INTRODUCING BAYESIAN INFERENCE WITH THE TAXICAB PROBLEM

BERG, Arthur and HAWILA, Nour
Pennsylvania State University, USA
berg@psu.edu

The taxi problem goes by many names in the literature including the Schrödinger problem, the German tank problem, the racing car problem, the horse-racing problem, and the taxicab problem. The basic problem goes like this: Suppose taxicabs in a certain city are numbered 1 to N, and one such taxicab is randomly selected, say number 1729. Based on this information, we wish to infer the total number of taxicabs, N, there are in the city. Rarely do we encounter problems in statistics with just a single data value, but as indicated by the many different names attributed to this problem, this seemingly simple problem can uncover a wealth of statistical inference. We would conclude there are at least 1729 taxicabs in the city, but we want to do better than just a lower bound. In order to do better, we need to introduce more information via reasonable assumptions on the potential number of taxicabs. Bayesian inference can be used to combine the observed data with the additional assumptions into a coherent estimate of N. This paper offers an introduction to Bayesian inference for students as part of an introductory probability and statistics course.

INTRODUCTION

Here we provide a resource for introducing Bayesian statistics to tertiary students as part of an introductory probability and statistics course. The simply stated taxicab problem has a rich history with many well-known statisticians, including M. S. Bartlett, R. Fisher, R. C. Geary, H. Jeffreys, P.-S. Laplace, J. Neyman, C.S. Pierce, E. J. G. Pitman, and E. Schrödinger having been associated with the problem in various contexts. Before we delve into discussing a Bayesian approach to the taxicab problem, we first present some of the rich historical context of this problem including various applications in which the problem arises.

After introducing the different applications, we outline the key approaches to solving the problem. At each step, we explicitly state the assumptions that are being made and students should be challenged to critically evaluate such assumptions in different contexts of the problem. Our first step is to write down the likelihood corresponding to the stated problem. With the likelihood in place, we explore classical non-Bayesian solutions (also referred to as "frequentist solutions") to the problem. We start with the maximum likelihood estimator and show that this standard estimation method is very conservative and underestimates the true value. We will also draw on a stick breaking model to intuitively calculate the expected values used to formulate an approximately unbiased estimator. Technical details are avoided; rather, we focus on key concepts. The article (Johnson, 1994), published in *Teaching Statistics*, is an excellent supplemental resource for this discussion.

What is lacking from the non-Bayesian solutions are intuitive or reasonable prior assumptions based on the context of the problem. For example, we can easily justify upper bounds for the number of taxicabs, and this additional information can be easily integrated into the analysis when a Bayesian approach is followed. Specifically, this information enters through the so-called prior in Bayesian inference. Different priors will be considered for different contexts of the taxicab problem.

Once the likelihood and prior are specified, we employ Bayes' theorem to update the prior based on observed data (the taxicab number that was randomly sampled) to produce the posterior distribution. The posterior distribution provides a probability for each possible value of the true parameter. We then discuss how we might reduce the distribution of values down to a single point estimate with a corresponding credible interval (the Bayesian version of the confidence interval).

A PROBLEM WITH A RICH HISTORY

In a letter written by the prolific American statistician Charles Pierce in 1911, Pierce attributes the problem to Pierre-Simon Laplace (Charles S, 1976):

"One of [Laplace's] problems professes to calculate from the fact that all balls in an urn are numbered 1, 2, 3, etc. and the fact that a ball has been drawn and found to bear a number N, what the probable number of balls in the urn is. But no deductive conclusion on the subject can be drawn from those premisses correctly."

Although Pierce clear attributes Laplace, the authors carefully explored Laplace's extensive writings and various English translations with no success any description of this problem. Laplace indeed analysed numerous ball-and-urn problems, but we simply could not find a description of the problem at hand. A couple decades after Pierce's letter, British statistician H. Jeffreys writes a letter in 1934 to another British statistician R. Fisher attributing the problem to Polish statistician J. Neyman (Fisher, 1990):

"[Neyman] once asked me the following: a man arrives at a railway junction in a town in a foreign country, which he has never heard of before. The first thing he sees is a tramcar numbered 100. Can he infer anything about the number of tramcars in the town? [Neyman] thought the question was significant and so did I, and we both had a feeling that there were probably about 200. I tried it on M.S. Bartlett, who thought it was meaningless but had the same feeling about 200."

Then in a 1944 paper, Irish statistician R. Geary attributes the problem to Nobeloriate E. Schrödinger (Geary, 1944):

"At a recent meeting of the Dublin University Mathematical Society, E. Schrödinger suggested the following ingenious problem as an illustration of Pitman's concept of closeness. In a town, cars are known to be numbered consecutively from 1. The numbers on r of the cars are noted: the problem is to find the closest estimate of the number of cars in the town."

