Journal of Education, Society and Behavioural Science

31(1): 1-9, 2019; Article no.JESBS.49580
ISSN: 2456-981X
(Past name: British Journal of Education, Society & Behavioural Science,
Past ISSN: 2278-0998)

Bayesian Benefits for Binomial Applications in Practice

Frank Tuyl[1*] and Peter Howley[1]

[1]School of Mathematical and Physical Sciences/Statistics, The University of Newcastle, Australia.

Authors' contributions

This work was carried out in collaboration between both authors. Author FT chose the case studies, performed the statistical analyses and wrote the first draft of the main body of the manuscript. Author PH managed and developed the introduction, conclusion, literature searches and contextualization. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JESBS/2019/v31i130138
Editor(s):
(1) Dr. Bakare Kazeem Kayode, Professor, Department of Education Foundations, Al-Medina International University, Malaysia.
Reviewers:
(1) Olumide Adesina, Olabisi Onabanjo University, Nigeria.
(2) Jackson Akpojaro, University of Africa, Nigeria.
(3) Kunio Takezawa, Institute for Agro-Environmental Sciences, Japan.
Complete Peer review History: http://www.sdiarticle3.com/review-history/49580

Original Research Article

Received 03 April 2019
Accepted 19 June 2019
Published 26 June 2019

ABSTRACT

Introduction: When it comes to the practice, and teaching, of statistics, the world has primarily focused on what is known as classical or frequentist methods, rather than Bayesian methods.
Scope of the Study: This paper demonstrates some beneficial properties of Bayesian methods within the commonly practiced domain of inference by utilizing consultancy case studies, one concerning an unusual sample size question and one on the detection of mail items with high biosecurity risk material.
Methods: We introduce through practical applications two more aspects of the Bayesian approach which we believe are invaluable to practitioners and instructors. Having in mind readers who may be less familiar with statistical software, we have added Excel instructions which are easy to translate for those who *are* familiar with any such software.
Findings: These cases reflect two valuable aspects for both practitioners and instructors which are unique to the Bayesian paradigm. They are: 1. including prior information to improve inference and how to apply sensitivity analysis to this inclusion and 2. the effortless inference for functions of parameters, compared with frequentist approaches. These examples involving the binomial

**Corresponding author: E-mail: frank.tuyl@newcastle.edu.au;*

parameter have not been considered from this perspective before, are of significant practical value and thus benefit students and instructors of courses teaching Bayesian techniques and endeavoring to include authentic learning experiences.

Keywords: Functions of parameters; informative priors; sensitivity analysis.

1. INTRODUCTION

When it comes to the practice, and teaching, of statistics, the world has primarily focused on what is known as Classical or Frequentist methods. Increasingly, however, Bayesian methods have been included in University programs and utilized in practice to overcome frequentist-based deficits [1]. Further, these tertiary education programs are increasingly being modified to address future workforce needs with the inclusion of practical contexts, enabling students to engage with practitioner-oriented aspects in preparation for their future workplace [2-4].

Many excellent introductions to the Bayesian approach to statistical inference exist, such as Berry [5]; however, in our experience, practitioners who are not also statisticians would typically not read such books, and instructors require authentic examples and applications. Recent articles have attempted to bridge the gap with a more concise and suitable introduction, such as Tuyl and Howley [6] who gave an introduction aimed at readers who are not statisticians, focusing on the attractive automatic treatment, and use, of known constraints within the estimation process to preclude impractical parameter estimates. The main example concerned variance components which, although known to be positive, may lead to negative estimates of such in the classical or frequentist estimation approach. Advantages of the Bayesian approach covered by Tuyl and Howley [6] are as follows:

