Revista Colombiana de Estadística - Applied Statistics
July 2021, volume 44, issue 2, pp. 243–252
http://doi.org/10.15446/rce.v44n2.89102

Bayesian Modeling Competitions for the Classroom

Concursos de modelos bayesianos para el salón de clases

Arthur Berg[a]

College of Medicine, Biostatistics and Bioinformatics, Penn State University, Hershey, USA

Abstract

Three educational and engaging competitions are described for students studying Bayesian statistics. These competitions are designed to help students explore the topics of James-Stein estimation, the German tank problem, and resampling inference. These competitions will inspire students to think creatively, challenge students to develop effective Bayesian models, and motivate students to pursue excellence in competition with their peers. The competition structures can be easily adapted for use in introductory or advanced Bayesian statistics courses.

***Key words*:** James-Stein estimation; German tank problem; Capture-recapture.

Resumen

Se describen tres concursos educativos y atractivos para los alumnos que estudian estadística bayesiana. Estos concursos están diseñados para ayudar a los estudiantes a explorar los temas del estimador de James-Stein, el problema de los tanques alemanes y la inferencia de remuestreo. Estos concursos inspirarán a los estudiantes a pensar de forma creativa, les desafiarán a desarrollar modelos bayesianos eficaces y les motivarán a buscar la excelencia en competencia con sus compañeros. Las estructuras de los concursos pueden adaptarse fácilmente para su uso en cursos de estadística bayesiana introductorios o avanzados.

***Palabras clave*:** Estimador de James-Stein; Problema de los tanques alemanes; Captura-recaptura.

[a]Penn State University. E-mail: berg@psu.edu

1. Introduction

In this article three friendly classroom-based competitions are introduced that are intended to

- encourage students to follow their passions,
- inspire students to think creatively, not just analytically,
- generate interest in Bayesian modeling approaches,
- stimulate insight into complex Bayesian methods, and
- motivate students to pursue excellence in a competition with their peers.

Each competition detailed here will allow students to be ranked based on their performance in the respective competition, allowing prizes to be offered up to the top ranked students. However, it is suggested that the students' final grades not be determined by the rankings, rather their grade would be based on their approach and explanation. This explains why the competitions are referred to here as "friendly" competitions. The rankings and prizes are simply rewards to provide a fun motivation and to some extent sense gamify the classroom learning experience. For classes with many students, these competitions work just the same with students working together in groups and having competitions with the different groups.

Details of each of the three competitions are outlined in the subsequent sections including various examples and approaches students have presented. No detailed solutions are presented in this article as that would defeat the purpose of the competition, but more to the point is that there are many different possible equally valid solutions to each of these competitions. Many different extensions and variations can also easily be incorporated to customize the competitions as appropriate for any given course. Practical suggestions are provided to help with the implementation of these competitions in the classroom, and these tips and resources may also prove helpful with other classroom activities. The author has implemented all three of these competitions in his graduate-level Bayesian statistics course, but these competitions could also be adapted and prepared for advanced-level undergraduate students.

2. Competition #1: James-Stein Estimation

The goal of this competition is to show that shrinkage, by borrowing information across the dataset, can provide substantial improvement in statistical inference. Students may have been exposed to the powerful effects of shrinkage in other contexts such as ridge regression, Lasso regression, or as random effects in mixed models. Shrinkage is a key component of Bayesian analysis and is powerfully exemplified in the empirical Bayes example of the James-Stein estimator (James & Stein, 1961).

The James-Stein estimator became a sensation in the statistical world when Professors Brad Efron and Carl Morris published an article in *Scientific American* (Efron & Morris, 1977) demonstrating how shrinkage with this estimator can yield substantial improvements in prediction performance over the more standard approach of using maximum likelihood estimation. In particular, the sample mean, with its extensive list of advantageous properties, was found to be inferior – inadmissible to be precise – in the multivariate normal setting.

