A Bayesian Treatment of the German Tank Problem

Cory M. Simon

The German tank problem has an interesting historical background and is an engaging problem of statistical estimation for the classroom. The objective is to estimate the size of a population of tanks inscribed with sequential serial numbers, from a random sample. In this tutorial article, we outline the Bayesian approach to the German tank problem, whose solution assigns a probability to each tank population size, thereby quantifying uncertainty, and which provides an opportunity to incorporate prior information and/or beliefs about the tank population size into the solution. We illustrate with an example. Finally, we survey problems in other contexts that resemble the German tank problem.

Background

To inform their military strategy during World War II (1939–1945), the Allies sought to estimate Germany's rate of production and capacity of various types of military equipment (tanks, tires, rockets, etc.). Conventional methods to estimate armament production, including extrapolating data on prewar manufacturing capabilities, obtaining reports from secret sources, and interrogating prisoners of war, were mostly unreliable or contradictory.

In 1943, British and American economic intelligence agencies exploited a German manufacturing practice in order to statistically estimate their armament production. Specifically, Germany marked their military equipment with serial numbers as well as codes for the date and/or place of manufacture. Their intention was to facilitate the handling of spare parts and to trace defective equipment and parts back to the manufacturer for quality control. However, these serial numbers and codes on a captured sample of German equipment conveyed information to the Allies about Germany's production.

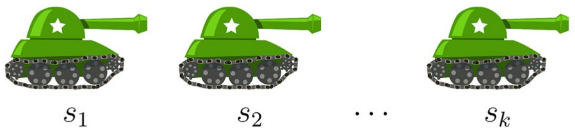
To estimate Germany's rate of production of tanks, the Allies collected serial numbers on the chassis, engines, gearboxes, and bogie wheels of samples of tanks by inspecting captured tanks and examining captured records.[1] Despite lacking an exhaustive sample, the sequential nature of these serial numbers and patterns in the samples enabled the Allies to estimate Germany's tank production. Postwar research has shown that serial number analysis gave more accurate estimates than the overestimates produced by conventional intelligence methods (Table 1).[2] See Richard Ruggles and Henry Brodie's article [44] for a detailed historical account of the serial number analysis used to estimate German armament production during World War II.

The German Tank Problem

Simplification of the historical context in which German tank production was estimated via serial number analysis [44] motivated the formulation of the textbook-friendly *German tank problem* [21]:

Problem statement. During World War II, the German military is equipped with n tanks. Each tank is inscribed with a unique serial number in the set $\{1, \dots, n\}$.

As the Allies, we do not know n, but we have captured (without replacement, of course) a sample of k German tanks with (ordered) inscribed serial numbers $(s_1, \dots, s_k)$.



Assuming that every tank in the population was equally likely to be captured and n is fixed, our objective is to estimate n in light of the data $(s_1, \dots, s_k)$.

In 1942, in a crowded restaurant in Washington, D.C., Alan Turing and Andrew Gleason discussed a variant of the German tank problem: "how to estimate the total number of taxicabs in a town after having seen a random selection of their license numbers" [13, 24]. Today, with its interesting historical background [44], the German tank problem is still a suitable dinner conversation topic and serves as an intellectually engaging, challenging, and enjoyable problem to illustrate combinatorics and statistical estimation in the classroom [3, 15, 27, 33].

[1] For example, captured records from tank repair depots listed serial numbers of the chassis and engine of repaired tanks, and records from divisional headquarters listed chassis serial numbers of tanks held by a specific unit.

[2] Gearboxes on captured tanks, for example, were inscribed with serial numbers belonging to an unbroken sequence. Chassis serial numbers, on the other hand, were broken into blocks to distinguish models/designs, leaving gaps between the serial numbers assigned to them.

© The Author(s) 2023
https://doi.org/10.1007/s00283-023-10274-6

Table 1. Monthly production rate of tanks by Germany [44].

Month	Estimates		German records
	Conventional American & British Intelligence	Serial number analysis	
June, 1940	1000	169	122
June, 1941	1550	244	271
August, 1942	1550	327	342

Uncertainty quantification. Any estimate of the tank population size n from the data $(s_1, \dots, s_k)$ is subject to uncertainty, since we (presumably) have not captured all of the tanks (i.e., $k \neq n$, probably). Quantifying uncertainty in our estimate of n is important because high-stakes military decisions may be made on the basis of it.

Our contribution. In this pedagogical article, we outline the Bayesian approach to the German tank problem, whose solution assigns a probability to each tank population size, thereby quantifying uncertainty, and which provides an opportunity to incorporate prior information and/or beliefs about the tank population size into the solution.

Survey of Previous Work on the German Tank Problem

The frequentist approach. Kim Border [7] calls the German tank problem a "weird case" in frequentist estimation. The maximum likelihood estimator of the tank population size n is the maximum serial number observed among the k captured tanks, $m^{(k)} := \max_{i \in \{1,\dots,k\}} s_i$. This is a biased estimator, since certainly $m^{(k)} \leq n$.

Leo Goodman [21, 22] derives the minimum-variance unbiased estimator of the tank population size

$$\hat{n} = m^{(k)} + \left(\frac{m^{(k)}}{k} - 1 \right). \tag{1}$$

To intuit $\hat{n}$, note that n must be greater than or equal to $m^{(k)}$, and if we observe large (small) gaps between the serial numbers $(s_1, \dots, s_k)$ after sorting them (including the gap preceding the smallest serial number), then n is likely (unlikely) to be much greater than $m^{(k)}$. The estimator of n in (1) quantifies how far beyond the maximum serial number $m^{(k)}$ we should estimate the tank population size, based on the gaps; $m^{(k)}/k - 1$ is the average size of the gaps. Goodman [21] also derives a frequentist two-sided $1 - a$ confidence interval $m^{(k)} \leq n \leq x$ for n, where x is the greatest integer satisfying $\left(m^{(k)} - 1\right)_k / (x)_k \geq a$ (the notation $(n)_k$ for the falling factorial is defined in (5)).

Use in pedagogy. Julian Champkin [23] highlights the application of statistics to estimate German tank production during WWII as a "great moment in statistics." Roger Johnson [27] lists and evaluates several intuitive point estimators for the size of the tank population. Richard Scheaffer et al. [45] propose a hands-on learning activity to illustrate the German tank problem by sampling chips labeled with numbers from 1 to n from a bowl. Inspired by the German tank problem, Arthur Berg [3] orchestrates a classroom-based competition to best estimate the size of a population of a city from a random sample. George Clark, Alex Gonye, and Steven J. Miller [10] explore the use of simulations of tank capturing and linear regression to discover the estimator in (1).

