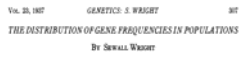
Revisiting allelic frequencies estimation: A decision theory approach to derive Bayes, minimax and admissible estimators

Carlos Alberto Martínez Niño[1,2], Kshitij Khare[2], Mauricio A. Elzo[1]
Department of Animal Sciences[1]
Department of Statistics[2]
University of Florida

UF UNIVERSITY of FLORIDA

Problem and objective

- Necessity of deriving point estimators of allelic frequencies with appealing statistical properties and biological soundness.
- Are allelic frequencies unknown constants or random variables?
- Random variation of allelic frequencies due to certain evolutionary forces (Wright, 1930; 1937).
- The aim of this study was to derive alternative estimators of allele frequencies with optimal statistical properties under a decision theory framework.



Vol. 23, 1937 GENETICS: S. WRIGHT 307

THE DISTRIBUTION OF GENE FREQUENCIES IN POPULATIONS

By Sewall Wright

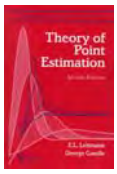
Elements of decision theory

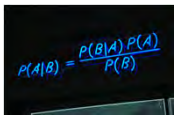
- Parameter space Θ , decision space D,observed data $\boldsymbol{X}$, loss function $L\big(\theta, \delta(\boldsymbol{X})\big)$
- Frequentist risk: $R(\theta, \delta) = E_\theta[L(\theta, \delta(\boldsymbol{X}))]$.
- Decision rules with uniformly smallest risk rarely exist (Lehmann and Casella, 1998): Use a weaker optimality criterion.
- Bayes decision rule:

$$r(\Lambda, \delta^*) = \int_\theta R(\theta, \delta^*) d\Lambda(\theta) = \inf_{\delta \in D} r(\Lambda, \delta).$$

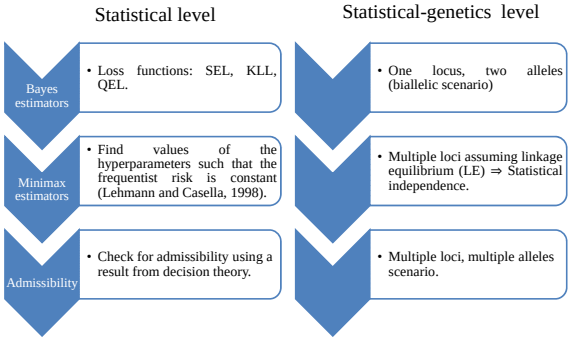
- Minimax decision rule:

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) = \inf_{\delta \in D} \sup_{\theta \in \Theta} R(\theta, \delta)$$





Approach



General setting and notation

Let X_1, X_2 and X_3 be random variables indicating the number of animals having genotypes AA, AB and BB and assume Hardy-Weinberg equilibrium.

Let $\theta :=$ frequency of the "reference" allele B.

$$\boldsymbol{X} := (X_1, X_2, X_3)$$

$$\boldsymbol{X}|\theta \sim Trinomial(n; (1-\theta)^2, 2\theta(1-\theta), \theta^2)$$

$$\theta \sim Beta(\alpha, \beta)$$

Bayes estimators and risks

Loss function	Functional form of loss function	Bayes estimator (BE)	Functional form of BE	Frequentist risk
SEL	$(\theta-\delta)^2$	Posterior mean	$\frac{x_2+2x_3+\alpha}{2n+\alpha+\beta}$	$\frac{2n\theta(1-\theta)+[\alpha(1-\theta)-\beta\theta]^2}{(2n+\alpha+\beta)^2}$
KLL	$E_\theta\left[ln\left(\frac{\pi(\boldsymbol{X}\|\theta)}{\pi(\boldsymbol{X}\|\delta)}\right)\right]$	$\underset{\delta\in D}{\operatorname{argmin}} \int_0^1 L_{KL}(\theta,\delta)\pi(\theta\|\boldsymbol{X})d\theta$	$\frac{x_2+2x_3+\alpha}{2n+\alpha+\beta}$	No closed form
QEL	$\frac{(\theta-\delta)^2}{\theta(1-\theta)}$	Mean of: $w(\theta)\pi(\theta\|\boldsymbol{X})$	See next slide	See next slide