1. Bayesian inference is based on deriving posterior *distributions* (or simply *posteriors*) of quantities ('parameters') of interest, given the data and *prior distributions* (or simply *priors*). This resulting *graphical* representation of uncertainty is valuable and provides a more holistic view of the situation.
2. As a description of uncertainty around point estimates of the parameters, intervals similar to classical confidence intervals (CIs) are calculated directly from the posterior, by simply taking a 95% (say) segment, and are referred to as credible intervals (CrIs). These credible intervals represent a *probability* statement of the parameter as opposed to the typical confidence statement associated with confidence intervals, the latter based on hypothetical repeated sampling and a source of confusion for many practitioners and non-statisticians. Further, these CrIs do not rely on Normal approximations, unlike CIs which so often do.
3. A posterior distribution is calculated by combining a statistical model of the data with a prior distribution about the parameter(s) of interest. This means that prior information, something which is commonly available, is included and thus contributes to the analysis. Prior information may simply concern known constraints on parameters; in the classical approach it is typically not possible to take such information into consideration, which may lead to inferior inference, including estimates of parameters outside the possible range. Thus the posterior distributions of the aforementioned variance components allowed for positive values only. As shown by Tuyl and Howley [6], whenever frequentist approaches lead to a negative estimate *for a particular data set*, the corresponding posterior simply, and elegantly, has its *mode* at zero, with a tail to the right, allowing the easy calculation of, e.g., a 95% upper limit.

We introduce through practical applications two more aspects of the Bayesian approach which we believe are invaluable to practitioners and instructors. Having in mind readers who may be less familiar with statistical software, we have added Excel instructions which are easy to translate for those who *are* familiar with any such software. The appendices contain examples of the power of statistical software. Note that the popular, but much more involved, technique of Markov chain Monte Carlo simulation (e.g., [7]) is not required here.

1. While the possible incorporation of 'subjective' prior information is arguably a

desirable feature, we recommend to always perform the analysis based on so-called *noninformative* or *objective* priors as well - aimed at 'letting the data speak for themselves' (Bernardo and Smith, sec. 5.6.2) [8], while permitting the additional value from Bayesian over classical methods. Comparing inference based on a subjective prior with inference based on an objective prior then provides a useful type of sensitivity analysis.
2. Bayesian-based inference for functions of parameters tends to be simple, compared with the classical approach. A common example of such functions is the *ratio* of two proportions, or relative risk.

2. TWO BINOMIAL EXAMPLES

The origin and importance of the binomial distribution was described by Tuyl and Howley [6], and is here extended, providing some key mathematical aspects and the use of Excel and R. Because the Normal approximation-based confidence interval for the binomial parameter θ (population proportion) is known to be poor for x close to 0 or close to n, where x represents the count of the events of interest and n the sample size, some introductory statistics books have added Agresti and Coull's [9] 'Plus Four' method (e.g., [10]). This interval is based on adding two successes and two failures to the sample on hand, before applying the Normal approximation. This interval and the Score interval it is based upon (recommended by Brown et al. [11]) appear to be primitive approximations of the Bayesian method described below, since they address skewness of the likelihood function in an ad hoc manner. Tuyl et al. [12] showed that these intervals are *still* inadequate for x = 0 or x = n.) As Tuyl and Howley [6] described, the likelihood function follows from viewing a model such as the binomial distribution, i.e.,

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

as a function of θ instead of x, for a *particular* value of x – the situation within which practitioners find themselves. When viewing (1) in this manner, it can be seen to be proportional to a beta(x+1,n-x+1) distribution, which is continuous between 0 and 1. (See Berry Ch. 6 [5] for a non-mathematical introduction to this distribution.) This formulation is due to a beta(a,b) distribution being defined as proportional to $\theta^{a-1}(1-\theta)^{b-1}$. It follows that when combining a beta(a,b) *prior* distribution with the aforementioned likelihood using Bayes' theorem, the posterior distribution is

$$p(\theta\,|\,x)=\frac{1}{B(x+a,n-x+b)}\theta^{x+a-1}(1-\theta)^{n-x+b-1}, \tag{2}$$