Other variants of this problem have also appeared more recently in the literature:

(Tenenbein, 1971): "A spectator at a race track is observing a car race in which the cars are numbered consecutively from one to some unknown number N. He wishes to estimate the number of cars on the race track after observing that M cars numbered $X_1, X_2, \dots, X_M$ have passed. Each car is equally likely to hold a given position in the race at any given time."

(Rosenberg & Deely, 1976): "Suppose we are at a horse race where we know there are no scratchings (*i.e.*, the number of horses on the track is equal to the highest number on any horse). We take a moving picture of a particular section of the track and stop the film after M horses have passed by. Assuming that it is possible to read the numbers of the horses in the movie, we wish to estimate the number of horses taking part in the race."

Arguably the most significant application of this problem was during the Second World War, in which the serial numbers of captured German tanks were found to be marked sequentially from 1 to N (Ruggles & Brodie, 1947). Applying the same statistical inference that we discuss in this paper led to an estimate of 246 German tanks being produced each month during the war, which is substantially lower than the conventional Allied intelligence estimates indicating a monthly production of 1,400 tanks. After the war, German records validated the statistical analyses by confirming the actual monthly production number to be 245.

This shows that this simple problem can have many different and diverse applications. In the subsequent discussion, we will stick with the formulation introduced in the abstract: taxicab number 1729 is randomly selected among taxicabs numbered 1 to N, and we wish to estimate the value of N. The specific number 1729 is chosen due to its historical significance as the Ramanujan-Hardy taxicab

number (Silverman, 1993). The first task in the inference – Bayesian and non-Bayesian alike – is to write down the likelihood corresponding to the data generating mechanism.

THE LIKELIHOOD

The likelihood simply encodes the probability of observing the data, which we will call M, given the true parameter N. Having observed just one taxicab, the likelihood is simply

$$\Pr(M|N) = \begin{cases} 1/N & if\ M \le N \\ 0 & otherwise. \end{cases}$$

If instead of just one taxicab, we observe k taxicab numbers $M_1, \dots, M_k$ independently sampled from the N total taxicabs (with replacement), then independence of the data allows us to write the likelihood as follows

$$\Pr(M_1, \dots, M_k|N) = \prod_{i=1}^{k} \Pr(M_i|N) = \begin{cases} 1/N^k & if M_{(k)} \le N \\ 0 & otherwise, \end{cases}$$

where $M_{(k)} = max\{M_1, \dots, M_k\}$. In particular, we see that this likelihood only depends on the maximum observed taxicab number, which we write as $M_{(k)}$, thus making $M_{(k)}$ a sufficient statistic.

Note that this likelihood assumes every value between 1 and N is possible and equally likely and that multiple samples are drawn independently with replacement. If the sampling was taken without replacement, say we observe three taxicabs at the same time, then we could modify the likelihood accordingly; see e.g. (Berg, 2021).

NON-BAYESIAN (FREQUENTIST) SOLUTIONS

The maximum likelihood estimator is the value of M, or more generally $M_{(k)}$, that maximizes the likelihood probability. Clearly, $1/N^k$ is a decreasing function in N, and, since the smallest possible value for N is $M_{(k)}$, the maximum likelihood estimator is $M_{(k)}$. However, this estimator is rather unsatisfying as it only reports the lower bound of the possible values. An alternative approach is to estimate N with an unbiased estimator.

In order to construct an unbiased estimator, we first calculate the expected value of $M_{(k)}$ and use that expected value to solve for N. Here, we only present a heuristic calculation of the expected value; a more rigorous calculation can be found in (Johnson, 1994). Let's suppose that instead of sampling k values from the discrete set $\{1, \dots, N\}$, we randomly sample k values from the continuous interval $[0, N]$. In this case, the expected value of $M_{(k)}$ can be intuitively calculated using a stick breaking analogy. If you randomly break a stick of length N at k randomly chosen positions, then the resulting k+1 pieces will have length $\frac{N}{k+1}$ on average (see Figure 1). Using this stick-breaking representation, we can intuitively see the expected value of $M_{(k)}$ is

$$\mathrm{E}\left[M_{(k)}\right] = \mathrm{N} - \frac{N}{k+1} = \mathrm{N}\left(1 - \frac{1}{k+1}\right) = \mathrm{N}\left(\frac{k}{k+1}\right) \quad \text{(continuous case)}$$

Note that the above expression applies when the samples are taken on the continuous interval $[0, N]$. A small correction is applied when sampling from the discrete set $\{1, \dots, N\}$ and depending on whether sampling is done with or without replacement.

