Stat 2911 Lecture Notes

Class 13, 2017

Uri Keich

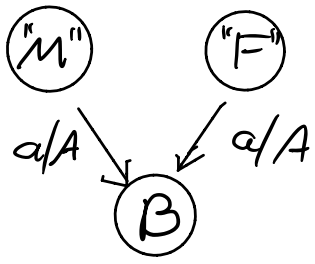
Estimating the minor allele frequency in a Hardy-Weinberg (HW) equilibrium, Comparing estimators: MSE and unbiased estimators (Rice 8.4, 8.5)

Hardy-Weinberg equilibrium (Rice p. 273)

If gene frequencies are at an equilibrium, the genotypes AA, Aa, aa occur in the population with corresponding frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, θ^2 where θ is the frequency of the (minor) allele "a".

How come?

Ignoring the diploid (2-chromosome copy) structure of the parents genome (imagine artificially creating an embryo by randomly drawing individual chromosomes from the population),



$P(M \to a,\ F \to a) =$

$P(M \to a,\ F \to A) = P(M \to A,\ F \to a) =$

Having established those frequencies, note that taking the diploid structure into account we have:

$$P(M \to a) = \theta^2 + \tfrac{1}{2} \cdot 2 \cdot \theta(1-\theta) = \theta,$$

which, reassuringly, is the same as when we ignored the diploidy.

Regardless, we will next assume the HW model applies.

We will analyze a sample from a Chinese population of HK in 1937 which surveys the genotype of the MN erythrocyte (red blood cell) antigens group.
For this gene, genotype=phenotype, or WYSIWYG. (alleles are not dominant/recessive)

The sample:

blood type	M	MN	N	
genotype	AA	Aa	aa	
count	342	500	187	(1029 in total)
	X_1	X_2	X_3	

Goal: estimate the frequency θ of the minor allele "a" in the population.

Model: the sample is taken from a multinomial distribution

with $p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$ and $n = 1029$

$\underline{X} = (\underline{X}_1, \underline{X}_2, \underline{X}_3)$

This is different from estimating $\underline{\theta} = (\theta_1, \theta_2, \theta_3)$ given a multinomial sample: our parameters here depend on a single θ.

We begin with estimating θ using the method of moments.
First, we express θ in terms of moments of the dist.

What is the dist. of X_3?

$X_3 \sim \text{Binom}(n, \quad)$

$\Rightarrow E(X_3) = \quad \Rightarrow \theta = \sqrt{E(X_3)/n}$

$\Rightarrow \tilde{\theta} = \sqrt{E(\tilde{X}_3)/n} = \sqrt{x_3/n} \overset{\text{in our case}}{=} \sqrt{187/1029} \simeq 0.4263.$

Does something look odd about $\tilde{\theta}$?

Clearly, $\tilde{\theta}$ does not take into account all available data!

Can we do better?

Imagine estimating a Bernoulli(p) using only half the data!

What is the proportion of the "a" allele in the sample?

$$\frac{x_1 \cdot 0 + x_2 \cdot 1 + x_3 \cdot 2}{2n}$$

Extending this to a random sample and taking expectations:

$$E\left[\frac{X_1 \cdot 0 + X_2 \cdot 1 + X_3 \cdot 2}{2n}\right] = \frac{1}{2n}\left[E(X_2) + 2E(X_3)\right]$$

$$= \frac{1}{2n}\left[2n\theta(1-\theta) + 2n\theta^2\right] = \theta \qquad X_2 \sim \text{Binom}(n, 2\theta(1-\theta))$$

$\Rightarrow$ (as expected) $\theta = \dfrac{E(X_2) + 2E(X_3)}{2n} \Rightarrow \hat{\theta} = \dfrac{x_2 + 2x_3}{2n} \overset{\text{in our case}}{\simeq} 0.4247$

Both $\tilde{\theta}$ and $\hat{\theta}$ are moment estimators for this problem.

Which one is better?

Intuitively, $\hat{\theta}$ because it uses <u>all</u> available data.

(why?)

