

Bayesian Statistics for Biological Data:

PEDIGREE ANALYSIS

WILLIAM D. STANSFIELD

MATTHEW A. CARLTON

In teaching biology, there may be a tendency to concentrate too much on the descriptive aspects of the subject. A well-rounded education in the biological sciences also requires experience in the gathering and statistical analysis (interpretation) of quantitative data from field or laboratory studies. There are numerous mathematical tools and computer programs to help us do this today. Introducing students to some of these tools and their practical applications should be part of every biology class. One of these tools is known as Bayesian analysis. The specific purposes of this report are to:

- Introduce Bayes' formula.
- Demonstrate its application to the biological problem of pedigree analysis.
- Illustrate that Bayes' formula and non-Bayesian or "classical" methods of probability calculation may yield different answers.

It is the authors' hope to alert biology teachers to this potential disparity and to underscore the importance of Bayes' formula in pedigree analysis and a wide range of other biological applications.

Typical applications of the Bayesian method involve estimation of an unobservable parameter that describes an entire population using observable (objective) data

derived by sampling techniques (Ledley, 1965). For example, a clinical trial might be designed to test the effectiveness of a drug in reducing the incidence of diabetes in a test group of individuals as compared with a control group of individuals who do not receive the drug, both groups being matched as closely as possible in all other respects (age, sex, lifestyles, health profiles, etc.). Bayesian methods are especially useful for analyzing more complex multivariate problems such as clinical trials designed to simultaneously gather data on two or more variables (e.g., age and drug treatment, or age, sex, and drug treatment).

Awareness of so-called Bayesian statistics certainly is appropriate at the introductory college level. It also could be introduced at the high school level were it not for the fact that many biology teachers have been short-changed in their formal statistical education. This unfortunate situation is likely to continue unless they receive help from sources like *The American Biology Teacher*. We believe our paper could be a first step toward providing the kind of help they need. In applying the information in this report, biology teachers should try to focus their students' attention on the fact that there often is more than one way to analyze biological data and that different analytical procedures may lead to different solutions, rather than focusing merely on the empirical results of a statistical analysis. They should also be made aware of the assumptions underlying the use of any statistical tool. For example, applying an analysis of variance to compare populations that do not roughly conform to normal distributions invalidates the results. Students should at least be made aware that there are

WILLIAM D. STANSFIELD is Professor Emeritus of Biology at California Polytechnic State University, San Luis Obispo, CA 93407. MATTHEW A. CARLTON is in the Department of Statistics at California Polytechnic State University.

statistical tools, e.g., the Kruskal-Wallis test, for analyzing non-normal data distributions. This report demonstrates that Bayesian and non-Bayesian analyses of pedigrees may or may not give the same results, forcing the investigator in the latter case to make value judgements. We also suggest criteria for determining which method is likely to be more appropriate, and we provide “how-to” examples of both types of analyses. A third method (Norton, 1937) for analyzing pedigrees is presented that yields the same result as the Bayesian method, thereby validating that approach. We are not aware of any basic genetics textbook that explains either the value of Bayes’ formula or how to use it in the analysis of pedigrees. Likewise, Norton’s formula does not appear in these books.

Historical Background

The English nonconformist minister and mathematician Thomas Bayes (1702-1761) has been called the Father of Inductive Probability (Anonymous 2). Ever since his pioneering work, the field of statistics seems to have been divided into two camps, the Bayesians and the Non-Bayesians (or frequentists).

[N]either side can clearly be shown to be wrong. When prior probabilities are given as data, the Non-Bayesean (sic) generally has no objection to the use of Bayes (sic) formula, but when prior probabilities are lacking he deplors the Bayesean’s tendency to make them up out of thin air.

(Anonymous 1)

When using Bayesian methods to quantify the probability that a hypothesis is correct, unknown quantities are described by a joint probability distribution. As each piece of evidence is brought into the equation, the effect is conditional on all previous evidence. Assuming that each piece of evidence gives no information about any other piece of evidence avoids this difficulty. However, conditional independence does not always hold (Anonymous 2). Therein lies the basis of much of the controversy over the use of Bayesian inference.