which is a beta(x+a,n-x+b) distribution. (Here B() is the so-called beta *function*, consisting of factorials similar to the ones occurring in (1), but allowing for non-integer values if necessary.) As Gelman et al. [7] pointed out, the beta(a,b) prior may thus be seen as representing a-1 prior successes and b-1 prior failures. This concept may be helpful when formulating informative priors, but it is typically easier to think about a prior mode and a quantile to derive reasonable values of a and b. See Chun-Lung's BetaBuster software as a user-friendly method of setting an informative prior (http://betabuster.software.informer.com/download). It also follows from the above beta(a,b) definition that the beta(1,1) or uniform prior is the recommended noninformative prior, representing zero prior successes and failures. This was adopted by Bayes [13], but other candidate priors have been suggested as well, arguably unnecessarily [14].

Most importantly, for extreme values of x, the beta posterior may be quite skewed. Significant skewness is present in the posterior distributions derived in the examples below. In such situations highest posterior density (HPD)-based intervals would seem preferable to central intervals. HPD intervals are the shortest of the many potential Bayesian credible intervals [15]). Such intervals are fairly easily implemented in e.g. Excel (using its BETA.DIST function and its Solver utility), and a useful reference is M'Lan et al. [16]. (Alternatively, see the second example and the Appendices for using simulation to obtain HPD intervals, and central intervals as well.) In the first example below, the HPD interval is additionally one-sided, so that the Solver utility is not required. Generally, given a required confidence level, both central and one-sided limits may be found directly by the Excel BETA.INV function.

The two binomial-based examples are as follows:

1. An actual case study based on x = 0 'successes' in n observations, which demonstrates how prior information may be incorporated quite naturally and how

straightforward it is to check the *effect* of this prior information on resulting inference about the true proportion θ, compared with *not* including prior information.

2. An example of a function of two binomial parameters, where classical inference appears mostly limited to using Normal approximations, but Bayesian inference follows directly from the posterior distributions of the two parameters.

2.1 Prior Information and Sensitivity Analysis

A consultation to an environmental engineering firm concerned a sample size determination, where the population proportion of ground samples having a certain property (presence of traces of a certain chemical) was known to be small. This confirmatory sample size was required to be such that various parties involved could be 95% confident that the true proportion of samples with traces of the chemical was below 0.05, i.e., $\Pr(\theta < 0.05 \mid x, n) = 0.95$. The engineers expected it to be smaller, but the other parties were satisfied with the aforementioned outcome from the confirmatory sample. The engineers knew about the relationship between sample size and confidence interval calculations and realized that an estimated proportion close to 0 caused a problem with their 'text book' (Normal approximation) formula – which is why they sought our assistance. They requested a sample size based on an anticipated $x = 0$ and understood that if in fact $x > 0$ were to eventuate, the resulting confidence interval would not satisfy the original requirement.

As mentioned, the true proportion was known to be small; it was conservatively stated to be below 0.1 with 90% confidence, as agreed upon by the multiple interested parties. It appears impossible to include this type of prior information in a classical analysis. In the Bayesian context of small proportions, informative priors are typically chosen from the beta(1,b) family [14], and here the above prior information could be well-represented by a beta(1,22) distribution: it may be found by quick trial and error that Excel's BETA.DIST(0.1,1,22,1) equals 0.9, where the 1 in the 4th field requests the 'cumulative' beta distribution, thus corresponding to $\Pr(\theta < 0.1 \mid a = 1, b = 22)$. The corresponding graph (Fig. 1) of this particular beta distribution for the prior was confirmed as reasonable, i.e., conservative, by the engineers. They liked the idea of this prior representing 21 prior samples without traces, and understood that the sample size needed to be such that the probability the true proportion was below 0.05 would be lifted from 68% (i.e., $\Pr(\theta < 0.05 \mid a = 1, b = 22) = 0.68$), under this prior, to at least 95%, based on $x = 0$. As mentioned, the one-sided intervals studied here, in the context of $x = 0$, are in fact HPD intervals, with the lower limit at zero, similar to the variance components example in Tuyl and Howley [6].