Bayes estimators and risks: QEL

$$\int_0^1 w(\theta)\left(\theta-\hat{\theta}^{QEL}\right)^2\pi(\theta|\boldsymbol{x})d\theta \propto \int_0^1\left(\theta-\hat{\theta}^{QEL}\right)^2\theta^{\alpha-2}(1-\theta)^{2x_1+\beta-2}d\theta$$
$$= \int_0^1\theta^{\alpha}(1-\theta)^{2x_1+\beta-2}d\theta - 2\hat{\theta}^{QEL}\int_0^1\theta^{\alpha-1}(1-\theta)^{2x_1+\beta-2}d\theta$$
$$+\left(\hat{\theta}^{QEL}\right)^2\int_0^1\theta^{\alpha-2}(1-\theta)^{2x_1+\beta-2}d\theta$$

Finite iff $\hat{\theta}^{QEL}=0$.

$$\hat{\theta}^{QEL}=\begin{cases}\frac{x_2+2x_3+\alpha-1}{2n+\alpha+\beta-2}, if\ x_2+2x_3+\alpha-1>0,\ 2x_1+x_2+\beta-1>0\\ 0, \quad if\ x_2+2x_3+\alpha-1\le 0\\ 1, \quad if\ 2x_1+x_2+\beta-1\le 0\end{cases}$$

$$R\left(\theta,\hat{\theta}^{QEL}\right)$$
$$=\begin{cases}\frac{2n}{(2n+\alpha+\beta-2)^2}+\frac{(-\theta(\alpha+\beta-2)+\alpha-1)^2}{\theta(1-\theta)(2n+\alpha+\beta-2)^2}, if\ x_2+2x_3+\alpha-1>0,\ 2x_1+x_2+\beta-1>0\\ \frac{\theta}{1-\theta}, \quad if\ x_2+2x_3+\alpha-1\le 0\\ \frac{1-\theta}{\theta}, \quad if\ 2x_1+x_2+\beta-1\le 0\end{cases}$$

Derivation of minimax rules

Theorem 1 (Lehmann and Casella, 1998). Let Λ be a prior and δ_Λ a Bayes rule with respect to Λ with Bayes risk satisfying $r(\Lambda,\delta_\Lambda)=\sup_{\theta\in\Theta}R(\theta,\delta_\Lambda)$. Then: $i)\ \delta_\Lambda$ is minimax and $ii)\ \Lambda$ is least favorable.

Loss function	Hyperparameters	Functional form of BE	Frequentist risk
SEL	$\alpha=\sqrt{\frac{n}{2}},\beta=\sqrt{\frac{n}{2}}$	$\frac{x_2+2x_3+\sqrt{\frac{n}{2}}}{\sqrt{2n}(\sqrt{2n}+1)}$	$\left(4\left(1+\sqrt{2n}\right)^2\right)^{-1}$
QEL[1]	$\alpha=1,\beta=1$	$\frac{x_2+2x_3}{2n}=MLE$	$\frac{1}{2n}$

[1] Provided $x_2+2x_3+\alpha-1>0,\ 2x_1+x_2+\beta-1>0$.

Extension to k loci (LE and independent priors)

$\boldsymbol{\theta}=(\theta_1,\theta_2,\dots,\theta_k),\boldsymbol{X}=(\boldsymbol{X_1},\boldsymbol{X_2},\dots,\boldsymbol{X_k}),\boldsymbol{X_i}=(X_{1i},X_{2i},X_{3i})$

Minimize: $R(\boldsymbol{\theta},\boldsymbol{\delta})=\int_{\theta_1}\dots\int_{\theta_k}L(\boldsymbol{\theta},\boldsymbol{\delta}(\boldsymbol{X}))\,\pi(\boldsymbol{\theta}|\boldsymbol{X})d\theta_1\cdots d\theta_k$, wrt $\delta_i,\forall\, i=1,2,\dots,k$

$$R(\boldsymbol{\theta},\boldsymbol{\delta})=\int_{\theta_1}\dots\int_{\theta_k}\left(\sum_{i=1}^{k}L(\theta_i,\delta_i(\boldsymbol{X}))\right)\pi(\boldsymbol{\theta}|\boldsymbol{X})d\theta_1\cdots d\theta_k$$
$$=\sum_{i=1}^{k}\int_{\theta_1}\dots\int_{\theta_k}L(\theta_i,\delta_i(\boldsymbol{X}))\prod_{j=1}^{k}\pi(\theta_j|\boldsymbol{X}_j)\,d\theta_1\cdots d\theta_k$$

