What is Bayesian statistics?

OSVALDO MARRERO

Introduction

Bayesian statistics is included in few elementary statistics courses, and many mathematicians have heard of it, perhaps through collateral readings from popular literature or [1], selected as an Editor's Choice in the *New York Times Book Review*. 'Bayesian statistics' provides for a way to incorporate prior beliefs, experience, or information into the analysis of data. Bayesian thinking is natural, and that is an advantage. For example, on a summer morning, if we see dark rain clouds up in the sky, we leave home for work with an umbrella because prior experience tells us that doing so is beneficial. In general, the idea is simple; schematically, it looks like this:

$$(\text{prior belief}) + (\text{data: new information}) \Rightarrow (\text{posterior belief}).$$

Thus, we begin with a prior belief that we allow to be modified or informed by new data to produce a posterior belief, which then becomes our new prior, and this process is never-ending. We are always willing to update our beliefs according to new information.

Bayesian thinking is very much a part of our time. A number of success stories are mentioned in [1]. Not limited to statistics, Bayesian ideas have also been used in physics to develop quantum Bayesianism (sometimes abbreviated to QBism) to try to explain or remove some of the strange, unusual things that come up in quantum mechanics; see the expository article [2].

Named after Thomas Bayes (1701 or 1702 to 1761), Bayesian statistics has been known – in theory – for a long time. But widespread applications of Bayesian methods had to wait until the 1980s, when useful, workable algorithms and software began to become available.

At present, the kind of statistics that is usually taught in elementary courses is called *frequentist*. The name comes from an interpretation of probability as the ratio of two frequencies or counts. For instance, based on assumed equally likely outcomes, we have the example of selecting counters from a bag.

Probability distributions come in families that are indexed by their parameters. That is, to select a particular member of a probability-distribution family, we need to specify the parameter value(s). A crucial, important difference between frequentist and Bayesian statistics is the way in which parameters are viewed. In the frequentist approach, parameters are considered to have exactly one true, unknown value which we seek to estimate. In the Bayesian approach, parameters are regarded as random variables whose behaviour we are interested in and want to understand.

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

Bayesian inference

In a world where uncertainty reigns, we express our beliefs through probability distributions. The mathematical transporter that takes our prior belief to the posterior—the heart and soul of a Bayesian analysis—is *Bayes's rule* or *Bayes's theorem*, an expression whose simplicity belies its significance. We stress that Bayesian statistics is much more than just using Bayes's rule.

In the discrete case, the result is simple enough to be discussed in elementary, freshman-level texts. Let $A_1, \ldots, A_n$ be a set-theoretic partition of the sample space, with the probability $\mathrm{P}(A_i) > 0$ for $i = 1, \ldots, n$. If B is another event with $\mathrm{P}(B) > 0$, then

$$\mathrm{P}(A_k|B) = \frac{\mathrm{P}(B|A_k)\,\mathrm{P}(A_k)}{\sum_{i=1}^{n} \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}, \qquad k = 1, \ldots, n. \tag{1}$$

The proof follows from the expressions for conditional probabilities

$$\mathrm{P}(A_k|B) = \frac{\mathrm{P}(A_k \cap B)}{\mathrm{P}(B)} \quad \text{and} \quad \mathrm{P}(B|A_k) = \frac{\mathrm{P}(B \cap A_k)}{\mathrm{P}(A_k)},$$

and the *law of total probability*

$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(B|A_i)\,\mathrm{P}(A_i).$$

Bayesian thinking is useful in medicine. Here is an example. A cardiac stress test is used to diagnose heart diseases or to guide treatment of heart disorders. The result from a stress test is not necessarily definitive. If a person fails a stress test, then the focus shifts, for example, to the chances that that person has coronary heart disease. Then we are interested in the probability

$$\mathrm{P}(\text{heart disease} \mid \text{fails test}) = \frac{\mathrm{P}(\text{fails test} \mid \text{heart disease})\,\mathrm{P}(\text{heart disease})}{\mathrm{P}(\text{fails test})}.$$

It is reasonable to believe that a practising cardiologist has a good idea about the size of the probabilities on the right-hand side of the preceding equation, for various subpopulations of patients.

