Bayesian analysis of the Hardy-Weinberg equilibrium model

Eduardo Gutiérrez Peña

Department of Probability and Statistics
IIMAS, UNAM

6 April, 2010

Outline

Statistical Inference

- Many scientific problems can be modelled in terms of a random variable X. The most common approach is to assume that the (unknown) distribution of X belongs to a parametric family

$$\mathcal{M} = \{p(x|\theta) : \theta \in \Theta\},$$

  indexed by a finite-dimensional parameter θ which characterizes the population under study.

- Given a random sample $X_1, X_2, \ldots, X_n$ from this population, typical statistical problems include:
  - Point estimation: $\hat{\theta}$
  - Interval estimation: $\theta \in (\underline{\theta}, \bar{\theta})$
  - Hypothesis testing: $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

Statistical Inference

- Many scientific problems can be modelled in terms of a random variable X. The most common approach is to assume that the (unknown) distribution of X belongs to a parametric family

$$\mathcal{M} = \{p(x|\theta) : \theta \in \Theta\},$$

  indexed by a finite-dimensional parameter θ which characterizes the population under study.

- Given a random sample $X_1, X_2, \ldots, X_n$ from this population, typical statistical problems include:
  - Point estimation: $\hat{\theta}$
  - Interval estimation: $\theta \in (\underline{\theta}, \bar{\theta})$
  - Hypothesis testing: $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

- Classical statistical methods solve these problems by constructing a *statistic* $T = t(X_1, X_2, \ldots, X_n)$ and optimizing a suitable criterion.
- The relative merits of the resulting methods is judged in terms of the sampling distribution of T for different values of the parameter θ and rely on a frequentist interpretation of probability.
- These methods typically ignore any external evidence (i.e, evidence not contained in the sample).

- Classical statistical methods solve these problems by constructing a *statistic* $T = t(X_1, X_2, \ldots, X_n)$ and optimizing a suitable criterion.
- The relative merits of the resulting methods is judged in terms of the sampling distribution of T for different values of the parameter θ and rely on a frequentist interpretation of probability.
- These methods typically ignore any external evidence (i.e, evidence not contained in the sample).

- Classical statistical methods solve these problems by constructing a *statistic* $T = t(X_1, X_2, \ldots, X_n)$ and optimizing a suitable criterion.
- The relative merits of the resulting methods is judged in terms of the sampling distribution of T for different values of the parameter θ and rely on a frequentist interpretation of probability.
- These methods typically ignore any external evidence (i.e, evidence not contained in the sample).

The Bayesian Approach

- From the Bayesian point of view, the value of the unknown parameter θ is regarded as a random variable whose distribution $p(\theta)$ (the *prior*) describes all the external information available.
- Bayesian inference is based on a subjective interpretation of probability.
- Given the sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from $p(x|\theta)$, the distribution that describes *all* the available information is given by the conditional distribution $p(\theta|\boldsymbol{x})$ (the *posterior*), which may be obtained by means of Bayes' rule, namely

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta)p(\boldsymbol{x}|\theta)}{\int p(\tilde{\theta})p(\boldsymbol{x}|\tilde{\theta})d\tilde{\theta}}.$$

The Bayesian Approach

- From the Bayesian point of view, the value of the unknown parameter θ is regarded as a random variable whose distribution $p(\theta)$ (the *prior*) describes all the external information available.
- Bayesian inference is based on a subjective interpretation of probability.
- Given the sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from $p(x|\theta)$, the distribution that describes *all* the available information is given by the conditional distribution $p(\theta|\boldsymbol{x})$ (the *posterior*), which may be obtained by means of Bayes' rule, namely

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta)p(\boldsymbol{x}|\theta)}{\int p(\tilde{\theta})p(\boldsymbol{x}|\tilde{\theta})d\tilde{\theta}}.$$

The Bayesian Approach

- From the Bayesian point of view, the value of the unknown parameter θ is regarded as a random variable whose distribution $p(\theta)$ (the *prior*) describes all the external information available.
- Bayesian inference is based on a subjective interpretation of probability.
- Given the sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from $p(x|\theta)$, the distribution that describes *all* the available information is given by the conditional distribution $p(\theta|\boldsymbol{x})$ (the *posterior*), which may be obtained by means of Bayes' rule, namely

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta)p(\boldsymbol{x}|\theta)}{\int p(\tilde{\theta})p(\boldsymbol{x}|\tilde{\theta})d\tilde{\theta}}.$$

- The marginal distribution of the data,

$$p(\boldsymbol{x}) = \int p(\theta)p(\boldsymbol{x}|\theta)d\theta,$$

is called the *predictive* distribution.