The h^{th} integral in the summation $(h=1,2,\dots,k)$ can be written as:

$$\int_{\theta_h}L(\theta_h,\delta_h(\boldsymbol{X}))\pi(\theta_h|\boldsymbol{X_h})d\theta_h\times\int_{\theta_1}\dots\int_{\theta_{h-1}}\int_{\theta_{h+1}}\dots\int_{\theta_k}\prod_{j\neq h}\pi(\theta_j|\boldsymbol{X}_j)\,d\theta_1\cdots d\theta_{h-1}d\theta_{h+1}\cdots d\theta_k$$
$$=\int_{\theta_h}L(\theta_h,\delta_h)\pi(\theta_h|\boldsymbol{X_h})d\theta_h$$

Bayes estimation of $\boldsymbol{\theta}$ reduces to that of its components.

Multiallelic loci

Let $\theta_{1_i},\theta_{2_i},\dots,\theta_{n_i}$ be the frequencies of the n_i alleles of locus i.

Let $Y_{1_i},Y_{2_i},\dots,Y_{n_i}$ be random variables indicating the counts of each one of the n_i allelic variants at locus $i,i=1,2,\dots,k$.

$$\boldsymbol{Y}_i:=\left(Y_{1_i},Y_{2_i},\dots,Y_{n_i}\right)\sim Multinomial(2n,\boldsymbol{\theta}_i)$$
$$\boldsymbol{\theta}_i=\left(\theta_{1_i},\theta_{2_i},\dots,\theta_{n_i}\right)\sim Dirichlet\left(\boldsymbol{\alpha}_i=\left(\alpha_{1_i},\alpha_{2_i},\dots,\alpha_{n_i}\right)\right)$$
$$\therefore\ \boldsymbol{\theta}_i|y_i\sim Dirichlet\left(\alpha_{1_i}+y_{1_i},\alpha_{2_i}+y_{2_i},\dots,\alpha_{n_i}+y_{n_i}\right)$$

Under the loss $\sum_{j_i=1}^{n_i}\left(\hat{\theta}_{j_i}-\theta_{j_i}\right)^2$, $\hat{\boldsymbol{\theta}}_i^{M-SEL}=\left(\hat{\theta}_{j_i}\right)_{n_i\times 1}=\frac{\alpha_{j_i}+Y_{j_i}}{2n+\sum_{j_i=1}^{n_i}\alpha_{j_i}}$

Under the loss $\sum_{j_i=1}^{n_i}\theta_{j_i}^{-1}\left(\hat{\theta}_{j_i}-\theta_{j_i}\right)^2$:

$$\hat{\boldsymbol{\theta}}_i^{M-QEL}=\left(\hat{\theta}_{j_i}^{M-QEL}\right)_{n_i\times 1}=\begin{cases}\frac{\alpha_{j_i}+y_{j_i}-1}{\sum_{j_i=1}^{n_i}\alpha_{j_i}+2n-1}, if\ \alpha_{j_i}+y_{j_i}-1>0\\ 0,\ if\ \alpha_{j_i}+y_{j_i}-1\le 0\end{cases}$$

Admissibility

Admissibility of one-dimensional and vector-valued estimators was established using this theorem (Lehmann and Casella, 1998).

Theorem 2. For a possibly vector-valued parameter $\boldsymbol{\theta}$, suppose that δ^π is a Bayes estimator having finite Bayes risk with respect to a prior density π which is positive for all $\boldsymbol{\theta}\in\Theta$, and that the risk function of every estimator δ is a continuous function of $\boldsymbol{\theta}$. Then δ^π is admissible.

Results and comments

$\hat{\theta}^{SEL},\hat{\theta}^{Minimax_1}$ and $\hat{\theta}^{Minimax_2}$ are admissible, for $\hat{\theta}^{QEL}$ the property holds provided $\alpha>1,\beta>1$.

If both alleles are observed: the MLE is also minimax and admissible. We have a Bayes, minimax, admissible and unbiased estimator.