The continuous version of (1) is

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\int_{\theta \in \Theta} p(y|\theta)\,p(\theta)\,d\theta}, \tag{2}$$

where Θ is the *parameter space*, the set containing the possible parameter values.

Note the standard abuse of notation in (2) . In general, the functions $p(\theta|y)$, $p(y|\theta)$ and $p(\theta)$ are all different, but we use the same symbol, p, for all of them. This is often a misuse of function notation, p being thought to stand for the probability operation; this is why many careful authors use the symbol **P**. However, this abuse of notation is popular and well entrenched.

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

We can tell the functions apart by their arguments. With this notation, (1) becomes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}. \tag{3}$$

The following terminology applies to (2) and (3).

- $p(\theta|y)$ is the *posterior distribution*, which uses the prior information and the observed data y to describe the parameter θ. This distribution is the basis for Bayesian inference.
- $p(\theta)$ is the *prior distribution*, what we believe about the parameter θ before we observe any data.
- $\int_{\theta \in \Theta} p(y|\theta)p(\theta)\,d\theta$ and $\sum_\theta p(y|\theta)p(\theta)$ are *normalising constants*, there to ensure that the posterior $p(\theta|y)$ is a bona fide probability distribution that integrates or sums to 1.
- $p(y|\theta)$, viewed as a function of θ for fixed y – that is, after the data y have been collected and observed, is the *likelihood function*. Note that 'likelihood' contrasts with the obvious 'viewed as function of y for fixed θ'. In line with its purpose in Bayes's theorem – (2) and (3), the likelihood function is defined up to multiplication by a constant, because such multiplication will not affect the posterior distribution of θ; the constant will cancel out by the normalising constant. Therefore, in general, the likelihood function is not a probability distribution; its integral or sum over all values need not equal 1. We sometimes write $L(\theta; y)$ to signify the likelihood.

In practice, we omit the normalising constant and abbreviate Bayes's theorem to either

$$p(\theta|y) \propto p(y|\theta)p(\theta) \qquad \text{or} \qquad p(\theta|y) \propto L(\theta; y)p(\theta), \tag{4}$$

where $\propto$ means *is proportional to*. In words,

$$\text{POSTERIOR} \propto \text{LIKELIHOOD} \times \text{PRIOR}.$$

Thus, the likelihood function is the medium through which we update our prior belief to obtain our posterior belief.

The beta and binomial distributions

These are two distributions that we will use later. For a continuous random variable Y, the *beta distribution* is a two-parameter family whose probability density p is given by

$$p(y \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha - 1}(1 - y)^{\beta - 1}, \qquad 0 < y < 1, \tag{5}$$

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

where the *gamma function* $\Gamma : (0, \infty) \to \mathbb{R}$ is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

When n is a positive integer,

$$\Gamma(n) = (n-1)!.$$

The two parameters, $\alpha > 0$ and $\beta > 0$, are *shape parameters* that affect the shape of the graph of the distribution.

For notation, if Y follows a beta distribution with parameters α and β, we write $Y \sim \text{Beta}(\alpha, \beta)$.

Next we state some basic information about the beta distribution.

Theorem 1: (mean and variance for a beta distribution)

If $Y \sim \text{Beta}(\alpha, \beta)$, then

$$\text{E}(Y|\alpha, \beta) = \mu = \frac{\alpha}{\alpha + \beta}$$

and

$$\text{Var}(Y|\alpha, \beta) = \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

We recall that the probability law for the familiar binomial random variable $X \sim \text{B}(n, \theta)$ is given by

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x = 0, \ldots, n.$$

It follows that the corresponding likelihood function

$$L(\theta; x) \propto \theta^x (1-\theta)^{n-x}, \qquad 0 < \theta < 1. \tag{6}$$

Comparing (5) and (6), we see an advantage of the beta distributions in this context. That is, if we choose a beta prior for the binomial parameter θ, then (4) makes it clear that the likelihood and the prior are multiplicative, so that the resulting posterior will have a beta distribution. It therefore makes sense to choose beta priors for θ, as this makes for simple, straightforward analysis. Note that, from (4), as we are dealing with proportionality, there is no need to exhibit the normalising constants. We summarise this in the following result.