The axioms of probability theory and the algebraic rules for manipulating probabilities that follow from them are generally accepted by both classical and Bayesian statisticians – even Bayes’s (sic) theorem is not questioned with respect to its algebraic validity. The controversy concerns only the definition and interpretation of probabilities, not their algebraic manipulation.

(Weber, 1973)

As long as all the pertinent data (pieces of evidence) become available, does it make any difference (to the probability that our hypothesis is

correct) whether these data are considered random or deterministic? A partial answer to this question will be presented at the conclusion of this report.

Bayes’ Formula

At the heart of Bayesian methodology is Bayes’ formula (also called Bayes’ theorem). To understand Bayes’ formula, one must first understand the notion of **conditional probability**. In words, for two random events A and B, the “conditional probability of A, given B” refers to the chance A will occur under the supposition that B has occurred. For example, shuffle a deck of cards, then define A = {we deal an ace} and B = {we deal a king}. Then while the probabilities of A and B are both 4/52, the *conditional* probability of A, given B, equals 4/51. Why? Because the supposition that B has occurred removes one king from the original deck of 52 cards, leaving only 51 cards (4 of which are aces). In mathematics, we denote these values by $P(A) = P(B) = 4/52$ and $P(A|B) = 4/51$, respectively. The vertical bar (|) may be read as “given.”

Note in the above example that the conditional probability of B given A is also 4/51: $P(B|A) = 4/51$. This is coincidental to the symmetry of events A and B. Bayes’ formula gives the general relationship between these two conditional probabilities:

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)}$$

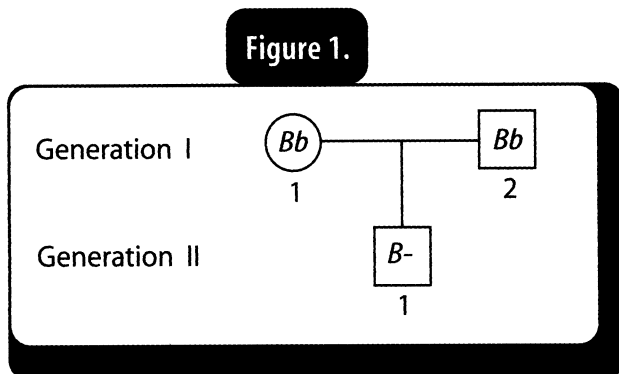
(One can easily verify that the four values from our previous example satisfy this equation.) The derivation of this formula is not complicated and appears in many standard probability and statistics texts (e.g., Peck et al., 2001; Mendenhall et al., 2003).

As noted before, the controversy surrounding so-called Bayesian methods stems not from the validity of this formula, but rather the appropriateness of viewing certain events as “random” and, thus, describable in terms of probability. In what follows, we will consider two examples: one in which Bayes’ formula and the “classical” (frequentist) approach yield the same mathematical result, and one in which they disagree, followed by an independent third method that validates the Bayesian approach.

Pedigree Analysis

The following example is not typical of statistical applications of the Bayesian method because it involves only a small amount of data concerning the genotype of a single individual in a specific pedigree rather than

estimation of a parameter in an entire population. It is, however, a real example of the simplest type of Bayesian analysis, where the *estimand* (an unobserved quantity for which statistical inferences are made) and an individual item of data each have only two possible values (Gelman, 1995). Suppose that black hair color in guinea pigs is governed by a dominant gene (B) and brown color is produced when its recessive allele (b) is in homozygous condition (bb). Consider the pedigree in Figure 1.



Both parents (I1 and I2) in generation I are phenotypically black and genetically heterozygous (Bb). Their male offspring (II1) is black, but its genotype is incompletely known ($B-$). Barring mutation, the process of meiosis should produce B and b gametes with equal frequencies in the parents, just like the tossing of a coin is expected to produce equally frequent heads or tails events. The **unconditional** *a priori* genotypic probabilities for all possible offspring in generation II are $1/4 BB$: $1/2 Bb$: $1/4 bb$. We note that among the black progeny, heterozygotes (Bb) are expected to be twice as frequent as homozygotes (BB); a 2:1 ratio respectively. Thus, once we see that the phenotype of the male offspring II1 is black, we can then predict the **conditional** *a posteriori* probability (among all possible black offspring) that II1 is heterozygous = $2/3$.