The premise of the James-Stein estimator is succinctly described in Efron (2012) as "learning from the experiences of others" in which individual data values are shrunk towards the grand mean. The optimal amount of shrinkage is a key question to be worked out, but in certain contexts an effective amount of shrinkage can be determined from the data alone.

There are several interesting properties of the James-Stein estimator. For one, shrinkage doesn't have to be towards the overall mean. Rather, it can be towards any arbitrary point, and the resulting James-Stein estimator will still do better than maximum-likelihood estimator (under appropriate assumptions). Also, even if there's "nothing to learn from the others" – that is, one is analyzing completely unrelated variables – the James-Stein estimator still won't be any worse than the maximum-likelihood estimator (again, under appropriate assumptions). A quirky example of this, as provided on Wikipedia, describes how estimating the speed of light, tea consumption in Taiwan, and hog weight in Montana, estimated all together, will be at least as good with the James-Stein estimator as with maximum likelihood estimation (though in this case probably no better).

The example published in the 1977 *Scientific American* article (Efron & Morris, 1977) remains the most popular example of the James-Stein estimator. This example predicted the batting averages of 18 major-league baseball players in the 1970 baseball season using only pre-season batting average data. Efron and Morris showed that the James-Stein estimator provided 3.5 times more accuracy in terms of mean-square prediction error compared to simply using just the pre-season batting average data directly as predictions.

One key notion for the students with regard to James-Stein estimation is that this method relies on *simultaneously* estimating/predicting several quantities. In the baseball example, we were interested in predicting the batting averages for all 18 players simultaneously, and not just evaluating the prediction for only one of the players. Indeed, the sample mean is an admissible estimator for the true mean with univariate and bivariate normally distributed data, but its admissibility is lost for higher dimensions.

There are several topics an instructor may choose to cover prior to this competition, depending on the level of the course and background of the students. Such topics could include loss functions, risk, Bayes risk, Bayes estimator, least favorable prior, minimax estimator, admissible estimator, and the James-Stein estimator.

Now that the stage is set, the actual competition is now introduced as the following student challenge.

The James-Stein Challenge

Find a dataset like the James-Stein baseball dataset in which you actually have two sets of data: pilot data and final/official data. Then make predictions/estimates of the final data using the maximum likelihood estimator and compare those predictions to the James-Stein estimator. Try to determine an appropriate amount of shrinkage in the James-Stein estimator using only the pilot data and not the final data. Calculate quantities such as the mean square prediction error to quantify the improvement (or lack there of) when using the James-Stein estimator. Finally, compare your prediction results against a range of shrinkage levels and determine the amount of shrinkage that appears to be optimal for your given dataset. You are encouraged to **be creative** and **follow your passions** when selecting your dataset.

Each student/group will give a short presentation to the rest of the class demonstrating their results, and all participants will subsequently rank the presentations from most favorite (1) to least favorite (n). The voting method of single transferable vote will be used to determine the top places.

As a side benefit to this competition, students will get to learn about the powerful voting method of single transferable vote that is starting to be adopted by progressive municipalities and other elections across the world. The function `stv` inside the R package `vote` (Sevcikova et al., 2018) provides a convenient implementation of this voting method. Creating a ballot is pretty straightforward, but I found the website `https://getvoting.aec.gov.au/ballotpaper/` to be a convenient and easy way of generating ranked choice voting ballots. Voting by the students can be replaced by a panel of judges, but I tend to like having the students select their own winners. Other students or faculty could also be invited to the classroom on the presentation day to promote enthusiasm for the competition. Example prizes for the winners could be theater tickets or gift cards to a local restaurant. For our competition, I gave out tickets to a production at our local playhouse, so the tickets weren't terribly expensive yet offered a fun and culturally enriching experience for the winning students.

Below are some example datasets that were presented in my classroom. Not all of the examples may be entirely appropriate for James-Stein estimation, but they all demonstrated creative exploration and careful consideration of how shrinkage estimation might be useful for their respective dataset. Even if a student's dataset isn't a great fit for James-Stein estimation, a thoughtful and passionate presentation could still win over the votes and collect a prize.