A similar trial and error calculation, assuming $x = 0$, resulted in a posterior beta(1,59), also shown in Fig. 1, that just satisfies this post-data collection requirement (i.e., $\Pr(\theta < 0.05 \mid x{+}a = 1, n{-}x{+}b = 59)$ = BETA.DIST(0.05,1,59,1) = 0.95). This implied that the parties could use a sample size of $n = 59 - 22 = 37$. Compared with a noninformative uniform (beta(1,1)) prior this was a saving of 21 samples, which was a substantial amount of money and time!

As expected, the actual number of samples with traces was zero. Everyone understood very well that, based on the stated prior information, the requirement had been satisfied with 95% probability. The sensitivity analysis referred to in the Introduction works as follows: the probability that $\theta < 0.05$, given the data $x = 0$, $n = 37$, *without* using any prior information (with the aim of letting the data speak for themselves), is given by the same calculation as before, but with $a = b = 1$ instead. It follows that $\Pr(\theta < 0.05 \mid x{+}a = 1, n{-}x{+}b = 38)$ = BETA.DIST(0.05,1,38,1) = 0.86, i.e., 86% instead of 95%, based on using the noninformative beta(1,1) prior instead of the informative beta(1,22). This result (known before collection of the 37 samples) was considered acceptable, where a noninformative prior-based probability of, say, 40% (instead of 86%) would have been less so.

2.2 Functions of Parameters

Decrouez and Robinson [17] were interested in the weighted sum of two proportions, in the context of developing performance indicators in the operation of quarantine inspection, with a view to detecting mail items with high biosecurity risk material. The observed numbers of 'successes' are represented by x_1 and x_2, with respective sample sizes n_1, the number of mail items inspected by x-ray or detector dogs, and n_2, the number of manually inspected items selected from all non-intercepted items.

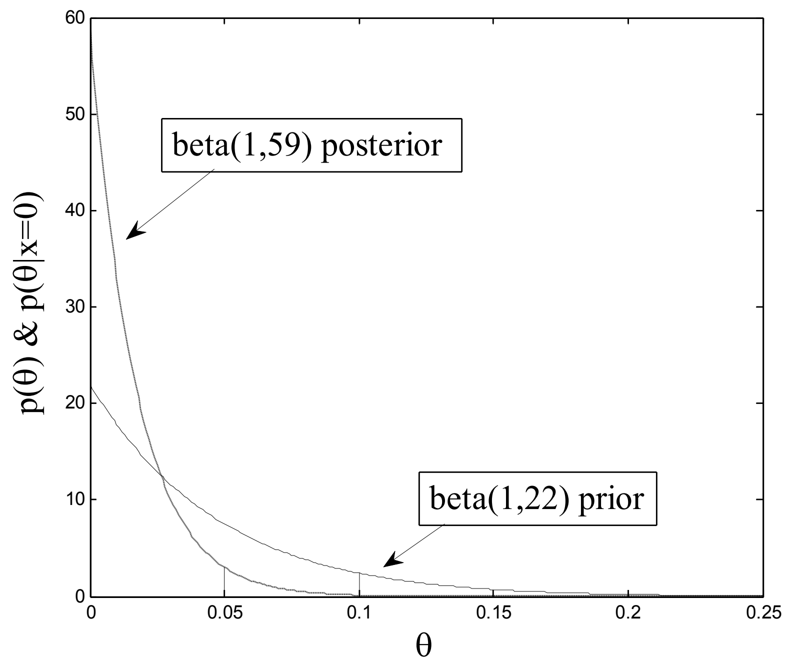

Fig. 1. Informative beta(1,22) prior and beta(1,59) posterior, based on x = 0 and n = 37