Bayes' formula will now be used to derive the same answer (although in other pedigrees this may not always be true, as will be shown later). To do so, we will rewrite the previous formula slightly. We may interpret Bayes' formula as a rule for revising belief in a hypothesis H (i.e., the probability of H) given certain evidence E and background information, or genetic context, G . Bayes' formula then states:

$$P(H|E,G) = \frac{P(H|G)P(E|H,G)}{P(E|G)}$$

Notice that, in addition to a notational change (H for B and E for A), all elements of this probability formula are now conditional upon the genetic context, G .

Let us pause here briefly to understand the elements of the formula.

- The left-hand term, $P(H | E, G)$, represents the *a posteriori* probability that the hypothesis H is true, given both the evidence E and the genetic context G .
- The $P(H | G)$ term is the *a priori* probability of H , given G . In Bayesian terms, $P(H | G)$ reflects our "prior belief" in H before the evidence E is considered.
- The term $P(E | H, G)$ is called the *likelihood*, and gives the probability that our evidence E would occur, assuming the hypothesis H and background information G are true.
- The denominator $P(E | G)$ is independent of H and can be regarded as a normalizing or scaling constant.
- The background information G is a conjunction of all other statements relevant to determining $P(H | G)$ and $P(E | G)$. (Stutz, 1994)

With regard to the pedigree in Figure 1, our context G is the fact that the genotypes of parents I1 and I2 are known for certain to be heterozygous (Bb); our hypothesis H is that II1 is heterozygous (Bb); and our evidence E is that II1 is black. Now, we must calculate the terms on the right-hand side of Bayes' formula.

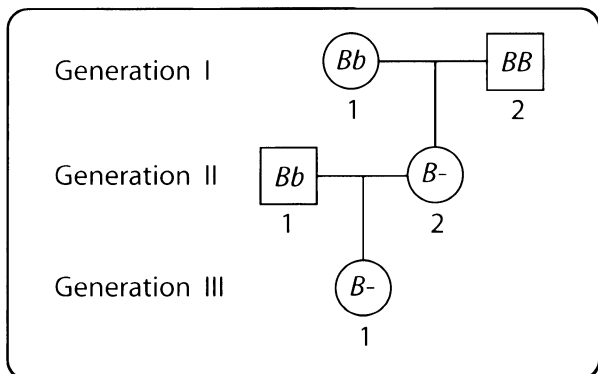
- Within the context G , the *a priori* probability of H (i.e., without considering the evidence E) is $1/2$, since half of all possible offspring from two heterozygote parents are expected to be heterozygous. That is, $P(H | G) = 1/2$.
- If our hypothesis H is true and II1 is heterozygous, then II1 is guaranteed to be black (since black dominates). That is, conditional on H being true, E must be true. Thus, the "likelihood" term $P(E | H,G)$ equals 1 in this case.
- Lastly, in the context G of two heterozygous parents, there is a $3/4$ probability an offspring will be black (all possible outcomes except bb). Hence, $P(E | G) = 3/4$.

Now the probability that our hypothesis H is correct, *after* considering the conditional evidence E that II1 is observed to be black, in the genetic context G of his parents being known heterozygotes, can be calculated using Bayes' formula:

$$P(H|E,G) = \frac{P(H|G)P(E|H,G)}{P(E|G)} = \frac{(1/2)(1)}{(3/4)} = \frac{2}{3}$$

That is, Bayes' formula gives a result identical to the "classical" (frequentist) approach we took previously.

Figure 2.



Without the use of Bayes' theorem, let us analyze the pedigree in Figure 2. Given the genetic context that II2 and III1 are both phenotypically black, our objective is to determine the probability that III1 is heterozygous (*Bb*). The calculation requires the four steps displayed below. In Steps 2 and 3, we will use a more basic fact from probability: if the probability of an event A is p_1 , and the conditional probability of event B given A is p_2 , then the probability of the joint event {A and B} equals the product of $(p_1)(p_2)$.