- **Chinese New Year Gala YouTube Data**

  The Chinese New Year Gala is a popular program broadcast on Chinese New Year's Eve including a variety of performances such as comedy sketches, singing, dancing, acrobatics, and magic shows. It just so happened that our classroom competition occurred shortly after Chinese New Year so this was a very timely topic and one of particular interest for many of our

students that grew up in China. Using the YouTube API and the R package `tuber`, 'like' and 'dislike' counts were extracted at two time points from the approximately 50 different gala performances published on YouTube. James-Stein estimation was applied to improve the prediction of the proportion of likes for each video. As an asside, it was also entertaining to the students to see which performances garnered the greatest numbers of likes.

- **Women's World Championship Volleyball: Faults and Aces**

  Data on the average number of faults and aces were extracted from the Women's Volleyball National League which were then used to predict their performances at the Women's World Championship.

- **Average Temperature Data**

  Average temperature data from 8 locations in the US from 1895-1985 was used to predict the respective average temperatures from 1986-2018.

- **Air Pistol and Air Riffle World Cup Data**

  A series of six qualification scores of the air pistol and air riffle world cups were used to predict the corresponding final competition scores.

- **Men's Singles Badminton: Win Percentages**

  The win percentages of the top singles men badminton players with six months of data was used to predict the win percentages of the same players over the course of the entire year.

- **Comparison of Classifiers**

  In a much different example, one student interested in machine learning considered 10 different classifiers, including nearest neighbors, linear support vector machine, random forest, naive Bayes, and others, on two sets of data – training and testing. James-Stein estimation was used to see if shrinkage could improve prediction of classification accuracy on the test data across all of the classifiers.

3. Competition #2: Traveling the World

This competition builds on the classic tramcar problem, which can be simply introduced with the following example problem.

> Suppose you enter an unfamiliar city and get in a random tramcar that is numbered 1729. Based just on this information, how many such tramcars might you estimate to be in the entire city?

Variants on this problem have had significant applications during World War II. For example, serial numbers extracted from captured tanks were used to estimate the number of tanks being produced by the Germans, as documented in several *JASA* articles shortly after the war ended (Goodman, 1952, 1954; Ruggles &

Brodie, 1947). In an article in *Teaching Statistics* (Johnson, 1994), presented a number of different solutions to this problem. The most straightforward solution to this problem involves constructing a Bayesian posterior from a carefully-formed prior, which is the focus of this competition, but with several complicating twists added. The following set of exercises/examples can be provided to the class to help prepare the class for the competition.

(A) Suppose the citizens of a city are numbered 1 to N and you randomly select a person from the city whose number is m. Write down the likelihood function, $f(m \mid N)$, that describes the probability of observing citizen number m out of a population of N citizens.

(B) Based on the likelihood in (A), what is the maximum likelihood estimator for N?

(C) Show that taking a "uniform prior" on N, i.e. $\pi(N) \propto 1$, leads to an improper posterior.

(D) Consider the prior $\pi(N) \propto \frac{1}{N}$. Show that the posterior distribution is well-defined but the posterior mean does not exist.

(E) With the prior in (D), show analytically the posterior median is approximately $2m$ and derive an approximate formula for the $(1-\alpha)100\%$ highest density interval. Hint: consider integral approximations of the summations.

(F) Consider the prior $\pi(N) = \frac{1}{N^\beta}$ with $\beta \sim \text{unif}(1, \infty)$. The posterior for N, when marginalized over β is equivalent to using an alternative prior $\tilde{\pi}(N)$. Derive $\tilde{\pi}(N)$ and determine if the posterior mean exists. Also, write a program to numerically approximate the posterior distribution of N given m.

(G) (Convergence practice) Consider the two priors $\pi_1(N) = \frac{1}{(\log(N))^2}$ and $\pi_2(N) = \frac{1}{\log(N)\log(\log(N))}$. Show that the posterior distribution of $N \mid m$ exists for $\pi_1(N)$ but not for $\pi_2(N)$.