The function of interest is $\varphi = a\theta_1 + b\theta_2$ (the "pathway-level leakage rate", based on two streams of leaked items), where *a* and *b* are known constants. Decrouez and Robinson [17] considered various Normal approximation-based intervals, some requiring quite lengthy calculations, but all ad hoc and unnecessary, from a Bayesian perspective. [Note that the frequentist approach typically requires sampling distributions of statistics corresponding to parameters of interest. If the parameters, or functions of interest thereof, are complex, Normal approximations are usually applied to proceed, which is arguably inadequate when the underlying likelihood function is skewed, for example.] The calculation required to move from $p(\theta_1, \theta_2 \mid x_1, x_2)$ to $p(\varphi \mid x_1, x_2)$ may be difficult mathematically, but is addressed through simple Monte Carlo simulation: when sampling from a given posterior, or posteriors, at the *same* time samples of *any* function $\varphi = g(\theta_1, \theta_2)$ can be calculated, and collected for the purpose of a histogram. See Hashemi et al. [18] for the mathematics required for the usual quantities resulting from a 2×2 contingency table, i.e., absolute risk $\theta_1 - \theta_2$, relative risk θ_1 / θ_2 and odds ratio $\frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}$. Again, practitioners may obtain intervals for these quantities more easily by simulation, as shown for the current weighted sum example; Hashemi et al.'s [18] methods are useful (faster) when calculation of many thousands of intervals is required, for theoretical purposes.

Decrouez and Robinson's [17] Table 3 shows inference for four data sets. Their statement that, due to the large sample sizes involved, "the normal approximation is very accurate" (p.292) is contentious: all four values for x_2 are relatively small, so such an approximation is not expected to work well necessarily. In the Bayesian context, however, no such assumptions are needed. For illustration purposes, the data set with $n_1 = 748559,\ x_1 = 139,\ n_2 = 4162,\ x_2 = 2$ is chosen; details are given in Appendix A, which show that the uncertainty of $\varphi = a\theta_1 + b\theta_2$, with $a = 0.114$ and $b = 0.995$, is almost entirely due to θ_2, such that x_2 being so close to zero seems all the more relevant to consideration of the appropriateness of Normal approximations.

The histogram for φ resulting from the simple simulation is shown in Fig. 2. With the point estimate at 0.0499%, the Normal approximation seems to have overshot the lower limit: the

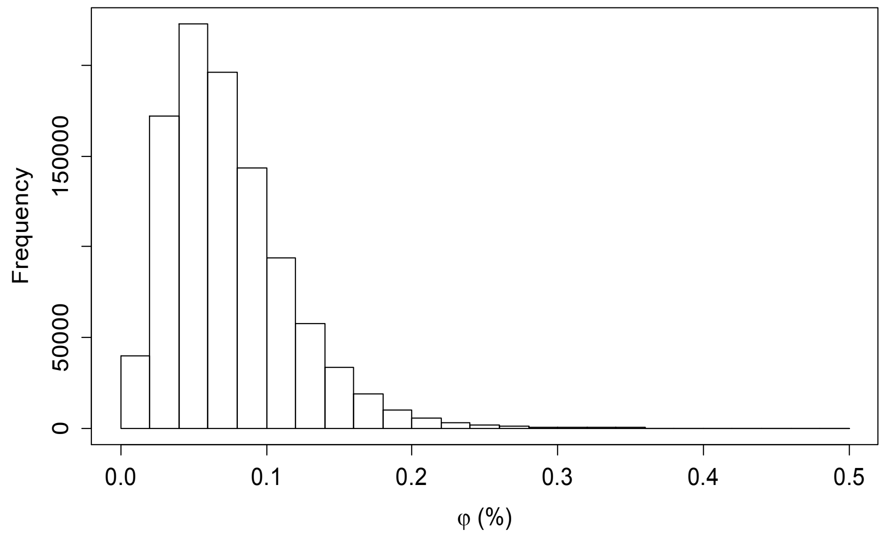

Fig. 2. Posterior distribution of the weighted sum of two binomial proportions, based on an example by Decrouez and Robinson [17]

