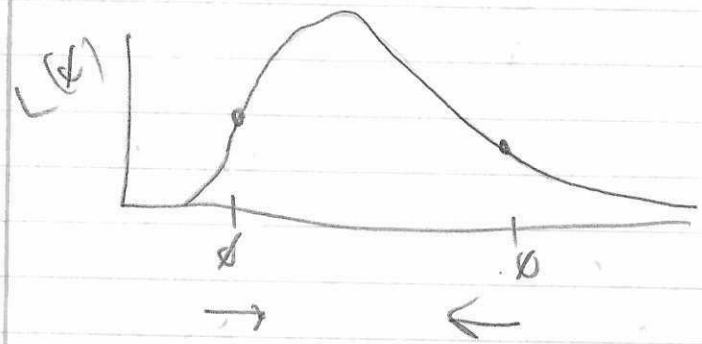
$$\hat{\phi} \sim N_s\left(\phi, \underbrace{[I(\hat{\phi})]^{-1}}_{\text{matrix inverse}}\right)$$

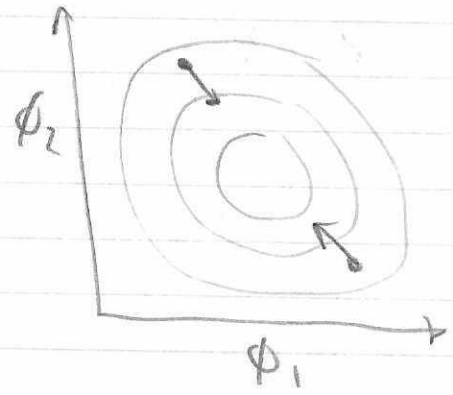
- Typically need to use iterative procedures to get MLE
- Gradient ascent `optim()` in R

$$\phi^{new} = \phi^{old} + \tau \frac{\partial \log L(\phi)}{\partial \phi}$$

τ: step size; $\frac{\partial \log L(\phi)}{\partial \phi}$: gradient

Gradient gives you direction of steepest ascent





- EM algorithm:

Θ - parameter
X - known data
z - latent (unobserved) data

$E_{z|x,\theta}(h|x)$ ← expected value of z given X

$$f_{z|x}(z) = \frac{f(x|z)\, f(z)}{f(x)}$$ depends on Θ

Example of EM

- Let $a_{ik} = Pr(\text{data for individual } i \mid \text{genotype for } i \text{ is } k)$

$$= Pr(d_i \mid g_i = k) \quad i = 1, \ldots, n; \; k = 0, 1, 2$$

g_i is unobserved (latent) $= z$
d_i is observed $= x$

Let π_k: Proportion of individuals w/ genotype k
$= Pr(g_i = k)$

$$Pr(d_i \mid \pi_k) = \sum_{k=0}^{2} Pr(d_i \mid g_i = k) \, Pr(g_i = k) \quad \text{"Law of total probability"}$$

$$= \sum_{k=0}^{2} a_{ik} \pi_k$$

↑ Goal is to maximize this likelihood over π

- Complete log-likelihood: Let z_i = one-of-3 vector indexing genotype

If $g_i = 0$ then $z_i = (1, 0, 0)$ so know $z_i \Leftrightarrow$ know g_i
$g_i = 1$ then $z_i = (0, 1, 0)$
$g_i = 2$ then $z_i = (0, 0, 1)$

then

$$Pr(d_i, z_i \mid \pi) = \prod_{k=0}^{2} \left[a_{ik} \pi_k \right]^{z_{ik}}$$

$$= a_{i0}\pi_0 \text{ if } g_i = 0$$
$$a_{i1}\pi_1 \text{ if } g_i = 1$$
$$a_{i2}\pi_2 \text{ if } g_i = 2$$

So complete log-likelihood:

$$\sum_{i=1}^{n} z_{ik} \sum_{k=0}^{2} \log(a_{ik} \pi_k) = \sum_{i=1}^{n} z_{ik} \sum_{k=0}^{2} \log a_{ik} + \sum_{i=1}^{n} z_{ik} \sum_{k=0}^{2} \log(\pi_k)$$

- EM says

$$\max_{\pi} E[\text{complete-log-likelihood}]$$

$$= \max_{\pi} E\left[\sum_{i=1}^{n} z_{ik} \sum_{k=0}^{2} \log \pi_k\right]$$

$$= \max_{\pi} \sum_{i=1}^{n} E[z_{ik}] \sum_{k=0}^{2} \log \pi_k$$

$$= \max_{\pi} \sum_{k=0}^{2} \left[\sum_{i=1}^{n} E[z_{ik}]\right] \log(\pi_k)$$

Need $E[z_{ik}] = P_r(z_{ik}=1 | d_i, \pi)$ since z_{ik} is an indicator

$$P_r(z_{ik} | d_i, \pi) = \frac{P_r(d_i | z_{ik}=1, \pi)\, P_r(z_{ik}=1)}{\sum_{k=0}^{2} P_r(d_i | z_{ik}=1, \pi)\, P_r(z_{ik}=1)}$$

$$= \frac{a_{ik} \pi_k}{\sum_{k=0}^{2} a_{ik} \pi_k} = w_{ik}$$

$$= \max_{\pi} \sum_{k=0}^{2} \left[\sum_{i=1}^{n} w_{ik}\right] \log(\pi_k) \qquad \text{let } w_k = \sum_{i=1}^{n} w_{ik}$$

$$= \max_{\pi} \sum_{k=0}^{2} w_k \log \pi_k$$

calculus →

$$\hat{\pi}_k = \frac{w_k}{\sum w_k} \quad //$$