The Bayesian approach. Closely related to our pedagogical exploration of the Bayesian approach to the German tank problem, Harry Roberts [41], Michael Höhle, and Leonhard Held [25], Wolfgang Von der Linden, Volker Dose, and Udo Von Toussaint [49], and Simona Cocco, Rémi Monasson, and Francesco Zamponi [11] undertake a Bayesian analysis of the German tank problem and provide an analytical formula for the mean and variance of the posterior distribution of the tank population size under an improper uniform prior distribution. Mark Andrews [1] outlines the Bayesian approach to the German tank problem in a blog post containing code in the R language. William Rosenberg and John Deely [43] outline an empirical Bayesian approach to estimate the number of horses in a race from a sample of numbered horses (the likelihood function here is equivalent to that in the German tank problem). Arthur Berg and Nour Hawila [4] use Bayesian inference for the closely related taxicab problem.

Generalizations and variants. Goodman [21, 22] and Clark, Gonye, and Miller [10] pose a variant of the German tank problem in which the initial serial number is not known; i.e., the n tanks are inscribed with serial numbers $\{b + 1, \dots, n + b\}$ with b and n unknown. Lee and Miller [31] generalize the German tank problem to the settings in which the serial numbers belong to a continuum and/or lie in two or more dimensions within a square or circle.

Overview of the Bayesian Approach to the German Tank Problem

Adopting a Bayesian perspective [6, 15, 46], we treat the (unknown) total number of tanks as a discrete random variable N to model our uncertainty about it. A probability mass function of N assigns a probability to each possible tank population size n. This probability is a measure of our degree of belief, perhaps with some basis in knowledge and data, that the tank population size is n [20]. The spread of the mass function of N over the integers reflects uncertainty.

The observed serial numbers $(s_1, \dots, s_k)$ convey information about the tank population size. Hence, the probability mass function of N changes after the data $(s_1, \dots, s_k)$ are collected and considered. That is, N has a prior and a posterior probability mass function.

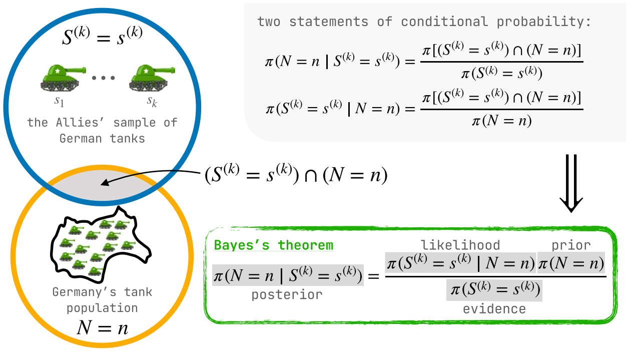

$S^{(k)} = s^{(k)}$
s_1
s_k
the Allies' sample of German tanks
Germany's tank population
$N = n$
$(S^{(k)} = s^{(k)}) \cap (N = n)$
two statements of conditional probability:
$\pi(N = n \mid S^{(k)} = s^{(k)}) = \frac{\pi[(S^{(k)} = s^{(k)}) \cap (N = n)]}{\pi(S^{(k)} = s^{(k)})}$
$\pi(S^{(k)} = s^{(k)} \mid N = n) = \frac{\pi[(S^{(k)} = s^{(k)}) \cap (N = n)]}{\pi(N = n)}$
$\Downarrow$
Bayes's theorem
likelihood
prior
$\pi(N = n \mid S^{(k)} = s^{(k)}) = \frac{\pi(S^{(k)} = s^{(k)} \mid N = n)\,\pi(N = n)}{\pi(S^{(k)} = s^{(k)})}$
posterior
evidence

Figure 1. Bayes's theorem applied to the German tank problem. An Euler diagram [32, 37] represents the two events $S^{(k)} = s^{(k)}$ and $N = n$ with circles. The area of each circle is proportional to the probability of the event, and the area of overlap is proportional to the probability of the intersection $(S^{(k)} = s^{(k)}) \cap (N = n)$ of the events. The Euler diagram rationalizes the two statements of conditional probability in terms of the intersection of the events, which in turn imply Bayes's theorem [30].

The three inputs to a Bayesian treatment of the German tank problem are as follows:

1. The prior mass function of N, which expresses a combination of our subjective beliefs and objective knowledge about the tank population size before we collect and consider the sample of serial numbers.
2. The data, namely the sample of serial numbers $(s_1, \dots, s_k)$, viewed as realizations of random variables $(S_1, \dots, S_k)$ owing to the stochasticity of tank-capturing.
3. The likelihood function, giving the probability of the data $(S_1, \dots, S_k) = (s_1, \dots, s_k)$ under each tank population size $N = n$, based on a probabilistic model of the tank-capturing process.

The output of a Bayesian treatment of the German tank problem is the posterior mass function of N, conditioned on the data $(s_1, \dots, s_k)$. The posterior follows from Bayes's theorem and can be viewed as an update to the prior in light of the data, as illustrated by Figure 1. The posterior mass function of N assigns to each possible tank population size n a probability according to a compromise between its likelihood, which invokes the probabilistic tank-capturing model to quantify the support lent by the observed serial numbers $(s_1, \dots, s_k)$ for the hypothesis that the tank population size is n, and its prior probability, which quantifies how likely we thought the tank population size was n before the serial numbers $(s_1, \dots, s_k)$ were collected and considered [46]. The posterior mass function of N is the raw Bayesian solution to the German tank problem; its spread quantifies our posterior uncertainty about N. We may summarize the posterior by reporting its median and a small subset of the integers on which most of the posterior mass sits—a *credible set* that likely contains the tank population size. Furthermore, from the posterior, we can answer questions such as, what is the probability that N exceeds some threshold quantity n' that would alter military strategy?

Table 2. List of parameters/variables.

Parameter/variable	$\in$	Description
n	$\mathbb{N}_{\geq 0}$	Size of population of tanks
k	$\mathbb{N}_{>0}$	Number of captured tanks
s_i	$\mathbb{N}_{>0}$	Serial number on captured tank i
$s^{(k)}$	$\mathbb{N}_{>0}^k$	Vector listing the serial numbers on the k captured tanks
$m^{(k)}$	$\mathbb{N}_{>0}$	Maximum serial number among the k captured tanks

The Bayesian Approach to the German Tank Problem

We now delve into the details of the Bayesian approach to the German tank problem and illustrate via an example. For reference, the variables are listed in Table 2. We use uppercase letters to represent random variables and lowercase letters to represent their realizations. Throughout, we employ the indicator function associated with a set A:

$$\mathcal{I}_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases} \tag{2}$$

The Prior Distribution

We construct the prior probability mass function $\pi_{\text{prior}}(N = n)$ to express a combination of our subjective beliefs and objective knowledge about the total number of tanks N before the data $(s_1, \dots, s_k)$ are collected and considered.