Theorem 2: (beta prior for binomial sample data)

For data assumed to come from $\text{B}(n, \theta)$, if we choose a prior

$$\theta \sim \text{Beta}(\alpha, \beta),$$

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

then the posterior

$$\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x),$$

where x is the number of observed successes throughout the n independent trials.

Example: Independence referendum in Scotland

On 18 September 2014, Scots voted in a referendum to determine whether Scotland was going to break away from the United Kingdom (UK), and thus become an independent, separate nation state. Of course, a number of surveys were done before the referendum date. We are going to do a Bayesian analysis of data from a survey of 1206 Scottish adults, conducted during 25-29 June 2014, and reported in [3]. The question was: 'Should Scotland be an independent country?' For our purposes, the answers are going to be 'yes' and 'other', a catchall category containing the non-affirmative answers. Thus we assume the data come from a binomial distribution with sample size $n = 1206$ and unknown parameter θ, the probability that, before the survey was done, a randomly selected such adult will answer 'yes' to the survey question.

Choosing the prior

In this section, I deliberately write in the first person singular, since I am writing about my beliefs. I encourage readers to choose their own priors.

Of course, the data I intend to analyse must not be used to inform my selection of prior, which should reflect my current belief about the parameter θ. There is more than one approach I could follow. There are, for example, a couple of things I could do.

- *Prior probability intervals*: This is a sort of trial-and-error visual approach. Hoping to find an adequate prior, I plot beta densities, as I look for one that matches my current thinking with respect to a prior probability interval for the parameter θ.
  For such an interval, the usual probability levels are 90%, 95%, and 99%, with 95% being the one most commonly used. Also, the probability interval is usually *central* or *symmetric*; for a 95% interval (a, b), this means that the probability mass below a is equal to 0.025, and the probability mass above b is also 0.025. It must be understood that these – the 95% and the symmetry – are not sacrosanct commands or laws that have to be followed; they are merely widely accepted habits.
  Before going on, I want to point out that what I've said applies to a continuous prior distribution. If the prior were discrete, then the *interval* would actually be a discrete set of values. Determining probability intervals for the beta – and other densities – is easy using the software R.

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

- *Prior belief about the mean and variance of* θ: Specify values for the mean and variance of the prior Beta(α, β), and then determine the resulting parameter values for α and β.

To determine my prior, I am going to use the second item above. I believe the referendum was organised because some politicians thought there was a non-trivial chance Scotland would break away from the UK. So, I believe it's plausible that more than 50% would vote 'yes'. At the same time, I believe Scots value tradition, and so they would hesitate breaking up such an old union. Not living in Scotland, I am uncertain about my estimates. To quantify my thinking, I expect 51% (a proportion of 0.51) will vote in favour, with a variance of 0.05, so that the standard deviation is $\sqrt{0.05} \approx 0.224$. Therefore, from Theorem 1, I set the mean of Beta(α, β)

$$\frac{\alpha}{\alpha + \beta} = 0.51$$

and the variance

$$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = 0.05,$$

and I proceed to solve for α and β. Doing so, I obtain $\alpha \approx 2.04$ and $\beta \approx 1.96$. Therefore, I choose the prior