1. Considering the genotypes of generation I, the probability that II2 is *BB* equals 1/2 and the probability that II2 is *Bb* equals 1/2.
2. In the case that II2 is *BB*, the conditional probability that III1 is *Bb* equals 1/2. Hence, $P(\text{II2 is } BB \text{ and III1 is } Bb) = (1/2)(1/2) = 1/4$.
3. In the case that II2 is *Bb*, the conditional probability that III1 is *Bb* equals 2/3, as calculated in the previous example. Hence, $P(\text{II2 is } Bb \text{ and III1 is } Bb) = (1/2)(2/3) = 2/6 = 1/3$.
4. If the approach taken in Steps 1, 2, and 3 is valid, then we find the total probability that III1 is *Bb* by adding the probabilities of Steps 2 and 3:

$$1/4 + 1/3 = 3/12 + 4/12 = 7/12.$$

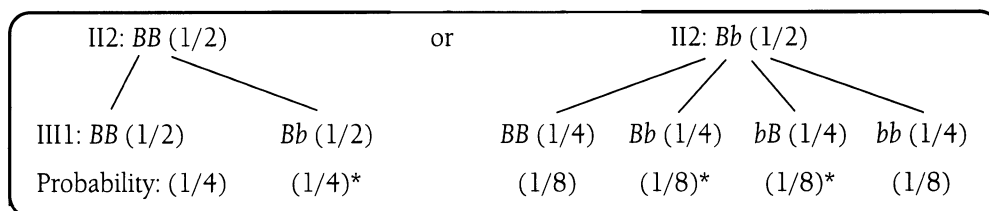
In the previous calculation, the statement that III1 is phenotypically black was not treated as a random event (which would carry a probability), but rather as a known fact. We now analyze the same pedigree with Bayes' formula, to find the probability of the

hypothesis $H = \{\text{III1 is } Bb\}$, given the evidence $E = \{\text{III1 is black}\}$. This will require one additional set of calculations. To that end, let us initially allow all possible genotypes for III1 to be produced, as shown in Figure 3.

The total *a priori* probability, among all possible offspring in generation III, of producing an individual of genotype *Bb* is found by summation of the three asterisk-labeled probabilities in Figure 3: $1/4 + 1/8 + 1/8 = 1/2$. That is, in the genetic context G provided by Figure 2, the *a priori* probability of the hypothesis H that III1 is *Bb* equals 1/2.

If now we are given the additional evidence E that III1 is black, Bayes' formula can be used to modify our

Figure 3.



belief in the previous hypothesis H (viz, the probability that III1 is *Bb* is 1/2). We have computed $P(H | G)$ in the preceding paragraph. By the same reasoning as in our first example—if III1 is of genotype *Bb*, then she must be phenotypically black—the likelihood term again equals 1. Finally, we use the branching diagram above to compute the denominator of Bayes' formula:

$$P(E | G) = P(\text{III1 is black} | G) = 1 - P(\text{III1 is brown} | G) = 1 - 1/8 = 7/8.$$

In the last step, we have used the fact that the brown allele (*b*) is recessive, and hence III1 is phenotypically brown only if she is genotypically *bb*. Therefore, using Bayes' formula, the probability that III1 is *Bb* (H), given evidence E that III1 is black (B-), in the genetic context G that II2 may be either *BB* or *Bb* with probabilities of 1/2 each, is:

$$P(H|E,G) = \frac{P(H|G)P(E|H,G)}{P(E|G)} = \frac{(1/4 + 1/8 + 1/8)(1)}{1-(1/8)} = \frac{1/2}{7/8} = \frac{4}{7}$$

Notice that Bayes' solution (4/7) disagrees with the "classical" solution (7/12).

continued on page 182

Bayesian Statistics for Biological Data: Pedigree Analysis

continued from page 180

An equivalent way to summarize the data for all black individuals in the branching diagram is as follows.