The prior mass function that we impose on N depends on the context. If we do not possess prior information about the tank population size, we may adopt the principle of indifference and impose a diffuse prior, e.g., a uniform distribution over a set of feasible tank population sizes. On the other hand, if we possess a rough estimate of the number of tanks from some other source of information or analysis, we may construct a more informative prior that concentrates its mass around this estimate. By definition, a diffuse prior admits more uncertainty (measured, for example, by entropy [35]) about the tank population size than a more informative prior [46].

Thinking ahead about the posterior mass function of N, which balances the prior and the likelihood (the latter based on the data), a more informative prior will have a larger impact on the posterior than a diffuse one [46], which "lets the data speak for itself" [15], and generally, as the number of captured tanks k increases, we expect the prior to have a smaller impact on the posterior [15] as the data "overwhelms" the prior.

The Data, Data-Generating Process, and Likelihood Function

The data. The data we obtain in the German tank problem is the vector

$$s^{(k)} := (s_1, \dots, s_k) \tag{3}$$

of serial numbers inscribed on the k captured tanks. We view the data $s^{(k)}$ as a realization of the discrete random vector $S^{(k)} := (S_1, \dots, S_k)$. At this point, we are entertaining the possibility that the order in which tanks are captured matters.

The data-generating process. The stochastic data-generating process consists in the sequential capture of k tanks from a population of n tanks, without replacement, and then inspecting their serial numbers to construct $s^{(k)}$. We assume that each tank in the population is equally likely to be captured at each step. Then mathematically, the stochastic data-generating process is a sequential uniform random selection of k integers, without replacement, from the set $\{1, \dots, n\}$.

The likelihood function. The likelihood function specifies the probability of the data $S^{(k)} = s^{(k)}$ given each tank population size $N = n$. Each outcome $s^{(k)}$ in the sample space $\Omega_n^{(k)}$ is equally likely, where

$$\Omega_n^{(k)} := \{(s_1, \dots, s_k)_{\neq} : s_i \in \{1, \dots, n\} \text{ for all } i \in \{1, \dots, k\}\}, \tag{4}$$

with $(\cdots)_{\neq}$ meaning that the elements of the vector $(\cdots)$ are unique. The number of outcomes $|\Omega_n^{(k)}|$ in the sample space is the number of distinct ordered arrangements of k distinct integers from the set $\{1, \dots, n\}$, given by the falling factorial:

$$(n)_k := n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}. \tag{5}$$

Under the data-generating process, then, the probability of observing data $S^{(k)} = s^{(k)}$ given the tank population size $N = n$ is the uniform distribution:

$$\pi_{\text{likelihood}}\left(S^{(k)} = s^{(k)} \mid N = n\right) = \frac{1}{(n)_k} \mathcal{I}_{\Omega_n^{(k)}}\left(s^{(k)}\right). \tag{6}$$

Interpretation. The likelihood quantifies the support provided by the serial numbers on the k captured tanks in $s^{(k)}$, when compared with our probabilistic model of the tank-capturing process, for the hypothesis that the tank population size is n [46]. We view $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$ as a function of n, since in practice we possess the data $s^{(k)}$ but not n.

The likelihood as a sequence of events. Alternatively, we may arrive at (6) from a perspective of sequential events $S_1 = s_1, S_2 = s_2, \dots, S_k = s_k$. First, the probability of a given serial number on the ith captured tank, conditioned on the tank population size and the serial numbers on the previously captured tanks, is the uniform distribution

$$\pi(S_i = s_i \mid N = n, S_1 = s_1, \dots, S_{i-1} = s_{i-1}) = \frac{1}{n-i+1} \mathcal{I}_{\{1,\dots,n\} \setminus \{s_1,\dots,s_{i-1}\}}(s_i), \tag{7}$$

since there are $n - (i-1)$ tanks to choose from, uniformly and randomly. By the chain rule of probability [29], the joint probability is

$$\pi_{\text{likelihood}}(S_1 = s_1, \dots, S_k = s_k \mid N = n) = \prod_{i=1}^{k} \pi(S_i = s_i \mid N = n, S_1 = s_1, \dots, S_{i-1} = s_{i-1}), \tag{8}$$

which gives (6) after simplifying the product of indicator functions.

The likelihood function in terms of the maximum observed serial number. We will find out below that only two independent features of the data $(s_1, \dots, s_k)$ provide information about the tank population size N: its size, k, and the maximum observed serial number

$$m^{(k)} = \max_{i \in \{1,\dots,k\}} s_i. \tag{9}$$

Thus, we also write a different likelihood: the probability $\pi_{\text{likelihood}}(M^{(k)} = m^{(k)} \mid N = n)$ of observing a maximum serial number $m^{(k)}$ given the tank population size $N = n$.

Because each outcome $s^{(k)} \in \Omega_n^{(k)}$ is equally likely, $\pi_{\text{likelihood}}(M^{(k)} = m^{(k)} \mid N = n)$ is the fraction of the sample space $\Omega_n^{(k)}$ in which the maximum serial number is $m^{(k)}$. To count the outcomes $s^{(k)} \in \Omega_n^{(k)}$ where the maximum serial number is $m^{(k)}$, consider that one of the k captured tanks has serial number $m^{(k)}$ and the remaining $k-1$ tanks have a serial number in $\{1, \dots, m^{(k)} - 1\}$. For each of the k possible positions of the maximum serial number in the vector $s^{(k)}$, there are $(m^{(k)} - 1)_{k-1}$ distinct outcomes specifying the other $k-1$ entries. Thus

$$\pi_{\text{likelihood}}(M^{(k)} = m^{(k)} \mid N = n) = \frac{k(m^{(k)} - 1)_{k-1}}{(n)_k} \mathcal{I}_{\{k,\dots,n\}}(m^{(k)}). \tag{10}$$

The Posterior Distribution

The posterior probability mass function of N assigns a probability to each possible tank population size n in consideration of its consistency with the data $(s_1, \dots, s_k)$, according to the likelihood in (6), and our prior beliefs/knowledge encoded in $\pi_{\text{prior}}(N = n)$.