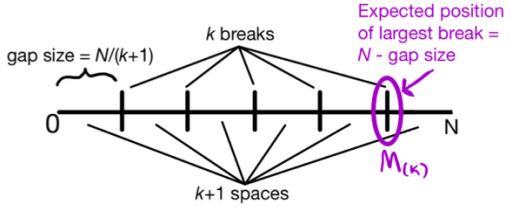

Figure 1: Using the stick breaking analogy to heuristically approximate the expected value of $M_{(k)}$.

Now we construct our approximately unbiased estimate of N by replacing the expectation in the above expression with $M_{(k)}$ and solving for N. This leads to the following approximate unbiased estimator of N:

$$\widehat{N} = M_{(k)}\left(\frac{k+1}{k}\right)$$

So, for having observed just one taxicab numbered M, this unbiased estimator simply doubles M, which is a much less conservative estimator than the maximum likelihood estimator. However, there is more information encoded in the problem than just the value of $M_{(k)}$. We explore this further in the next section.

THE PRIOR

The problem formulation we are focusing on is estimating the number of taxicabs in a certain city. Although we are not told the size of the city, we can describe some broad bounds that would cover most cities. For example, New York City has about 13,600 licenced taxis, London has about 70,600 licenced taxis, and Mexico City, the city with the most taxicabs, has approximately 140,000 taxicabs. Therefore, we could with reasonable certainty conclude the number of taxis in the unspecified city could range from 50 to 140,000. This is certainly a very wide range, but it is still information that could be utilized to augment the statistical inference. In this example, with such a wide range, the additional information would not provide much improvement, but if we knew more about the city, such as its population, more precise bounds could be constructed leading to more precise estimates.

This prior in Bayesian inference encodes the probabilities of each possible value before observing the data. So, if the feasible range for N is in the interval [50, 140000], then we need to specify a probability for each possibility. Lacking any insight as to the potential number of taxis, a natural prior would be the uniform distribution on [50, 140000] in which each value is equally likely. A somewhat more sophisticated approach would be to utilize published reports, such as (Schaller, 2005), that contain data on the number of taxis in a large sampling of cities to approximate a more realistic prior distribution.

Mathematically, we will denote the prior distribution as $\pi(N)$. So, if we take the prior to be a uniform distribution on [50, 140000], then $\pi(N) = 1/(140{,}000 - 50 + 1) = 1/139951$. In this case the prior mean is quite high – close to 70,000 – so instead we modify the prior probabilities to decrease proportionally with N. Specifically, we take $\pi(N) = c/N$, where the constant c is chosen so that the prior sums to one (in this case, $c \approx 0.126$). The prior mean for this "decaying prior" is 17,616, which is still quite high – close to the number of taxis in New York City – but far better than 70,000. It is often the case that different priors are considered to understand how the results vary with the priors.

We finally note that for different applications, such as estimating the number of German tanks or estimating the number horses at a horse race or estimating the number of cars at a racetrack, a different prior would be called for depending on the context of the given application. We would certainly use much smaller numbers when modelling the number of horses or the number of race cars, yet the non-Bayesian estimators presented above do not change according to these substantial differences across the applications.

THE POSTERIOR AND BAYESIAN INFERENCE

Once the prior distribution has been pinned down, the posterior distribution is calculated using Bayes' theorem:

$$\pi\left(N \mid M_{(k)}\right) = \frac{\pi(N)\, Pr\left(M_{(k)} \mid N\right)}{\sum_N \pi(N)\, Pr\left(M_{(k)} \mid N\right)}$$

This seems like a monstrous formula, but it's not so bad; it's basically just the product of the prior $\pi(N)$ with the likelihood $Pr\left(M_{(k)} \mid N\right)$, but then normalized (dividing by its sum) so that it adds up to one.

The approach to calculate $Pr(M_{(k)} \mid N)$ can be found in (Berg, 2021). Here, we will assume just one taxicab was observed ($k = 1$), so we will use the likelihood $Pr(M \mid N)$ above. After multiplying $\pi(N)$ by $Pr(M \mid N)$ and then normalizing the vector so that it sums to one, we obtain the posterior probability of N given M. This is an entire distribution of values for N, but if we wanted to report a single estimate, we would report a central tendency like the mean or the median of the posterior distribution. Additionally, we can summarize the posterior distribution with a $(1 - \alpha)\%$ credible interval by identifying values of N with posterior probabilities that sum to $1 - \alpha$ for a user-specified value of α.

SHINY APPLICATION

Implementing the computations required for the Bayesian analysis may be a barrier to some, so we developed an interactive Shiny application (Chang et al., 2021) to assist with these computations. The Shiny app, accessible at https://glow.shinyapps.io/taxicab/, implements the Bayesian and non-Bayesian estimators of the taxicab problem. The source code is accessible at https://github.com/NourHawila/taxicab. After the user provides the data (e.g., observed taxicab numbers), smallest and largest feasible values of N (parameters $N_{\min}$ and $N_{\max}$), and level α, the application calculates and graphs the posterior distribution for N and displays the maximum likelihood estimate, approximate unbiased estimate, prior mean, posterior mean, posterior median, and the lower and upper bounds of a $(1 - \alpha)\%$ credible interval.