(H) Suppose k citizens of the city are sampled with replacement, yielding numbers $m_1, \ldots, m_k$. Derive the likelihood in terms of the sufficient statistic $m_{(k)} = \max\{m_1, \ldots, m_k\}$; i.e. compute the probability $\Pr(m_k = x \mid N)$. Hint: notice that $\Pr\left(m_{(k)} = x \mid N\right) = \Pr\left(m_{(k)} \leq x \mid N\right) - \Pr\left(m_{(k)} < x \mid N\right)$.

(I) Calculate the same probability as in (H) under the assumption the citizens are sampled without replacement.

Once the students are familiar with the above calculations, they are prepared to start thinking about the competition, which is now introduced.

Population Estimation Challenge

In this competition, we will virtually travel the world by using `https://randomcity.net` to select random cities across the globe. In particular, we will first randomly select two cities and let the quantities N_1 and N_2 represent the most recent population estimates of those cities on Wikipedia. Without loss of generality, let's assume $N_1 \leq N_2$. Then ten random integers will be sampled with replacement from the interval $[N_1, N_2]$. The goal of this challenge is to most accurately estimate the values N_1 and N_2, say with estimates $\hat{N}_1$ and $\hat{N}_2$, under the following loss function

$$\text{loss} = \frac{|\hat{N}_1 - N_1|}{N_1} + \frac{|\hat{N}_2 - N_2|}{N_2}.$$

The loss will be calculated over three rounds, and the final rankings will be based on minimizing the total loss over all three rounds.

If the entirety of the randomcity.net database is available with associated populations, then unique optimal Bayes estimates can be produced. But the complexities of the likelihood expression, the intricate loss function, as well as the ambiguities of the randomcity.net database provide plenty of challenges for the students to explore. The loss function can also be varied to explore how different loss functions can affect the estimates.

The sampled data can either be provided in advance of the competition or the students can be asked to produce "live" estimates in the classroom. Either way, the students should also have an opportunity to at some point present their solutions to the class.

Of course this competition isn't modeled after any real practical scenario, but it provides a fun way to explore random cities of the world in the context of learning statistics with a friendly competition.

4. Competition # 3: Capture-Recapture

This competition was initially inspired by a recent popular YouTube (Parker, 2018) in which Matt Parker traveled to the 2018 Annual Conference of the Royal Statistical Society to estimate the number of statisticians attending that conference using a capture-recapture design. This YouTube video could serve as an entertaining introduction to the concept of capture-recapture experiments. For an introduction to Bayesian models for capture-recapture designs, Chapter 5 (Capture-Recapture Experiments) of Marin & Robert (2014) is recommended.

A number of complexities have been incorporated into this next competition to allow students to creatively explore different modelling ideas into their solutions. It is quite straightforward to modify the parameters of this competition as needed for a given classroom. Although the other competitions didn't require any special materials, except possibly the prizes, this competition requires several orange and

white ping pong balls (say 30 to 300) to be acquired. Practice ping pong balls can be readily found on the internet in packs of 144 for under $10 so these material costs are not substantial. We now introduce the last competition challenge.

Ball Counting Challenge

A box is prepared containing an unspecified number of orange and white ping pong balls. The orange balls are labelled sequentially starting at some unspecified integer and the white balls start off unlabelled. The ultimate goal will be to predict the number of orange balls, the number of white balls, and the total number of balls (orange and white), given the sampled data.

Students in class will take turns taking out a handful of balls at a time from the box. For each sample, the integer labels on the orange balls are recorded, the unlabelled white balls will be labelled (starting with 1), and any labels already labelled white balls will be recorded. This data will be made available to the students for subsequent modelling.

There are a number of ways the sampled data can be structured and recorded. Below is one example of an actual dataset that was collected in the classroom.