The posterior distribution is a conditional distribution related to the likelihood and prior mass functions by Bayes's theorem [30] (see Figure 1):

$$\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)}) = \frac{\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)\pi_{\text{prior}}(N = n)}{\pi_{\text{evidence}}(S^{(k)} = s^{(k)})}. \tag{11}$$

The denominator, the *evidence* [30], is the probability of the data $s^{(k)}$:

$$\pi_{\text{evidence}}(S^{(k)} = s^{(k)}) = \sum_{n'=0}^{\infty} \pi_{\text{likelihood}}\left(S^{(k)} = s^{(k)} \mid N = n'\right)\pi_{\text{prior}}(N = n'). \tag{12}$$

We view $\pi_{\text{posterior}}\left(N = n \mid S^{(k)} = s^{(k)}\right)$ as a probability mass function of N, since in practice, we have $s^{(k)}$. Then $\pi_{\text{evidence}}\left(S^{(k)} = s^{(k)}\right)$, which is independent of n, is just a normalizing factor for the numerator in (11).

In interpreting (11), the prior mass function of N is updated, in light of the data $(s_1, \dots, s_k)$, to yield the posterior mass function of N. The posterior probability that $N = n$ is proportional to the product of the likelihood at and prior probability of $N = n$, giving a compromise between the likelihood and prior.

We simplify the posterior mass function of N in (11) by substituting (6), restricting the sum in (12) to tank population sizes where the likelihood is nonzero, and noting that the only two features of the data $(s_1, \dots, s_k)$ that appear are its size k and the maximum serial number $m^{(k)}$:

$$\begin{aligned}\pi_{\text{posterior}}&(N = n \mid S^{(k)} = s^{(k)}) \\ &= \pi_{\text{posterior}}(N = n \mid M^{(k)} = m^{(k)}) \\ &= \frac{(n)_k^{-1}\pi_{\text{prior}}(N = n)}{\sum_{n'=m^{(k)}}^{\infty} (n')_k^{-1}\pi_{\text{prior}}(N = n')}\mathcal{I}_{\{m^{(k)}, m^{(k)}+1, \dots\}}(n).\end{aligned} \tag{13}$$

Note, we may arrive at (13) through (10) as well.

Interpretation. The posterior probability mass function of N in (13) assigns a probability to each tank population size n in consideration of the serial numbers $(s_1, \dots, s_k)$ observed on the captured tanks, our probabilistic model of the tank-capturing process, and our prior beliefs and knowledge about the tank population size expressed in the prior mass function of N. The spread (measured, e.g., by entropy) of the posterior mass function of N reflects remaining epistemic (reducible with more data) [17, 47] uncertainty about the tank population size.

A remark on "uncertainty." The source of posterior uncertainty is a lack of complete data: we have not captured all of the tanks[3] and observed their serial numbers to be certain of the tank population size. In practice, an additional source of posterior uncertainty about the tank population size is the possible inadequacy of the model of the tank-capturing process (uniform sampling) in (6). That is, selection bias could be present in the tank-capturing process. Our analysis here neglects this source of uncertainty.

Summarizing the posterior mass function of N . We may summarize the posterior mass function of N with a point estimate of the tank population size and a credible subset of the integers that contains the tank population size with a high probability.[4] A suitable point estimate of the tank population size is a median of the posterior mass function of N; by definition, the posterior probability that the tank population size is greater (less) than or equal to a median is at least 0.5. A suitable credible subset, which entertains multiple tank population sizes, is the α-high-mass subset [26]

$$\mathcal{H}_\alpha := \left\{n' : \pi_{\text{posterior}}\left(N = n' \mid M^{(k)} = m^{(k)}\right) \geq \pi_\alpha\right\}, \tag{14}$$

where π_α is the largest mass to satisfy

$$\pi_{\text{posterior}}\left(N \in \mathcal{H}_\alpha \mid M^{(k)} = m^{(k)}\right) \geq 1 - \alpha. \tag{15}$$

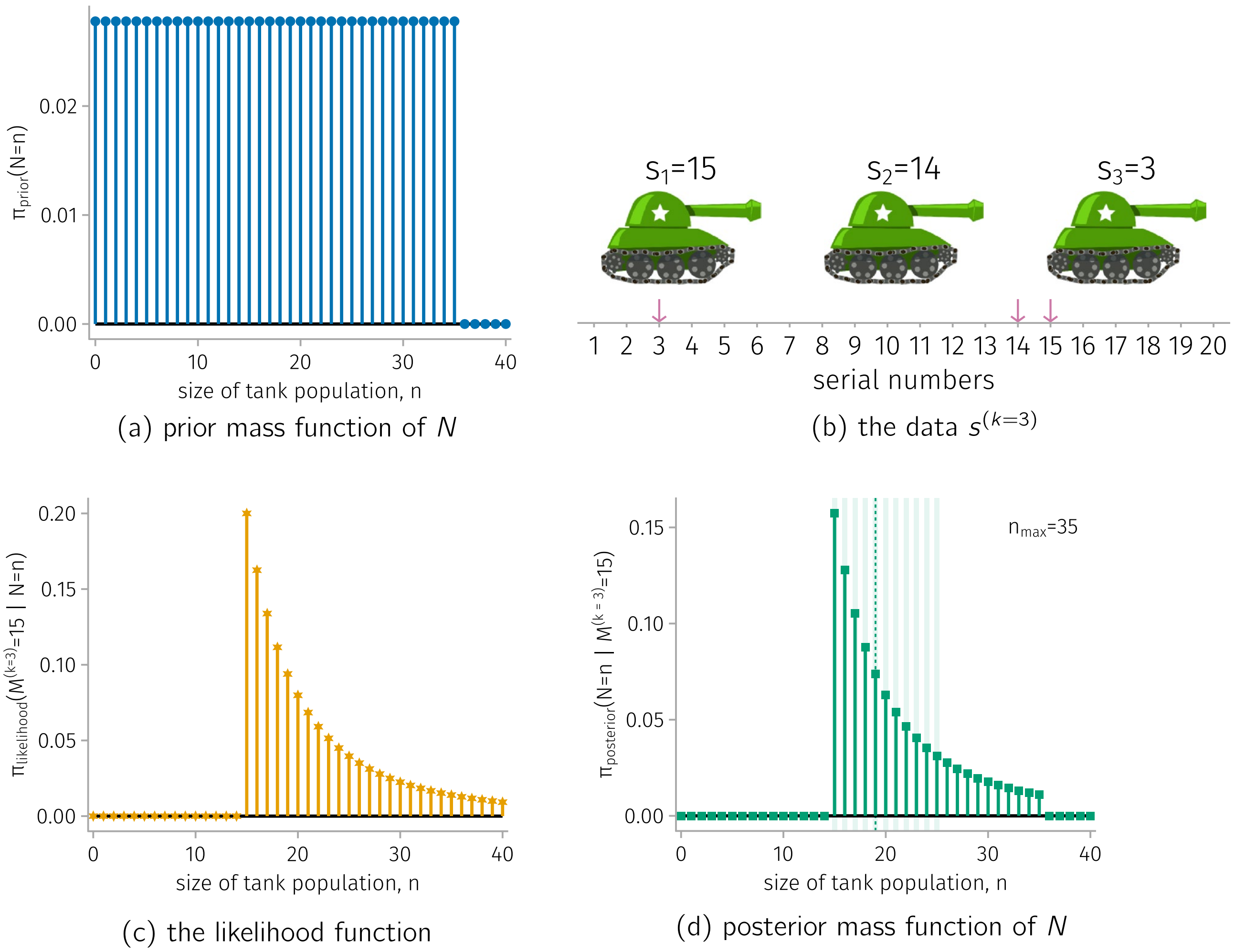
In words, the α-high-mass subset $\mathcal{H}_\alpha$ is the smallest that contains at least a fraction $1 - \alpha$ of the posterior mass of N and ensures that every tank population size belonging to it is more probable than any population size outside of it.