- Given two alternative models (hypotheses)

$$\mathcal{M}_0 = \{p_0(x|\theta_0) : \theta_0 \in \Theta_0\} \text{ and } \mathcal{M}_1 = \{p_1(x|\theta_1) : \theta_1 \in \Theta_1\},$$

with corresponding priors $p_0(\theta_0)$ and $p_1(\theta_1)$, the Bayes factor in favour of $\mathcal{M}_0$ is defined as

$$B_{01} = \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}$$

and interpreted as the *evidence* contained in the sample against model $\mathcal{M}_1$.

- The marginal distribution of the data,

$$p(\boldsymbol{x}) = \int p(\theta)p(\boldsymbol{x}|\theta)d\theta,$$

is called the *predictive* distribution.

- Given two alternative models (hypotheses)

$$\mathcal{M}_0 = \{p_0(x|\theta_0) : \theta_0 \in \Theta_0\} \text{ and } \mathcal{M}_1 = \{p_1(x|\theta_1) : \theta_1 \in \Theta_1\},$$

with corresponding priors $p_0(\theta_0)$ and $p_1(\theta_1)$, the Bayes factor in favour of $\mathcal{M}_0$ is defined as

$$B_{01} = \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}$$

and interpreted as the *evidence* contained in the sample against model $\mathcal{M}_1$.

The Bayesian Revolution in Genetics

- Beaumont and Rannala (2004) review the use of Bayesian methods is genetic problems such as population genetics, genomics and human genetics (gene mapping).
- They highlight the aspects of many genetic problems that make Bayesian reasoning particularly attractive.
- They also discuss some of the current problems and limitations of Bayesian inference in genetics and outline potential future applications.

The Bayesian Revolution in Genetics

- Beaumont and Rannala (2004) review the use of Bayesian methods is genetic problems such as population genetics, genomics and human genetics (gene mapping).
- They highlight the aspects of many genetic problems that make Bayesian reasoning particularly attractive.
- They also discuss some of the current problems and limitations of Bayesian inference in genetics and outline potential future applications.

The Bayesian Revolution in Genetics

- Beaumont and Rannala (2004) review the use of Bayesian methods is genetic problems such as population genetics, genomics and human genetics (gene mapping).
- They highlight the aspects of many genetic problems that make Bayesian reasoning particularly attractive.
- They also discuss some of the current problems and limitations of Bayesian inference in genetics and outline potential future applications.

The Hardy-Weinberg Equilibrium Model

- The Hardy-Weinberg (HW) model of equilibrium has been of interest to population geneticists in a variety of contexts, most notably evolutionary theory and forensic science.
- At a single locus with two alleles (a dominant "A" and a recessive "a", say), a diploid individual can be one of three possible genotypes, namely: "AA", "Aa", "aa" ("aA" being indistinguishable from "Aa"). Let $\theta_1, \theta_2, \theta_3$ with $\theta_i \geq 0$ and $\theta_3 = 1 - \theta_1 - \theta_2$ be the genotype frequencies in the population.
- Alternatively, θ_i may be thought of as the probability that an individual, randomly chosen from the population, be of genotype i.

The Hardy-Weinberg Equilibrium Model

- The Hardy-Weinberg (HW) model of equilibrium has been of interest to population geneticists in a variety of contexts, most notably evolutionary theory and forensic science.
- At a single locus with two alleles (a dominant "A" and a recessive "a", say), a diploid individual can be one of three possible genotypes, namely: "AA", "Aa", "aa" ("aA" being indistinguishable from "Aa"). Let $\theta_1, \theta_2, \theta_3$ with $\theta_i \geq 0$ and $\theta_3 = 1 - \theta_1 - \theta_2$ be the genotype frequencies in the population.
- Alternatively, θ_i may be thought of as the probability that an individual, randomly chosen from the population, be of genotype i.

The Hardy-Weinberg Equilibrium Model

- The Hardy-Weinberg (HW) model of equilibrium has been of interest to population geneticists in a variety of contexts, most notably evolutionary theory and forensic science.
- At a single locus with two alleles (a dominant "A" and a recessive "a", say), a diploid individual can be one of three possible genotypes, namely: "AA", "Aa", "aa" ("aA" being indistinguishable from "Aa"). Let $\theta_1, \theta_2, \theta_3$ with $\theta_i \geq 0$ and $\theta_3 = 1 - \theta_1 - \theta_2$ be the genotype frequencies in the population.
- Alternatively, θ_i may be thought of as the probability that an individual, randomly chosen from the population, be of genotype i.