We now return to the originally stated taxicab problem having observed the taxicab number 1729. The estimates of the total number of taxicabs based on the maximum likelihood estimator and the approximate unbiased estimator are 1729 and 3458, respectively. In Table 1 we present the Bayesian for different prior parameters. We see that the Bayesian inference is indeed sensitive to the prior parameters. The more accurate the prior distribution can be specified, the more accurate the Bayesian inference becomes. We also see that the posterior median is consistently smaller than the posterior mean as the posterior distribution is right-skewed.

Table 1: Bayesian analysis of the taxicab problem having observed taxicab number 1729 with four different priors.

Parameters			Bayesian results			
Nmin	Nmax	Prior	Prior Mean	Posterior Mean	Posterior Median	95% Credible Interval Upper Bound
50	140,000	Uniform	70,025	31,518	15,561	112,388
50	140,000	Decaying	17,610	15,610	3,417	28,015
50	10,000	Uniform	5,025	4,762	4,159	9,160
50	10,000	Decaying	1,875	4,208	2,949	8,071

DISCUSSION

This taxicab problem, with its simplistic structure and historical roots, provides an excellent gateway problem for students to be introduced to Bayesian methods. We highlighted several different applications related to this problem, presented non-Bayesian solutions, and detailed a Bayesian approach to solving the problem. The Bayesian solution allows for more information of the problem to be utilized but is also computationally more complex. To facilitate the computation of the Bayesian solution, an accompanying interactive application is described and provided online.

When used in the classroom, additional Bayesian examples and applications that could follow the taxicab problem include (Eadie et al., 2019), which applies Bayesian statistics to modelling the colours of M&M's candies, (Bárcena et al., 2019), which applies Bayesian statistics in finding the sunken nuclear submarine USS Scorpion that sank in 1968, and (Kuindersma & Blais, 2007), which uses

Bayesian statistics in analysing the probability a flipped cylinder (representing a thick coin) comes to rest on its edge.

BIBLIOGRAPHY

Bárcena, M. J., Garín, M. A., Martín, A., Tusell, F., & Unzueta, A. (2019). A Web Simulator to Assist in the Teaching of Bayes' Theorem. *Journal of Statistics Education*, *27*(2), 68–78. https://doi.org/10.1080/10691898.2019.1608875

Berg, A. (2021). Bayesian Modeling Competitions for the Classroom. *Colombian Journal of Statistics*, *in press*, 1–11.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *shiny: Web Application Framework for R* (1.6.0). https://cran.r-project.org/package=shiny

Charles S, P. (1976). The New Elements of Mathematics. In C. Eisele (Ed.), *Mathematical Miscellanea* (Vol. 3, Issue 3). https://doi.org/10.1016/0315-0860(77)90070-2

Eadie, G., Huppenkothen, D., Springford, A., & McCormick, T. (2019). Introducing Bayesian Analysis With m&m's®: An Active-Learning Exercise for Undergraduates. *Journal of Statistics Education*, *27*(2), 60–67. https://doi.org/10.1080/10691898.2019.1604106

Fisher, R. A. (1990). *Statistical Inference and Analysis* (J. Bennett (ed.)). Clarendon Press.

Geary, R. C. (1944). *Comparison of the Concepts of Efficiency and Closeness for Consistent Estimates of a Parameter*. *33*(2), 123–128.

Johnson, R. W. (1994). Estimating the Size of a Population. *Teaching Statistics*, *16*(2), 50–52. https://doi.org/10.1111/j.1467-9639.1994.tb00688.x

Kuindersma, S. R., & Blais, B. S. (2007). Teaching Bayesian Model Comparison with the Three-Sided Coin. *The American Statistician*, *61*(3), 239–244.

Rosenberg, W. J., & Deely, J. J. (1976). The Horse-Racing Problem-A Bayesian Approach. *The American Statistician*. https://doi.org/10.2307/2682883

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.

Schaller, B. (2005). A Regression Model of the Number of Taxicabs in U.S. Cities. *Journal of Public Transportation*, *8*(5), 63–78. https://doi.org/10.5038/2375-0901.8.5.4

Silverman, J. H. (1993). Taxicabs and Sums of Two Cubes. *The American Mathematical Monthly*, *100*(4), 331–340.

Tenenbein, A. (1971). The Racing Car Problem. *The American Statistician*, *25*(1), 38–40.