Querying the posterior distribution. We may find the posterior probability that the tank population size belongs to any set of interest by summing the posterior mass over it; e.g., the probability that the tank population size exceeds some number n' is

$$\pi_{\text{posterior}}\left(N > n' \mid M^{(k)} = m^{(k)}\right) = \sum_{n=n'+1}^{\infty} \pi_{\text{posterior}}\left(N = n \mid M^{(k)} = m^{(k)}\right). \tag{16}$$

[3] Certainly, $k < n$ if there are gaps in the observed serial numbers $(s_1, \dots, s_k)$. Even if there are no gaps in $(s_1, \dots, s_k)$, we cannot be certain we have captured the tank with the largest serial number.

[4] Under our assumptions embedded in the likelihood and prior mass functions.



(a) prior mass function of N

(b) the data $s^{(k=3)}$

(c) the likelihood function

(d) posterior mass function of N

Figure 2. A Bayesian approach to the German tank problem. (a) The prior mass function. (b) The data $s^{(3)}$, with maximum observed serial number $m^{(3)} = 15$. (c) The likelihood function associated with the data $s^{(3)}$. (d) The posterior mass function of N; $\mathcal{H}_{0.2}$ is highlighted, and the median is marked with a vertical dashed line.

An Example

We illustrate the Bayesian approach to the German tank problem through an example.

The prior probability mass function of N. Suppose we have an upper bound n_{max} for the possible number of tanks, based on, e.g., the supply of some raw material needed for tank production, but no other information. Then we may impose a diffuse prior, a uniform prior probability mass function

$$\pi_{\text{prior}}(N = n) = \frac{1}{n_{\text{max}} + 1}\mathcal{I}_{\{0,\dots,n_{\text{max}}\}}(n). \tag{17}$$

This prior mass function expresses that in the absence of any data $(s_1, \dots, s_k)$ (i.e., no serial numbers, and not even k), we believe that the total number of tanks N is equally likely to be any value in $\{0, \dots, n_{\text{max}}\}$. Particularly, suppose $n_{\text{max}} = 35$. Figure 2a visualizes $\pi_{\text{prior}}(N = n)$.

The data $(s_1, \dots, s_k)$ and the likelihood function. Now suppose we capture $k = 3$ tanks, with serial numbers $s^{(3)} = (15, 14, 3)$. See Figure 2b. So the maximum observed serial number is $m^{(3)} = 15$. The likelihood function $\pi_{\text{likelihood}}(M^{(3)} = 15 \mid N = n)$ in (10) is displayed in Figure 2c. Note that the likelihood function is maximal at $n = m^{(3)} = 15$ and decreases monotonically.

The posterior probability mass function of N. Under the uniform prior in (17), the posterior probability mass function of N in (13) becomes

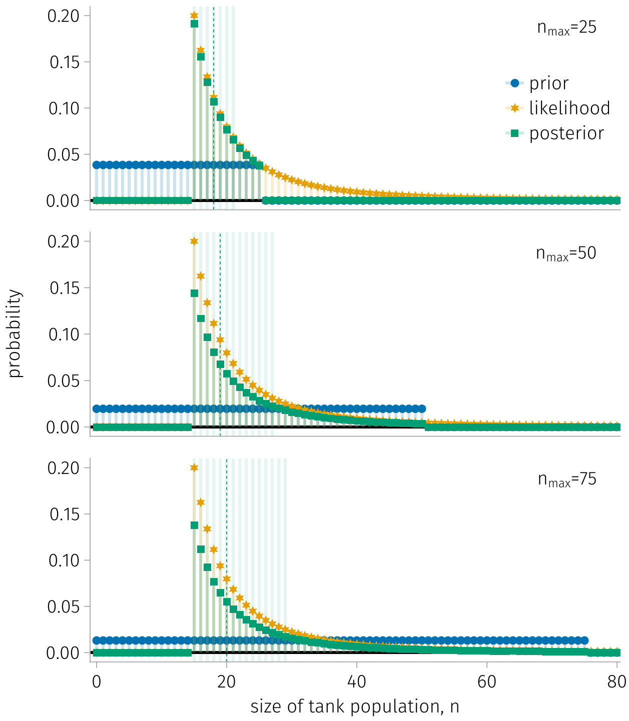


Figure 3. Evaluating the sensitivity of the posterior mass function of N to the upper bound $n_{\max}$ imposed by the prior mass function of N.