Sampler	Orange Count	White Count	Orange Labels	White Labels
Sample 1	4	2	8,53,56,59	
Sample 2	3	2	20,35,64	1
Sample 3	4	1	3,15,56,58	
Sample 4	4	1	38,39,46,60	
Sample 5	3	3	8,57,62	3,4
Sample 6	4	2	19,28,46,47	
Sample 7	0	4		8
Sample 8	1	4	30	9
Sample 9	2	2	53,64	
Sample 10	3	2	16,35,38	13

In the above example, there were 28 orange balls drawn of which 7 balls (labelled 8, 35, 38, 46, 53, 56, and 64) were drawn twice, and there were 23 white balls drawn of which 6 balls (labelled 1, 3, 4, 8, 9, and 13) were drawn twice. In this particular competition, the orange balls were numbered from -5 to 64 (total of 70) and there were 42 white balls, thus yielding a total of 112 orange and white balls in the box. Random objects may also be added to the box to discourage the use of weight or volume in the inference. For the dataset collection above, sweet potatoes and onions were added to the box as a distraction.

There are a number of features to this competition. First is the multiple-stage capture-recapture component with the white balls. Inference for this part might follow the T-stage capture-recapture model as described in Section 5.2.3 of Marin & Robert (2014) but possibly with a carefully devised informative prior. We also have elements of the population estimation challenge here with regard to the orange ball inference but on a smaller scale and different considerations for the priors. Having the orange and white balls drawn in the same sample also adds an extra layer of complexity. Finally, if the balls are not sufficiently mixed up

between samples, students looking for a competitive edge in their inference may attempt to account for the lack of randomness in the sampling.

5. Conclusions

The first competition presented on James-Stein estimation inspires students to explore data they may be passionate about while also requiring them to consider how the James-Stein methodology may be applicable. Students will also be fully engaged in the presentations by having them vote for their favorite presented application. The use of single transferable vote to select the winners offers another valuable learning opportunity for the students.

The second competition starts off by presenting a Bayesian approach to the classic tramcar/German tank problem with a series of exercises. The competition builds on the ideas of the tramcar problems but with added complexities to be considered in the modelling. Not only will students be curious to see whose model performed best, but they will be able to virtually visit different cities around the world while engaging with this competition.

The third competition, relating to a capture-recapture problem, builds on a fun experiment conducted by Matt Parker in one of his popular YouTube videos. This competition uses physical props in the classroom – numbered ping pong balls in a box – allowing for an interactive classroom data generating experience. Although a complex challenge is presented in this article, it can easily simplified to suit the level of the students.

These are all fun and engaging competitions for students learning Bayesian statistics. With these competitions, students are challenged to think creatively and match their creative ideas with analytical methods. Finally, these competitions can be easily be adapted and tailored to the appropriate level of the students and various classroom sizes.

[Received: July 2020 — Accepted: January 2021]

References

Efron, B. (2012), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press.

Efron, B. & Morris, C. (1977), 'Stein's paradox in statistics', *Scientific American* **236**(5), 119–127.

Goodman, L. A. (1952), 'Serial number analysis', *Journal of the American Statistical Association* **47**(260), 622–634.

Goodman, L. A. (1954), 'Some practical techniques in serial number analysis', *Journal of the American Statistical Association* **49**(265), 97–112.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, pp. 361–379.

Johnson, R. W. (1994), 'Estimating the size of a population', *Teaching Statistics* **16**(2), 50–52.

Marin, J.-M. & Robert, C. P. (2014), *Bayesian essentials with R*, Vol. 48, Springer.

Parker, M. (2018), 'How to estimate a population using statisticians'. https://www.youtube.com/watch?v=MTmnVBJ9gCI

Ruggles, R. & Brodie, H. (1947), 'An empirical approach to economic intelligence in World War II', *Journal of the American Statistical Association* **42**(237), 72–91.

Sevcikova, H., Silverman, B. & Raftery, A. (2018), *vote: Election Vote Counting*, R package version 1.1-0 edn, https://CRAN.R-project.org/package=vote.