MLE

Recall the multinomial pmf: $p_{\underline{N}}(n_1,\dots,n_r) = \binom{n}{n_1 \dots n_r} p_1^{n_1} \cdots p_r^{n_r}$

In our case $p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$ and $n = 1029$ so

$$\ell(\theta) = x_1 \underbrace{\log[(1-\theta)^2]}_{2\log(1-\theta)} + x_2 \underbrace{\log[2\theta(1-\theta)]}_{\log 2 + \log\theta + \log(1-\theta)} + x_3 \underbrace{\log(\theta^2)}_{2\log\theta} + \log\binom{n}{x_1\, x_2\, x_3}$$

$$\dot{\ell}(\theta) = -\frac{2x_1}{1-\theta} + \frac{x_2}{\theta} - \frac{x_2}{1-\theta} + \frac{2x_3}{\theta}$$

$$= \frac{(x_2+2x_3)(1-\theta) - (2x_1+x_2)\theta}{\theta(1-\theta)}$$

$$= \frac{(x_2+2x_3) - 2\theta(x_1+x_2+x_3)}{\theta(1-\theta)}$$

$$= \frac{(x_2+2x_3)/2n - \theta}{\theta(1-\theta)/2n}$$

$\operatorname{sgn}\ell'$: $+$ then 0 then $-$, as $\theta \in (0,1)$ increases, with the sign change at $(x_2+2x_3)/2n$

if $x_2+2x_3 > 0$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{x_2+2x_3}{2n}$$ (same if $x_2+2x_3 = 0$)

Same as $\hat{\theta}$, boosting our confidence in $\hat{\theta}$ but how do we rigorously compare estimators?

Note that even a poor estimator might "get lucky" for any particular sample.
For example, given a Poisson(λ) sample $x_1, \dots, x_n$ we can estimate $\hat{\lambda} = x_1$ instead of $\bar{x}$. Still, if $\lambda = 2$ and $x_1 = 2$ it would probably do better than $\bar{x}$.
Hence, to compare estimators in a meaningful way we need to consider their performance on all possible samples.
So far our setup was: given a single sample $x_1, \dots, x_n$ from a dist. $F(\theta)$ we construct an estimate of θ by $\tilde{\theta} = \varphi(x_1, \dots, x_n)$, e.g., $\hat{\theta} = \bar{x}$.
To consider all possible samples we recall our assumption the sample $x_1, \dots, x_n$ is a particular realization of the $F(\theta)$-distributed RVs $X_1, \dots, X_n$, so we define the RV

$$\Theta = \varphi(X_1, \dots, X_n).$$

Example. Our estimate for the Bernoulli(θ) case was $\bar{x}$. If we now consider the sample consisting of the ind. $X_1, \dots X_n$ Bernoulli(θ) RVs, then our estimator $\bar{x}$ corresponds to the RV $\Theta = \frac{1}{n} \sum_1^n X_i$. Strictly speaking, $\Theta = \Theta(n)$

How can we gauge the quality of an estimator?
What does an estimator need to do?

The estimator $\hat{\Theta}$ needs to estimate the unknown θ.

When is it doing a good job?

The error $\hat{\Theta} - \theta$ needs to be small.

Which samples should we consider?

All, each weighted by its probability, and in order to avoid cancellations between positive and negative errors we need to take $|\cdot|$ or $(\cdot)^2$ of the error.

The mean square error (MSE) of an estimator $\hat{\Theta}$ is defined as,

$$MSE(\hat{\Theta}) := E_\theta[\underbrace{(\hat{\Theta}-\theta)}_{Y}^2]$$

$$= V(\hat{\Theta}-\theta) + [E(\hat{\Theta}-\theta)]^2$$

$$= V(\hat{\Theta}) + [\underbrace{E(\hat{\Theta})-\theta}_{\text{bias}}]^2$$

both positive

An estimator is called unbiased if $E_\theta(\hat{\Theta}) = \theta$ (bias $= 0$).

For unbiased estimators, $MSE(\hat{\Theta}) = V(\hat{\Theta})$.

Examples

1) $X_1, \ldots, X_n$ are ind. Bernoulli(θ) RVs.

$$\hat{\Theta} = \hat{\Theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\hat{\Theta}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \theta$$

$\Rightarrow \hat{\Theta}$ is estimator

$$MSE(\hat{\Theta}_n) = V(\hat{\Theta}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$\overset{2}{=} \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

$$= \frac{\theta(1-\theta)}{n}$$

Do you see a problem?

θ is, of course, unknown, but we can:

(i) use $\theta(1-\theta) \le 1/4$ for $\theta \in [0,1]$ to conclude that $MSE(\hat{\Theta}) \le 1/4n$

(ii) estimate $MSE(\hat{\Theta})$ by plugging in $\hat{\theta} = \bar{x}$

Regardless, clearly $MSE(\hat{\Theta}_n) \to$. (does it make sense?)