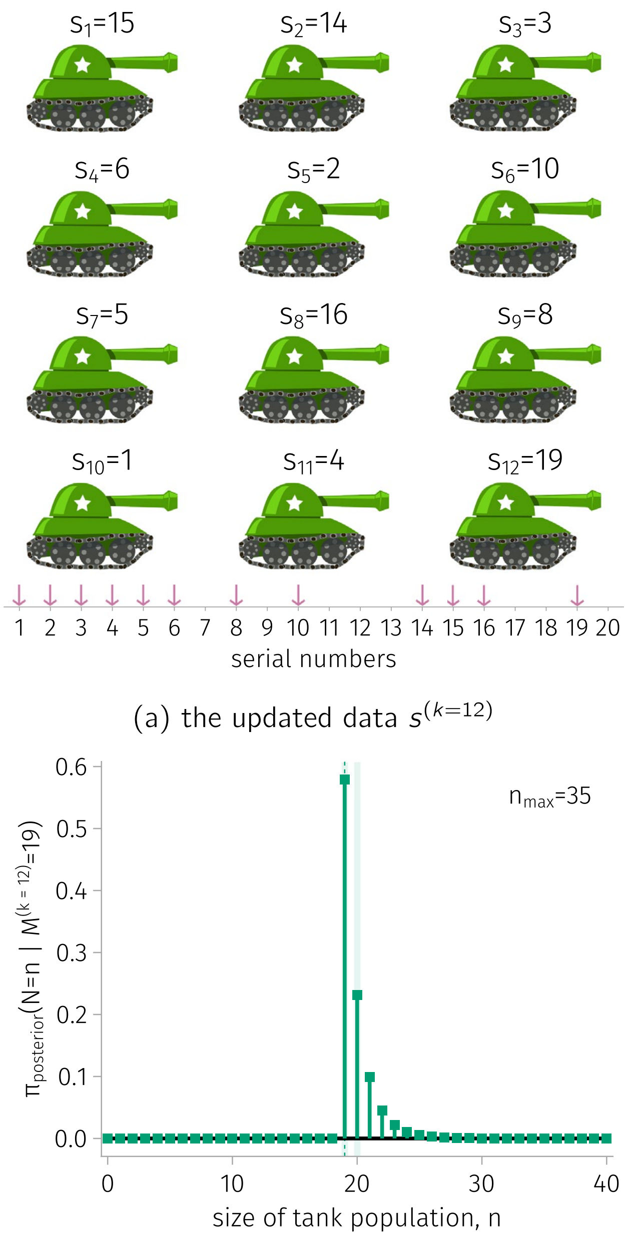
$$\pi_{\text{posterior}}\left(N = n \mid M^{(k)} = m^{(k)}\right) = \frac{(n)_k^{-1}}{\sum\limits_{n'=m^{(k)}}^{n_{\max}} (n')_k^{-1}} \mathcal{I}_{\{m^{(k)},m^{(k)}+1,\ldots,n_{\max}\}}(n). \tag{18}$$

Figure 2d visualizes the posterior probability mass function of N for the data $s^{(3)}$ in Figure 2b and the prior in (17) ($n_{\max} = 35$).

Summarizing the posterior. The posterior mass function of N has median 19 and high-mass credible subset $\mathcal{H}_{0.2} = \{15, \ldots, 25\}$ (highlighted in Figure 2d). For what it's worth, the data in Figure 2b were generated from a tank population size of $n = 20$ (explaining the choice of scale in Figure 2b).

Querying the posterior. Suppose our military strategy would change if the size of the tank population were to exceed 30. From the posterior distribution of N, we calculate $\pi_{\text{posterior}}(N > 30 \mid M^{(3)} = 15) \approx 0.066$.

Sensitivity of the posterior to the prior. Because of the subjectivity involved in constructing the prior, checking the sensitivity of the posterior to the prior is good practice [46]. Figure 3 shows how the posterior mass function of N changes with the upper bound on the tank population $n_{\max}$ that we impose via the prior mass function of N in (17). For example, under $n_{\max} = 75$, the high-mass subset $\mathcal{H}_{0.2}$ expands to $\{15, \ldots, 29\}$.



(a) the updated data $s^{(k=12)}$

(b) the updated posterior mass function of N

Figure 4. The posterior distribution of N after we capture more tanks. (a) We capture an additional nine tanks. (b) The updated posterior mass function of N.

Capturing more tanks. Suppose we capture an additional nine tanks and rerun the Bayesian analysis. Figure 4 shows

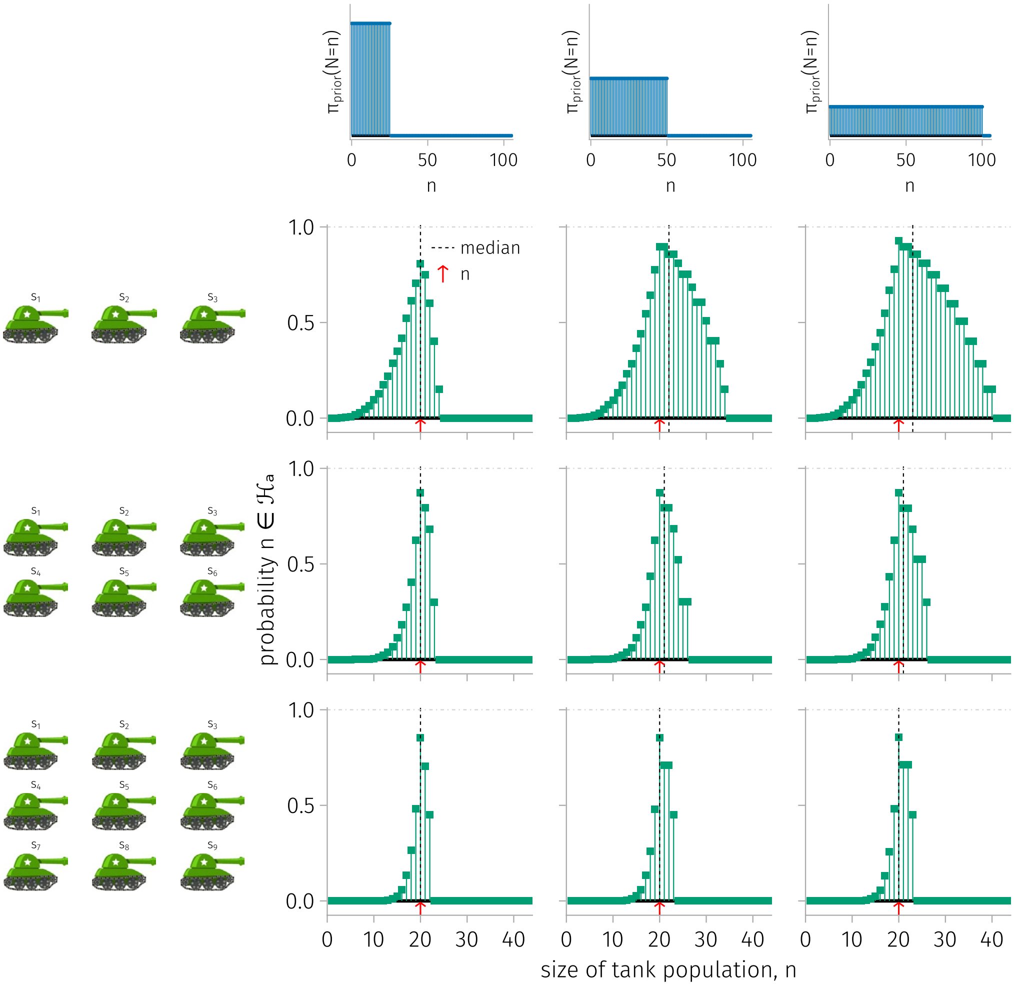


Figure 5. The average high-mass subset and median of the median of the posterior over tank-capturing under a fixed tank population size. Rows: different numbers k of tanks captured (left). Columns: different maxima of tank population sizes entertained by the uniform prior $n_{\max}$ (top). For a particular $(k, n_{\max})$, the stem plots show the probability of each tank population size belonging to the high-mass subset $\mathcal{H}_{a=0.2}$ of the posterior. The vertical dashed line shows the median of the median of the posterior. The red arrow shows the true tank population size, 20.