$$\theta \sim \text{Beta}(\alpha = 2.04, \beta = 1.96),$$

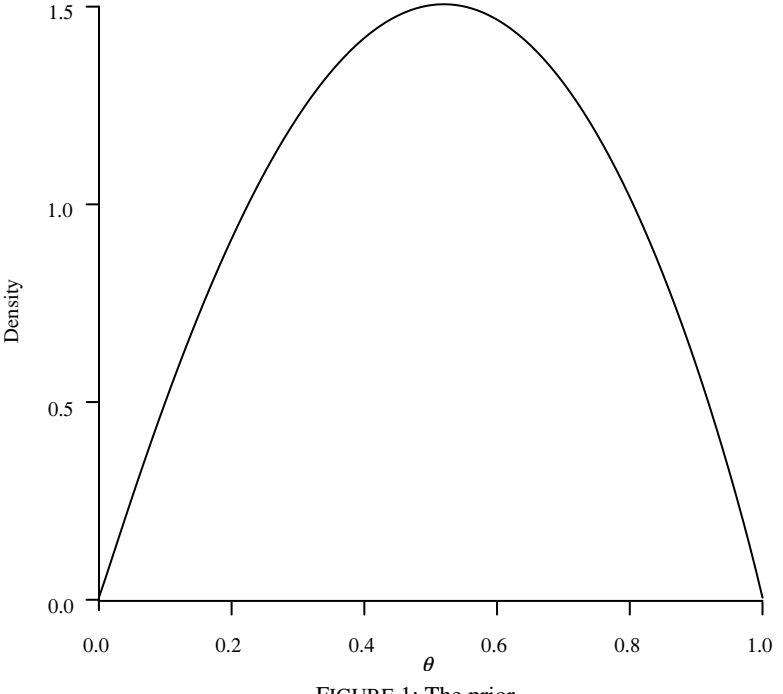


FIGURE 1: The prior

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

whose graph is shown in Figure 1. The large spread in that graph is a result of my uncertainty in selecting the prior. Using R, I obtain the 95% central prior probability interval (0.100, 0.911); thus, my prior belief is that there is a 95% chance that the value of θ lies in this interval.

The data

As reported in [3], of the 1206 people in the survey, 35% voted 'yes'; 54% voted 'no'; 2% 'would not vote'; and 9% said 'don't know'. For our purposes, we pool the data to obtain two classes: 35% in the 'yes' category and 65% in the 'other' category. Thus, of the 1206 people, (0.35)(1206) = 422.1 ≈ 422 said 'yes', and (0.65)(1206) = 783.9 ≈ 784 said something 'other than yes'.

The resulting posterior

From Theorem 2, it follows that the posterior is

$$\theta|x \sim \text{Beta}(2.04 + 422 = 424.04,\ 1.96 + 784 = 785.96),$$

whose graph is shown in Figure 2, along with the prior's graph. Notice that, because it incorporates the sample information, the posterior's graph is much narrower – less variation, less uncertainty – than the prior's graph. Using the software R, we compute the 95% central posterior confidence interval (0.324, 0.378); thus, once we have seen the data, we now believe there is a 95% chance the parameter θ is between 0.324 and 0.378.