interval given by Decrouez and Robinson [17] is (0.0000%, 0.1749%). Instead, the 95% central CrI (with 2.5% in each tail) is (0.0169%, 0.1747%), which effectively has the same upper limit but a more realistic and useful nonzero lower limit. Although a correct 95% probability interval, the disadvantage of this central interval is that arguably the lower limit is too large, in the sense that it sacrifices small values of φ with relatively high density for large values of φ with lower density (that *are* inside the interval). As stated earlier, the common approach to take into account such skewness of a posterior is to calculate the HPD interval, which is (0.0094%, 0.1551%) in this case. Note that with nonlinear transformations of θ_1 and θ_2, such as the relative risk θ_1 / θ_2, more care must be taken with how to choose an 'appropriate' 95% CrI [19,20], which is beyond the scope of this article's aims. The important point here is that the HPD interval easily sits inside the confidence interval, i.e., is less conservative, with a more useful positive lower limit. This interval also appears to be an improvement on the central CrI, by redistributing the probability in the latter's tails, each of 2.5%. In fact, it may be checked that the HPD interval's tails are approximately 0.4% and 4.6%, respectively.

In short, in contrast with confidence intervals, this credible interval is not based on hypothetical repeated sampling, or on assumptions of Normality, and allows the more natural statement that there is a 95% probability that the pathway-level leakage rate is between 0.0094% and 0.1551%, based on the two sample proportions in question. It appears that the approximate 95% confidence interval (0.0000%, 0.1749%) quoted by Decrouez and Robinson [17] is too conservative, and arguably reflects a *probability* of approximately 97.5% of containing the true rate. Frequentist coverage calculations (work in progress), such as performed by Agresti and Min [19], appear to confirm the above claims, but are beyond the scope of this article.

3. CONCLUSION

The importance of an increased toolkit of methods for the practitioner is becoming increasingly acknowledged in the literature, as part of a growing focus on statistical or data science techniques to best utilize and interrogate data sets. There is also a great need to provide examples of such to practitioners who may otherwise not be aware, and to tertiary education instructors to enable students to be better prepared for the practical rigours of the workforce. To this end, focus has been on exemplifying accessible Bayesian techniques we consider significant for practitioners, and for instructors who may also utilize these in their courses, and in a manner not usually considered. The Bayesian view is important for our first example from practice, which solved an important problem for an engineering firm by incorporating prior information in a statistical analysis (followed by a sensitivity analysis of the use of this information). We also give an example of the ease with which seemingly difficult quantities of interest may be analysed by Monte Carlo simulation, and illustrate the additional visual benefits arising from the posterior distributions inherent to Bayesian methods, for both examples.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Zyphur MJ, Oswald FL, Rupp DE. Rendezvous overdue: Bayes analysis meets organizational research. Journal of Management. 2015;41(2):387–389.
2. Howley P. Keeping it real, keeping them interested and keeping it in their minds. Journal of Statistics Education. 2008;16(1).
3. Simpson JM, Ryan P, Carlin JB, Gurron L, Marschner I. Training a new generation of biostatisticians: A successful consortium model. Journal of Statistics Education. 2009;17(2).
4. Tran TT. Enhancing graduate employability and the need for university-enterprise collaboration. Journal of Teaching and Learning for Graduate Employability. 2016;7(1):58-71.
5. Berry DA. Statistics - A Bayesian perspective. Belmont, California: Duxbury; 1996.
6. Tuyl F, Howley P. Simplifying life through Bayes: Hints for practitioners new to Bayesian inference. Quality Management Journal. 2016;23:22-28.
7. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis, 3rd Edition, New York, NY: Chapman & Hall/CRC; 2013.
8. Bernardo JM, Smith AFM. Bayesian Theory. New York, NY: Wiley; 1994.
9. Agresti A, Coull BA. Approximate is better than ‘exact’ for interval estimation of binomial proportions. The American Statistician. 1998;52:119-126.
10. Moore DS, McCabe GP, Alwan LC, Craig BA, Duckworth WM. The practice of statistics for business and economics. W.H. Freeman and Company; 2011.
11. Brown LD, Cai T, DasGupta A. Interval estimation for a binomial proportion (with discussion). Statistical Science. 2001;16: 101-133.
12. Tuyl F, Gerlach R, Mengersen K. The rule of three, its variants and extensions. International Statistical Review. 2009;77: 266-275.
13. Bayes TP. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London. 1763;53:370-418.
14. Tuyl F, Gerlach R, Mengersen K. A comparison of Bayes–Laplace, Jeffreys, and other priors: The case of zero events. The American Statistician. 2008a;62:40-44.
15. Box GEP, Tiao GC. Bayesian inference in statistical analysis. New York, NY: Wiley; 1973.
16. M'Lan CE, Joseph L, Wolfson DB. Bayesian sample size determination for binomial proportions. Bayesian Analysis. 2008;3:269-296.
17. Decrouez G, Robinson AP. Confidence intervals for the weighted sum of two independent binomial proportions. Australian & New Zealand Journal of Statistics. 2012;54:281-299.
18. Hashemi L, Nandram B, Goldberg R. Bayesian analysis for a single 2×2 table. Statistics in Medicine. 1997;16:1311-1328.
19. Agresti A, Min Y. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. Biometrics. 2005;61: 515-523.
20. Tuyl F, Gerlach R, Mengersen K. Inference for proportions in a 2×2 contingency table: HPD or not HPD? Biometrics. 2008b;64: 1293-1296.
21. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; 2019.
Available:http://www.R-project.org/