- Consider a random sample of n individuals from the population. Conditionally on (θ_1, θ_2), let X_1 and X_2 represent counts of genotype 1 and 2 whose sampling distribution is trinomial with index n and probabilities (θ_1, θ_2):
- The population is said to be in HW-equilibrium if

$$\theta_1 = \theta^2,\ \theta_2 = 2\theta(1-\theta),\ \theta_3 = (1-\theta)^2,$$

for some $0 < \theta < 1$.

- Consider a random sample of n individuals from the population. Conditionally on (θ_1, θ_2), let X_1 and X_2 represent counts of genotype 1 and 2 whose sampling distribution is trinomial with index n and probabilities (θ_1, θ_2):
- The population is said to be in HW-equilibrium if

$$\theta_1 = \theta^2,\ \theta_2 = 2\theta(1-\theta),\ \theta_3 = (1-\theta)^2,$$

for some $0 < \theta < 1$.

- Equilibrium is obtained under the following assumptions: random mating, no mutation, no migration, infinitely large populations size and no selective pressure for or against a particular trait.
- In the simple case described above, θ is the population frequency of allele “A”.
- The HW-model can be used in two ways: either a population is assumed to be in HW-equilibrium, from which the genotype frequencies can be calculated, or, if the genotype frequencies of all three genotypes are assumed known, they can be tested for deviations that are statistically significant.

- Equilibrium is obtained under the following assumptions: random mating, no mutation, no migration, infinitely large populations size and no selective pressure for or against a particular trait.
- In the simple case described above, θ is the population frequency of allele “A”.
- The HW-model can be used in two ways: either a population is assumed to be in HW-equilibrium, from which the genotype frequencies can be calculated, or, if the genotype frequencies of all three genotypes are assumed known, they can be tested for deviations that are statistically significant.

- Equilibrium is obtained under the following assumptions: random mating, no mutation, no migration, infinitely large populations size and no selective pressure for or against a particular trait.
- In the simple case described above, θ is the population frequency of allele “A”.
- The HW-model can be used in two ways: either a population is assumed to be in HW-equilibrium, from which the genotype frequencies can be calculated, or, if the genotype frequencies of all three genotypes are assumed known, they can be tested for deviations that are statistically significant.

- It is often convenient to reparametrize the general trinomial model so as to show more explicitly the departure from the HW-model by means of disequilibrium parameters.
- One such parametrization uses the inbreeding coefficient within populations, here denoted by ϕ. It is given by

$$\theta_1 = \theta^2+\theta(1-\theta)\phi,\ \ \theta_2 = 2\theta(1-\theta)(1-\phi),\ \ \theta_3 = (1-\theta)^2+\theta(1-\theta)\phi.$$

The constraints on f are

$$\max\{-\theta/(1-\theta), -(1-\theta)/\theta\} \leq \phi \leq 1,$$

and $\phi = 0$ corresponds to HW-equilibrium.

- It is often convenient to reparametrize the general trinomial model so as to show more explicitly the departure from the HW-model by means of disequilibrium parameters.
- One such parametrization uses the inbreeding coefficient within populations, here denoted by ϕ. It is given by

$$\theta_1 = \theta^2+\theta(1-\theta)\phi,\ \ \theta_2 = 2\theta(1-\theta)(1-\phi),\ \ \theta_3 = (1-\theta)^2+\theta(1-\theta)\phi.$$

The constraints on *f* are

$$\max\{-\theta/(1-\theta), -(1-\theta)/\theta\} \leq \phi \leq 1,$$

and $\phi = 0$ corresponds to HW-equilibrium.

- Lindley (1988) suggests the following reparametrization

$$\alpha = \frac{1}{2}\log\frac{4\theta_1\theta_3}{\theta_2^2}, \quad \beta = \frac{1}{2}\log\frac{\theta_1}{\theta_3}.$$

  HW-equilibrium obtains then $\alpha = 0$ and $\beta = \log\{\theta/(1-\theta)\}$.

- An important advantage of the (α, β) parametrisation is that the two parameters are variation independent, as opposed to the awkward dependence between θ and ϕ.