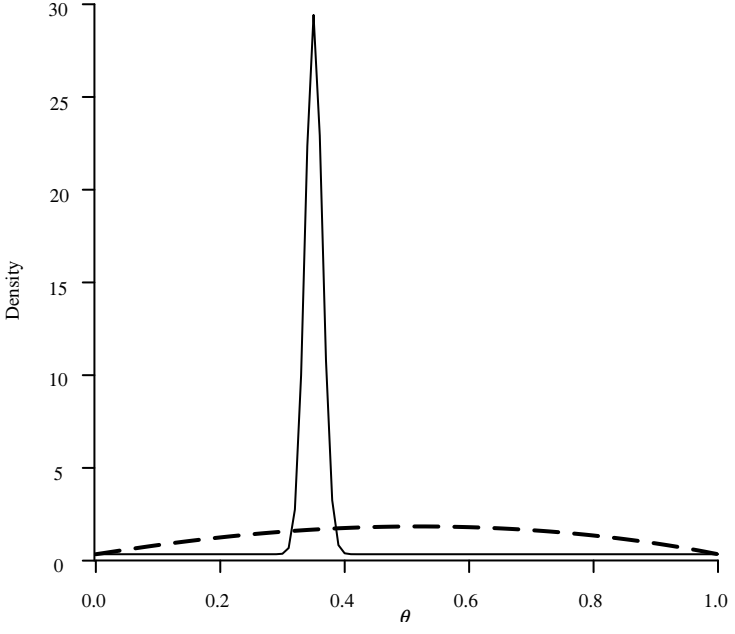


FIGURE 2: The prior (dashed) and posterior

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

Another advantage of the Bayesian approach

In Bayesian statistics, a confidence interval (CI) is an actual probability interval. Thus, when we say that (0.324, 0.378) is a 95% CI for θ, it means that the probability that θ lies between 0.324 and 0.378 is 95%; this is a clear, straightforward meaning. Such interpretation is not valid for frequentist confidence intervals, whose correct meaning appears unnatural to many people; ask anyone who has taught frequentist statistics!

Commentary on the posterior

The posterior in the example is obtained using data collected about three months before the referendum was held. We all know that, with respect to the definitive election-day results, the accuracy of election opinion polls tends to increase as the polls are taken closer to the actual election date. Moreover, the headline in the *Times* article [3] suggests that, in this poll, the decrease in the support for an independent Scotland was 'dramatic', and so perhaps somewhat unrealistically excessive. Thus, it is reasonable to believe that data from a later poll would produce a posterior CI that better matches the actual result.

Robustness of the prior

The posterior depends in part upon the prior, which is for us to choose. Therefore, it is desirable to investigate how changes in the prior affect the resulting concomitant posterior; that is, we want to assess the *robustness* of the prior. As before, we encourage readers to do their own robustness analyses; then it may be useful to note that the expressions in Theorem 1 can be written in the form

$$\alpha = \mu\left\{\frac{\mu(1-\mu)}{\sigma^2} - 1\right\} \qquad \text{and} \qquad \beta = \alpha\left(\frac{1}{\mu} - 1\right).$$

For our robustness analysis, in addition to our original prior, we choose priors with mean 0.6 (*for*), 0.5 (*draw*), and 0.4 (*against*) and, for each such set of three means, we consider variances that decrease progressively from 0.05 to 0.00001; these values were suggested by the referee.

The results of the robustness analysis are summarised in Table 1. The effect of the prior variance upon the posterior distribution is clear and makes sense: strongly held beliefs are not easily modified, and it doesn't take much to change a loosely held opinion. A larger prior variance means less certainty in our prior, which then does not have much effect upon the posterior CI. However, to modify appreciably a prior that we hold with much certainty, we need strong evidence from the data. For instance, in Table 1, consider a prior with mean 0.6. When the variance is 0.05, the corresponding posterior CI is (0.324, 0.378), away from 0.6, and so the prior has virtually no effect upon the CI; but when the variance is 0.00001, the interval is (0.582, 0.594), close to 0.6 and, moreover, this interval is now 4.5 times shorter—in this case, the prior has a clear and definite effect upon the CI.

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

Prior Distribution				Posterior Distribution				
μ	σ^2	α	β	α	β	μ	σ^2	95% CI for θ
0.51	0.05	2.03898	1.95902	424.039	785.959	0.3504	0.000188	(0.324, 0.378)
0.6	0.05	2.28	1.52	424.28	785.52	0.3507	0.000188	(0.324, 0.378)
0.5	0.05	2	2	424	786	0.3504	0.000188	(0.324, 0.378)
0.4	0.05	1.52	2.28	423.52	786.28	0.3501	0.000188	(0.323, 0.377)
0.6	0.002	71.4	47.6	493.4	831.6	0.3724	0.000176	(0.347, 0.399)
0.5	0.002	62	62	484	846	0.3639	0.000174	(0.338, 0.390)
0.4	0.002	47.6	71.4	469.6	855.4	0.3544	0.000173	(0.329, 0.380)
0.6	0.0005	287.4	191.6	709.4	975.6	0.4210	0.000145	(0.398, 0.445)
0.5	0.0005	249.5	249.5	671.5	1033.5	0.3938	0.000140	(0.371, 0.417)
0.4	0.0005	191.6	287.4	613.6	1071.4	0.3642	0.000137	(0.341, 0.387)
0.6	0.00001	14399.4	9599.6	14821.4	10383.6	0.5880	9.6×10^{-6}	(0.582, 0.594)
0.5	0.00001	12499.5	12499.5	12921.5	13283.5	0.4931	9.5×10^{-6}	(0.487, 0.499)
0.4	0.00001	9599.6	14399.4	10021.6	15183.4	0.3976	9.5×10^{-6}	(0.392, 0.404)