the updated posterior mass function of N. The high-mass credible subset $\mathcal{H}_{0.2}$ shrinks considerably, to $\{19, 20\}$. This shows how more data—a larger number k of tanks captured—generally reduces our uncertainty about the tank population size.

Simulations to Investigate the Behavior of the Posterior of N Under a Known Population Size

We now investigate how, on average, over the stochastic outcomes of the tank-capturing process for a fixed tank population size, the posterior distribution $\pi_{\text{posterior}}\left(N = n \mid S^{(k)} = s^{(k)}\right)$ depends on the number k of tanks captured and the maximum $n_{\max}$ of the support of the uniform prior.

For a given k and $n_{\max}$, we conduct 50,000 simulations, in each of which k random tanks are captured from a population of $n = 20$ tanks, giving data $s^{(k)}$; computing the posterior mass function $\pi_{\text{posterior}}\left(N = n \mid S^{(k)} = s^{(k)}\right)$; then finding the high-mass subset $\mathcal{H}_a$ $(a = 0.2)$ and median of the posterior. Figure 5 displays (1, stems), the probability of each tank population size n belonging to $\mathcal{H}_a$, and (2, vertical line), the median of the median of the posterior, for $(k, n_{\max}) \in \{3, 6, 9\} \times \{25, 50, 100\}$.

As k increases, the high-mass subset $\mathcal{H}_a$ tends to be less sensitive to $n_{\max}$, since the data overrides the prior, and to shrink, since uncertainty decreases with a larger sample. As $n_{\max}$ increases, larger population sizes become more likely to be included in $\mathcal{H}_a$. The median of the median of the posterior matches the true tank population size of 20 when $n_{\max} = 25$ or $k = 9$. For $k \in \{3, 6\}$, the larger $n_{\max}$ values pull the median above the true tank population size.

Discussion

Selection bias. A strict assumption in the textbook-friendly German tank problem, which enables us to estimate the size of the population of tanks from a random sample of their (sequential) serial numbers, is that sampling is uniform. To check consistency of the sample with this model of the tank-capturing process, Goodman [22] demonstrates a test of the hypothesis that the sample of serial numbers is from a uniform distribution. Interesting extensions of the textbook German tank problem could involve modeling selection bias in the tank-capturing process. For example, such bias could arise hypothetically if older tanks with smaller serial numbers were more likely to be deployed in the fronts opened earlier in the war, where capturing tanks is more difficult than at less fortified fronts opened more recently. Selection bias could also manifest in clusters in the observed serial numbers.

The German tank problem in other contexts. The Bayesian probability theory used to solve the German tank problem applies (perhaps with modification) to many other contexts in which we wish to estimate the size of some finite hidden set [9], such as the number of taxicabs in a city [19, 23], racing cars on a track [48], accounts at a bank [25], pieces of furniture purchased by a university [22], aircraft operations at an airport [34], cases in court [50], or electronic devices produced by a company [2]. And also the extent of leaked classified government communications [18], the time needed to complete a project deadline [16], the time-coverage of historical records of extreme events like floods [39], the length of a short-tandem repeat allele [51], the size of a social network [28], the lifetime of a flower of a plant [38], or the duration of existence of a species [42]. In addition, mark and recapture methods in ecology to estimate the size of an animal population [8, 36] are tangentially related to the German tank problem.

The practice of inscribing sequential serial numbers on military equipment. Germany adopted the practice of marking their military equipment with serial numbers and codes to trace the equipment/parts/components back to the manufacturer. However, the sequential nature of those serial numbers was exploited by the Allies to estimate their armament production. To reduce vulnerability to serial number analysis for estimating production while maintaining the advantages of tracing equipment back to the manufacturer, serial numbers and codes could instead be encrypted [14] or obfuscated, for instance by the method known as chaffing [40].

Acknowledgments

This work has been supported by the United States Department of Homeland Security Countering Weapons of Mass Destruction under CWMD Academic Research Initiative Cooperative Agreement #21CWDARI00043. This support does not constitute an expressed or implied endorsement on the part of the government. Thanks to Bernhard Konrad and Edward Celarier for providing detailed feedback on the manuscript, to my students Gbenga Fabusola, Adrian Henle, and Paul Morris for feedback on the introduction, the anonymous reviewer for providing the idea for the section "Simulations to Investigate the Behavior of the Posterior of N Under a Known Population Size," and to Paul Campbell for informing me about the example of iPhone estimation.

Data and Code Availability The Julia [5] code to reproduce all of our visualizations drawn using `Makie.jl` [12] is available on Github at https://www.github.com/SimonEnsemble/the_German_tank_problem.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

[1] Mark Andrews. German tank problem: a Bayesian analysis. Available at https://www.mjandrews.org/blog/germantank. Accessed 2022-12-03.

[2] Charles Arthur. Why iPhones are just like German tanks. Available at https://www.theguardian.com/technology/blog/2008/oct/08/iphone.apple, 2008.

[3] Arthur Berg. Bayesian modeling competitions for the classroom. *Revista Colombiana de Estadística* 44:2 (2021), 243–252.

[4] Arthur Berg and Nour Hawila. Introducing Bayesian inference with the taxicab problem. In *Proceedings of the Tenth Australian Conference on Teaching Statistics*, pp. 55–60, 2021.

[5] Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. arXiv:1209.5145, 2012.

[6] William M. Bolstad and James M. Curran. *Introduction to Bayesian Statistics*. John Wiley & Sons, 2016.

[7] Kim C. Border. Lecture 18: Estimation. Available at https://healy.econ.ohio-state.edu/kcb/Ma103/ (2021 version), 2017.