$\hat{\boldsymbol{\theta}}^{M-SEL},\hat{\boldsymbol{\theta}}^{M-Minimax_1}$ and $\hat{\boldsymbol{\theta}}^{M-Minimax_2}$ are admissible, as well as $\hat{\boldsymbol{\theta}}^{M-QEL}$ when $\alpha_{j_i}>1,\forall\, j_i=1,2,\dots,n_i,\forall\, i=1,2,\dots,k$.

The estimators proposed here always have uniformly smaller variance than the MLE, except for those derived from QEL which require: $\alpha+\beta>2$ (biallelic case) and $\sum_{k_i=1}^{n_i}\alpha_{k_i}>1$ (multiallelic case) to meet this property.

Frequentist variances

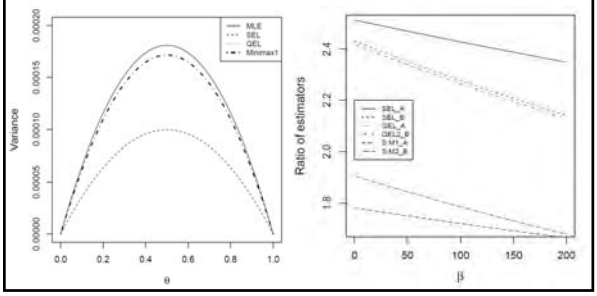
Biallelic case	Multiallelic case
$Var_\theta[\hat{\theta}^{ML}] = \frac{\theta(1-\theta)}{2n}$	$Var_{\theta_{j_i}}\left[(\hat{\boldsymbol{\theta}}_i^{ML})_j\right] = \frac{\theta_{j_i}(1-\theta_{j_i})}{2n}$
$Var_\theta[\hat{\theta}^{SEL}] = \frac{2n\theta(1-\theta)}{(2n+\alpha+\beta)^2}$	$Var_{\theta_{j_i}}\left[(\hat{\boldsymbol{\theta}}_i^{M-SEL})_j\right] = \frac{2n\theta_{j_i}(1-\theta_{j_i})}{(2n+\alpha^*)^2}$
$Var_\theta[\hat{\theta}^{Minimax_1}] = \frac{\theta(1-\theta)}{(\sqrt{2n}+1)^2}$	$Var_{\theta_{j_i}}\left[(\hat{\boldsymbol{\theta}}_i^{M-Minimax_1})_j\right] = \frac{\theta_{j_i}(1-\theta_{j_i})}{(\sqrt{2n}+1)^2}$
$Var_\theta[\hat{\theta}^{Minimax_2}] = Var_\theta[\hat{\theta}^{ML}]$	$Var_{\theta_{j_i}}\left[(\hat{\boldsymbol{\theta}}_i^{M-Minimax_2})_j\right] = \frac{2n\theta_{j_i}(1-\theta_{j_i})}{(2n+n_i-1)^2}$
If $x_2+2x_3+\alpha-1>0$, $2x_1+x_2+\beta-1>0$: $Var_\theta[\hat{\theta}^{QEL}] = \frac{2n\theta(1-\theta)}{(2n+\alpha+\beta-2)^2}$	If $\alpha_{j_i}+y_{j_i}-1>0$: $Var_{\theta_{j_i}}\left[(\hat{\boldsymbol{\theta}}_i^{M-QEL})_j\right] = \frac{2n\theta_{j_i}(1-\theta_{j_i})}{(2n+\alpha^*-1)^2}$

Example

Beta distribution with $\alpha = 240, \beta = 240, n = 691$

Sample sizes: 1382 (A), 691 (B)

$(x_2, x_3) = (10, 25)$, $\alpha = 96$



Results and comments

- For all decision rules derived from SEL, the form of the risk functions shows that they converge to zero as $n \to \infty$. QEL: When all hyperparameters are greater than one, all the derived risk functions converge to zero as $n \to \infty$. When some alleles are not observed and the hyperparameters corresponding to their frequencies are smaller or equal to one, the result does not hold.
- The impact of the use of these estimators in the many applications they could have should be assessed either empirically or theoretically and is an area for further research.

Acknowledgements

- Dr. Malay Ghosh, Department of Statistics, University of Florida.
- Fulbright Colombia and COLCIENCIAS.