- Lindley (1988) suggests the following reparametrization

$$\alpha = \frac{1}{2}\log\frac{4\theta_1\theta_3}{\theta_2^2}, \quad \beta = \frac{1}{2}\log\frac{\theta_1}{\theta_3}.$$

  HW-equilibrium obtains then $\alpha = 0$ and $\beta = \log\{\theta/(1-\theta)\}$.

- An important advantage of the (α, β) parametrisation is that the two parameters are variation independent, as opposed to the awkward dependence between θ and ϕ.

- Consonni et al. (2008) use the HW model to illustrate the concept of compatible priors in the context of model comparison and provide a detailed analysis using Bayes factors.
- The main idea is as follows. When models are nested within a unique encompassing model $\mathcal{M}$, it is natural to perform inference using the prior assigned on the parameter $\theta \in \Theta$ under $\mathcal{M}$, since all models under investigation are obtained through a suitable restriction of Θ.
- When model comparison is performed through the Bayes factor, a specific prior under each submodel is still required. If each of these priors is derived from that on θ under $\mathcal{M}$, we achieve "compatibility" of prior distributions across models, thus alleviating the sensitivity of the Bayes factor to prior specification, and we reduce the burden of the elicitation procedure, which can be especially heavy when the collection of models is large.

- Consonni et al. (2008) use the HW model to illustrate the concept of compatible priors in the context of model comparison and provide a detailed analysis using Bayes factors.
- The main idea is as follows. When models are nested within a unique encompassing model $\mathcal{M}$, it is natural to perform inference using the prior assigned on the parameter $\theta \in \Theta$ under $\mathcal{M}$, since all models under investigation are obtained through a suitable restriction of Θ.
- When model comparison is performed through the Bayes factor, a specific prior under each submodel is still required. If each of these priors is derived from that on θ under $\mathcal{M}$, we achieve "compatibility" of prior distributions across models, thus alleviating the sensitivity of the Bayes factor to prior specification, and we reduce the burden of the elicitation procedure, which can be especially heavy when the collection of models is large.

- Consonni et al. (2008) use the HW model to illustrate the concept of compatible priors in the context of model comparison and provide a detailed analysis using Bayes factors.
- The main idea is as follows. When models are nested within a unique encompassing model $\mathcal{M}$, it is natural to perform inference using the prior assigned on the parameter $\theta \in \Theta$ under $\mathcal{M}$, since all models under investigation are obtained through a suitable restriction of Θ.
- When model comparison is performed through the Bayes factor, a specific prior under each submodel is still required. If each of these priors is derived from that on θ under $\mathcal{M}$, we achieve "compatibility" of prior distributions across models, thus alleviating the sensitivity of the Bayes factor to prior specification, and we reduce the burden of the elicitation procedure, which can be especially heavy when the collection of models is large.

- Consonni et al. (2008) assume that, under the general model, (θ_1, θ_2) is distributed according to a Dirichlet prior, with hyperparameters $m_i > 0$, written $\text{Di}(m_1, m_2, m_3)$.

- The Dirichlet family is the standard conjugate for the general model and allows a closed-form expression for the marginal distribution of the data, which is especially useful in order to compute the Bayes factor. Moreover, it covers a wide range of possible prior specifications.

- Note that the prior distribution may be elicited in terms of the more meaningful parameters θ and ϕ and then translated into (θ_1, θ_2) or any other convenient parametrization such as (α, β).

- Consonni et al. (2008) assume that, under the general model, (θ_1, θ_2) is distributed according to a Dirichlet prior, with hyperparameters $m_i > 0$, written $\text{Di}(m_1, m_2, m_3)$.

- The Dirichlet family is the standard conjugate for the general model and allows a closed-form expression for the marginal distribution of the data, which is especially useful in order to compute the Bayes factor. Moreover, it covers a wide range of possible prior specifications.

- Note that the prior distribution may be elicited in terms of the more meaningful parameters θ and ϕ and then translated into (θ_1, θ_2) or any other convenient parametrization such as (α, β).

- Consonni et al. (2008) assume that, under the general model, (θ_1, θ_2) is distributed according to a Dirichlet prior, with hyperparameters $m_i > 0$, written $\text{Di}(m_1, m_2, m_3)$.

- The Dirichlet family is the standard conjugate for the general model and allows a closed-form expression for the marginal distribution of the data, which is especially useful in order to compute the Bayes factor. Moreover, it covers a wide range of possible prior specifications.

- Note that the prior distribution may be elicited in terms of the more meaningful parameters θ and ϕ and then translated into (θ_1, θ_2) or any other convenient parametrization such as (α, β).