[8] Anne Chao. An overview of closed capture–recapture models. *Journal of Agricultural, Biological, and Environmental Statistics* 6:2 (2001), 158–175.

[9] Si Cheng, Daniel J. Eck, and Forrest W. Crawford. Estimating the size of a hidden finite set: large-sample behavior of estimators. *Statistics Surveys* 14 (2020), 1–31.

[10] George Clark, Alex Gonye, and Steven J. Miller. Lessons from the German tank problem. *Mathematical Intelligencer* 43:4 (2021), 19–28.

[11] Simona Cocco, Rémi Monasson, and Francesco Zamponi. *From Statistical Physics to Data-Driven Modelling, with Applications to Quantitative Biology*. Oxford University Press, 2022.

[12] Simon Danisch and Julius Krumbiegel. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software* 6:65 (2021), 3349.

[13] Peter Donovan. Alan Turing, Marshall Hall, and the alignment of WW2 Japanese naval intercepts. *Notices of the AMS* 61:3, 2014.

[14] Hans Delfs, Helmut Knebl, and Helmut Knebl. *Introduction to Cryptography*, volume 2. Springer, 2002.

[15] Allen B. Downey. Think Bayes 2. Available at https://allendowney.github.io/ThinkBayes2/index.html, 2021.

[16] Thomas M. Fehlmann and Eberhard Kranich. A new approach for continuously monitoring project deadlines in software development. In *Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement*, pp. 161–169, 2017.

[17] Craig R. Fox and Gülden Ülkümen. Distinguishing two dimensions of uncertainty. Chapter 1 of *Perspectives on Thinking, Judging, and Decision Making*, 2011.

[18] Michael Gill and Arthur Spirling. Estimating the severity of the WikiLeaks US diplomatic cables disclosure. *Political Analysis* 23:2 (2015), 299–305.

[19] John Goebel and Dan Teague. How many taxis? *Consortium for Mathematics and Its Applications* 72, 1999.

[20] Jayanta K Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian Analysis: Theory and Methods*. Springer, 2006.

[21] Leo A Goodman. Serial number analysis. *Journal of the American Statistical Association* 47:260 (1952), 622–634.

[22] Leo A Goodman. Some practical techniques in serial number analysis. *Journal of the American Statistical Association* 49:265 (2954), 97–112.

[23] Carlos Gómez Grajalez, Eileen Magnello, Robert Woods, and Julian Champkin. Great moments in statistics. *Significance* 10:6 (2013), 21–28.

[24] Andrew Hodges. Alan Turing: the enigma. In *Alan Turing: The Enigma*. Princeton University Press, 2014.

[25] Michael Höhle and Leonhard Held. Bayesian estimation of the size of a population. Technical Report 499, LMU Munich, Discussion Paper, 2006.

[26] Rob J Hyndman. Computing and graphing highest density regions. *American Statistician* 50:2 (1996), 120–126.

[27] Roger W. Johnson. Estimating the size of a population. *Teaching Statistics* 16:2 (1994), 50–52.

[28] Liran Katzir, Edo Liberty, and Oren Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 597–606, 2011.

[29] Karl-Rudolf Koch. *Introduction to Bayesian Statistics*. Springer, 2007.

[30] John Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[31] Anthony Lee and Steven J. Miller. Generalizing the German tank problem. *PUMP Journal of Undergraduate Research* (2023), 59–95.

[32] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics* 18:12 (2012), 2536–2545.

[33] Frederick Mosteller. *Fifty Challenging Problems in Probability with Solutions*. Courier Corporation, 1987.

[34] John H Mott, Margaret L. McNamara, and Darcy M. Bullock. Estimation of aircraft operations at airports using nontraditional statistical approaches. In *2016 IEEE Aerospace Conference*, pp. 1–11. IEEE, 2016.

[35] Kevin P Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[36] James D. Nichols. Capture–recapture models. *BioScience* 42:2 (1992), 94–102.

[37] Alvitta Ottley, Blossom Metevier, P. K. Han, and Remco Chang. Visually communicating Bayesian statistics to laypersons. In *Technical Report*. Tufts University, 2012.

[38] William D. Pearse, Charles C. Davis, et al. A statistical estimator for determining the limits of contemporary and historic phenology. *Nature Ecology & Evolution* 1:12 (2017), 1876–1882.

[39] Ilaria Prosdocimi. German tanks and historical records: the estimation of the time coverage of ungauged extreme events. *Stochastic Environmental Research and Risk Assessment* 32:3 (2018), 607–622.

[40] Ronald L. Rivest et al. Chaffing and winnowing: Confidentiality without encryption. *CryptoBytes (RSA Laboratories)* 4:1 (1998), 12–17.

[41] Harry V. Roberts. Informative stopping rules and inferences about population size. *Journal of the American Statistical Association* 62:319 (1967), 763–775.

[42] David L. Roberts and Andrew R. Solow. When did the dodo become extinct? *Nature* 426:6964 (2003), 245.

[43] W. J. Rosenberg and J. J. Deely. The horse-racing problem, a Bayesian approach. *American Statistician* 30:1 (1976), 26–29.

[44] Richard Ruggles and Henry Brodie. An empirical approach to economic intelligence in World War II. *Journal of the American Statistical Association* 42:237 (1947), 72–91.

[45] Richard L Scheaffer, Ann Watkins, Mrudulla Gnanadesikan, and Jeffrey Witmer. *Activity-Based Statistics: Student Guide*. Springer, 2013.

[46] Rens van de Schoot, Sarah Depaoli, Ruth King, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers* 1:1 (2021), 1–26.

[47] Robin Senge, Stefan Bösner, Krzysztof Dembczyński, et al. Reliable classification: learning classifiers that

distinguish aleatoric and epistemic uncertainty. *Information Sciences* 255 (2014), 16–29.

[48] Aaron Tenenbein. The racing car problem. *American Statistician* 25:1 (1971), 38–40.

[49] Wolfgang Von der Linden, Volker Dose, and Udo Von Toussaint. *Bayesian Probability Theory: Applications in the Physical Sciences*. Cambridge University Press, 2014.

[50] Xiaohan Wu, Margaret E. Roberts, Rachel E. Stern, Benjamin L. Liebman, et al. Augmenting serialized bureaucratic data: the case of Chinese courts. *21st Century China Center Research* 11, 2022.

[51] Haibao Tang, Ewen F Kirkness, et al. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *American Journal of Human Genetics* 101:5 (2017), 700–715.

Cory M. Simon, School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR, USA. E-mail: cory.simon@oregonstate.edu

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.