TABLE 1: Robustness analysis of the prior distribution

The referendum results

On 18 September 2014, approximately 45% voted 'yes' and 55% said 'no'. Thus Scotland voted to remain in the UK.

Further reading and study

Our short list of references is not meant to be comprehensive.

In terms of Bayesian methods and theory, the books [4, 5, 6, 7] are used in courses on Bayesian statistical methods. The more mathematical, theoretical aspects of Bayesian statistics are presented in [8, 9, 10]. In [11] , the author discusses different approaches to statistical inference. We have [1, 12] which are aimed at a wider audience, .

In terms of software, statistics has always been attuned to computing, which, in turn, has software as an indispensable partner. Widely used by the statistical community, the software `R` is available from www.r-project.org; the `R` package `LearnBayes` is discussed in [4]. A principal reference for the software `BUGS` (Bayesian inference using Gibbs sampling) is [13]; `BUGS` is also discussed in [4, 5]. There is other software, for example, `Stan`, which is discussed in [6].

In terms of Quantum Bayesianism, the article [2] is intended for a general audience. Within the philosophy of present-day physics, [14] has many references.

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press

Acknowledgement

The author thanks the referee for constructive comments that helped improve the paper.

References

1. S. B. McGrayne, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. Yale University Press, New Haven, Connecticut (2011).
2. H. C. von Baeyer, Quantum weirdness? It's all in your mind, *Scientific American* **308** (6) (2013) pp. 46-51.
3. L. McIntosh, S. Coates, Voters turn away from a Scottish nation state: dramatic poll shows slump in Salmond's support, *The Times*, 1 July 2014.
4. J. Albert, *Bayesian computation with* R (2nd edn.), Springer, New York (2009).
5. M. K. Cowles, *Applied Bayesian statistics: with R and OpenBUGS examples*, Springer, New York (2013).
6. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian data analysis* (3rd edn.), CRC Press, Boca Raton, Florida (2014).
7. P. D. Hoff, *A first course in Bayesian statistical methods*. Springer, New York (2009).
8. G. E. P. Box, G. C. Tiao, *Bayesian inference in statistical analysis*, Addison-Wesley, Reading, Massachusetts (1973).
9. D. V. Lindley, *Introduction to probability and statistics from a Bayesian viewpoint*. Part 1 *Probability*, Cambridge University Press (1965).
10. D. V. Lindley, *Introduction to probability and statistics from a Bayesian viewpoint*. Part 2 *Inference*, Cambridge University Press (1965).
11. V. Barnett, *Comparative statistical inference* (2nd edn.), Wiley, New York (1982).
12. D. V. Lindley, *Understanding uncertainty* (Revised edn.), Wiley, Hoboken, New Jersey (2014).
13. D. Lunn, C. Jackson, N. Best, A. Thomas, D. Spiegelhalter, *The BUGS book: a practical introduction to Bayesian analysis*. CRC Press, Boca Raton, Florida (2013).
14. C. G. Timpson, Quantum Bayesianism: A study, *Studies in history and philosophy of modern physics* **39** (2008) pp. 579-609.

10.1017/mag.2016.61

OSVALDO MARRERO

Department of Mathematics and Statistics, Villanova University, 800 Lancaster Avenue, Villanova, Pennsylvania 19085-1699, USA

e-mail: *Osvaldo.Marrero@villanova.edu*

https://doi.org/10.1017/mag.2016.61 Published online by Cambridge University Press