Appendix A

R code [21] to produce the histogram in Fig. 2 follows below, with comments. After setting data values, posteriors of the two proportions are derived as follows:

```
n1=748559; x1=139; n2=4162; x2=2; k=1000000
theta1=rbeta(k,x1+1,n1-x1+1); theta2=rbeta(k,x2+1,n2-x2+1)
```

Here n_1 is the number of mail items inspected by x-ray or detector dogs, n_2 is the number of manually inspected items selected from all non-intercepted items, and x_1 and x_2 are items intercepted as having high biosecurity risk material. The rbeta function is then used for single statement Monte Carlo simulations, to create large vectors (size *k*) of samples from the posteriors (normalized likelihood functions) of θ_1 and θ_2. After setting the total number of mail items *N*, and *a* and *b* as defined by Decrouez and Robinson [17], only one more line of code is required:

```
N=845007; a=1-n1/N; b=1-n2/N
phi = a*theta1 + b*theta2
```

The beautiful simplicity is that a vector of samples from the posterior of φ may be obtained in this manner. Code for the histogram in Fig. 2 (with φ as a percentage), and the central and HPD intervals (see Appendix B) mentioned in the text, follows below.

```
phi = 100*phi; hist(phi)
Central = quantile(phi,c(.025,.975))
HPD = HPDsample(phi,.95)
```

Appendix B

The R code [21] below describes a function that produces a highest posterior density (HPD) interval, from a large number of samples from a posterior distribution. The first argument of the function contains the vector of samples, the second the required confidence level. Appendix A contains an example of how this function is called. Calling this function for multiple simulations is recommended, to check the number of correct decimal places.

The function is based on the simple idea that a, say, 95% HPD interval is such that it contains 95% of ordered samples with minimum difference between the smallest and largest values. That is, having approximately equal posterior height at the lower and upper limits, the HPD interval is by definition the *shortest* interval.

```
HPDsample = function(y,conflev)
{
m=length(y)
lag=conflev*m
x=sort(y,decreasing=FALSE)
length=diff(x,lag-1)
k=which.min(length)
hpd=array(0,2)
hpd[1]=x[k]
hpd[2]=x[k+lag-1]
return(hpd)
}
```


Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle3.com/review-history/49580