- Under these conditions, all compatible priors for θ assuming HW-equilibrium are of the form Beta(a, b).
- As a consequence, for trinomial data (x_1, x_2, x_3), it can be shown that the Bayes factor in favour of the HW-model takes the form

$$B_{01} = \frac{2^{x_2}\Gamma(M+n)\Gamma(a+2x_1+x_2)\Gamma(b+2n-2x_1-x_2)\Gamma(a+b)\Gamma(m_1)\Gamma(m_2)\Gamma(m_3)}{\Gamma(M)\Gamma(a)\Gamma(b)\Gamma(a+b+2n)\Gamma(m_1+x_1)\Gamma(m_2+x_2)\Gamma(m_3+n-x_1-x_2)},$$

where $M = m_1 + m_2 + m_3$.

- Under these conditions, all compatible priors for θ assuming HW-equilibrium are of the form Beta(a, b).

- As a consequence, for trinomial data (x_1, x_2, x_3), it can be shown that the Bayes factor in favour of the HW-model takes the form

$$B_{01} = \frac{2^{x_2}\Gamma(M+n)\Gamma(a+2x_1+x_2)\Gamma(b+2n-2x_1-x_2)\Gamma(a+b)\Gamma(m_1)\Gamma(m_2)\Gamma(m_3)}{\Gamma(M)\Gamma(a)\Gamma(b)\Gamma(a+b+2n)\Gamma(m_1+x_1)\Gamma(m_2+x_2)\Gamma(m_3+n-x_1-x_2)},$$

where $M = m_1 + m_2 + m_3$.

- Using both real and simulated data, Consonni et al. (2008) compared three procedures to construct compatible priors on θ for a variety of prior specifications on (θ_1, θ_2) under the general trinomial model, including a weakly informative one.

- In this latter case they confirmed an observation by Lindley (1988), regarding the fact that the Bayes factor may favour the HW-model in situations where the classical tests reject the HW-model.

- Using both real and simulated data, Consonni et al. (2008) compared three procedures to construct compatible priors on θ for a variety of prior specifications on (θ_1, θ_2) under the general trinomial model, including a weakly informative one.

- In this latter case they confirmed an observation by Lindley (1988), regarding the fact that the Bayes factor may favour the HW-model in situations where the classical tests reject the HW-model.

Concluding Remarks

- An attractive feature of the Bayesian approach is its ability to incorporate background information into the specification of the model.
- Also, its flexibility allows the researcher to focus on questions and quantities of actual scientific interest.
- However, it can be argued that its current popularity is largely pragmatic and has been made possible by the recent development of computationally intensive Monte Carlo methods.
- Finally, it is important to check the sensitivity of the models to the choice of priors, a task that can be difficult in complicated multiparametric models.

Concluding Remarks

- An attractive feature of the Bayesian approach is its ability to incorporate background information into the specification of the model.
- Also, its flexibility allows the researcher to focus on questions and quantities of actual scientific interest.
- However, it can be argued that its current popularity is largely pragmatic and has been made possible by the recent development of computationally intensive Monte Carlo methods.
- Finally, it is important to check the sensitivity of the models to the choice of priors, a task that can be difficult in complicated multiparametric models.

Concluding Remarks

- An attractive feature of the Bayesian approach is its ability to incorporate background information into the specification of the model.
- Also, its flexibility allows the researcher to focus on questions and quantities of actual scientific interest.
- However, it can be argued that its current popularity is largely pragmatic and has been made possible by the recent development of computationally intensive Monte Carlo methods.
- Finally, it is important to check the sensitivity of the models to the choice of priors, a task that can be difficult in complicated multiparametric models.

Concluding Remarks

- An attractive feature of the Bayesian approach is its ability to incorporate background information into the specification of the model.
- Also, its flexibility allows the researcher to focus on questions and quantities of actual scientific interest.
- However, it can be argued that its current popularity is largely pragmatic and has been made possible by the recent development of computationally intensive Monte Carlo methods.
- Finally, it is important to check the sensitivity of the models to the choice of priors, a task that can be difficult in complicated multiparametric models.

References

- Beaumont, M.A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews (Genetics)* Vol. 5, 251–261.

- Consonni, G., Gutiérrez-Peña, E. and Veronese, P. (2008) Compatible priors for the Hardy-Weinberg equilibrium model. *Test* 17, 585–605.

- Lindley, D.V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.) pp. 307–326. University Press: Oxford.