Types of Pedigrees		Combined Probabilities	Ratio
II2	III1		
BB (1/2)	BB (1/2)	1/4 = 2/8	2
BB (1/2)	Bb (1/2)	1/4 = 2/8	2*
Bb (1/2)	BB (1/4)	1/8	1
Bb (1/2)	Bb (1/2)	1/4 = 2/8	2**
Total = 7			

Thus, the *a posteriori* hypothesis that III1 is *Bb* is expected to be true in 4 of every 7 pedigrees of this kind. Step 3 of the non-Bayesian procedure did not allow the unconditional *a priori* production of all possible genotypes in generation III before combining probabilities under the condition that III1 is black, as illustrated in Figure 3 and verified by the use of Bayes' formula. In this particular pedigree, the non-Bayesian solution ($7/12 = 0.5833$) and the solution using Bayes' formula ($4/7 = 0.5714$) are very nearly the same. However, there is a net probability decrease of 0.0119 using the Bayesian approach. It might seem intuitively that Bayesian probabilities should always be greater than probabilities derived by non-Bayesian methods. However, this one example illustrates that this may not always be true.

Comparison with a Third Method

In 1937, H. W. Norton developed general formulas for calculating the probability of homozygosity among individuals exhibiting dominant phenotypes in pedigrees. These formulas were generated independently of Bayes' theorem. According to Norton, if two parental *B*-individuals in a pedigree have probabilities *u* and *v*, respectively, of being homozygous (*BB*), the proportion *p* of *BB* individuals among their *B*-progeny is:

$$p = \frac{1 + u + v + uv}{3 + u + v - uv}$$

In applying this formula to the pedigree in Figure 2, let *p* be the probability that III1 is *BB*, and *u* and *v* be the probabilities that II1 and II2 are *BB*, respectively. Then Norton's formula gives:

$$p = \frac{1 + 0 + (1/2) + 0(1/2)}{3 + 0 + (1/2) - 0(1/2)} = \frac{3}{7}$$

Thus, the probability that III1 is heterozygous (*Bb*) is $1 - (3/7) = 4/7$, in agreement with the solution that was derived above by use of Bayes' theorem.

Conclusion

As noted above, the disparity between the two solutions for the heterozygosity of III1 ($7/12$ vs. $4/7$) depends upon whether we consider the event {III1 is black} to be initially a known fact (the non-Bayesian method) or an event which carries a certain probability (the Bayesian method). The Bayesian method is most useful for revising a prior hypothesis based on new data. Under the assumption that no new data will become available that might cause us to reconsider the prior hypothesis that III1 could be either *BB* or *Bb*, a classical analysis might be considered to be more justified. However, if neither the genotype nor the phenotype of III1 were known initially, but later we learn that III1 is black, then a Bayesian analysis might be considered more justified because the *a priori* possibility that III1 might be brown (*bb*) could be eliminated from consideration. Thus, solutions to problems of this kind may vary depending on whether all pertinent data are initially available or only become available piecemeal.

Acknowledgment

The authors would like to thank John Walker for reviewing an earlier version of this paper.

References

- Anonymous 1. Bayeseans vs. Non-Bayeseans. *The Rancocas Valley Journal of Applied Mathematics*. Retrieved August 2, 2002, from <http://members.tripod.com/Probability/bayes02.htm>.
- Anonymous 2. Thomas Bayes. University of Minnesota. Retrieved August 2, 2002 from <http://www.mrs.umn.edu/~sungurea/introstat/history/w98/Bayes.html>.
- Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis* (pp. 5, 10-11). New York: Chapman & Hall.
- Ledley, R.S. (1965). Scope of computer applications. In T. H. Waterman & H. J. Morowitz (Eds.), *Theoretical and Mathematical Biology* (pp. 284-287). New York: Blaisdell Publishing Company.
- Mendenhall, W., Beaver, R. & Beaver, B. (2003). *Introduction to Probability and Statistics*. Pacific Grove: Duxbury.
- Norton, H.W. III. (1937). General formulae for homozygosity. *Iowa Academy of Science*, XLIV, 139-143.
- Peck, R., Olsen, C. & Devore, J. (2001). *Introduction to Statistics and Data Analysis*. Pacific Grove: Duxbury.
- Stutz, J. & Cheeseman, P. (1994). A short exposition on Bayesian inference and probability. Retrieved August 2, 2002, from <http://ic.arc.nasa.gov/ic/projects/bayes-group/html/bayes-theorem-long.html>.
- Weber, J. D. (1973). *Historical Aspects of the Bayesian Controversy*. Tucson: University of Arizona.