Lessons from the German Tank Problem

George Clark, Alex Gonye, and Steven J. Miller

In this paper we revisit a famous and historically important problem that has become a staple in many probability and statistics classes: the German tank problem. This case study illustrates that one does not need to use the most advanced mathematics to have a tremendous impact on real-world problems; the challenge is frequently in creatively using what one knows.[1]

By the end of 1941, most of continental Europe had fallen to Nazi Germany and the other Axis powers, and by 1942, their forces had begun their significant advance in the eastern front deep into the Soviet Union; see Figure 1 for an illustration of their rapid progress, or https://www.youtube.com/watch?v=WOVEy1tCnk for an animation of territorial gains day by day. A key component of their rapid conquests was their revolutionary use of tanks in modern warfare. While other militaries, most notably that of France, used tanks as a modern armored form of cavalry, the Germans were the first to make full use of tanks' speed and strength. Tanks would move rapidly and punch through enemy lines, creating gaps that German infantry would stream through. Once through the holes in the line, the Germans would wreak havoc on lines of communication, creating logistical nightmares for the combatants left on the front lines. This lightning-fast warfare has been dubbed blitzkrieg (or lightning war).

With the Nazis utilizing tanks with such devastating results, it was essential for the Allies to stop them. A key component of the solution was figuring out how many tanks the Germans were building, or had deployed in various theaters, in order to allocate resources effectively. As expected, they tried espionage (both with agents and through decrypting intercepted messages) to estimate these numbers. It was essential that the Allies obtain accurate values, since there was a tremendous danger both in underestimating and in overestimating the enemy's strength. The consequence of underestimating is clear, since one could suddenly be outnumbered in battle. Overestimating is also bad, since it can lead to undue caution and failure to exploit advantages, or to committing too many resources in one theater and thus not having enough elsewhere. The United States Civil War provides an example of these consequences. General George McClellan did not allow his Union army take the field against the Confederates, because in his estimation, his forces were greatly outnumbered, though in fact, they were not. This situation led to one of President Abraham Lincoln's many great quips: "If General McClellan isn't going to use his army, I'd like to borrow it for a time."[2] Considering how close Pickett's charge came to succeeding at Gettysburg, or what would have happened if Sherman hadn't taken Atlanta before the 1864 elections (where McClellan, now as the Democratic Party's nominee for president, was running against Lincoln on a platform of a negotiated peace), the paralysis from incorrect analysis could have changed the outcome of the war.

Returning to World War II and the problem of determining the number of tanks produced and facing them in the field, the Allies understandably wanted to have a way to evaluate the effectiveness of their estimates. They realized that the tanks destroyed and captured during a battle had serial numbers on their gearboxes that could help with this problem. Assuming that the numbers were sequential and began with number 1, one could try to estimate the largest value given a string of observed numbers. This discovery contributed to the birth of the statistical method and the use of a population maximum formula, and it changed both the war and science.[3]

[1] Another example is the famous Battle of Midway and the role that cryptographers played in figuring out the Japanese target; see, for example, [1].

[2] There are many different phrasings of this remark; this one is taken from https://thehistoriansmanifesto.wordpress.com/2013/05/13/best-abraham-lincoln-quotes/.

https://doi.org/10.1007/s00283-021-10082-w

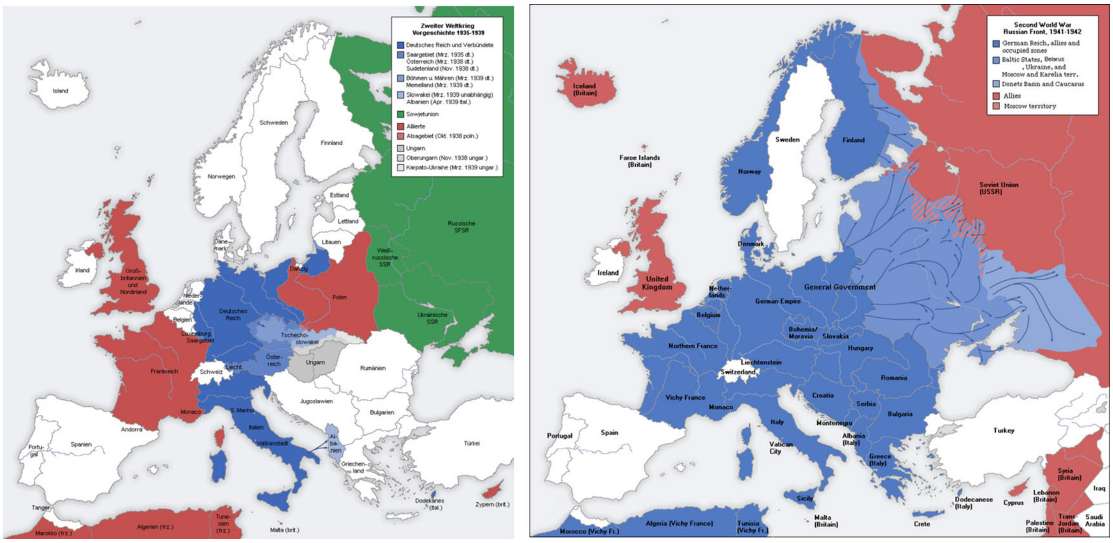

Figure 1. Left: Europe before the start of major hostilities. Right: Europe in 1942. (Images courtesy of Wikimedia Commons from author San Jose.)

The original variant of the problem assumes that the first tank is numbered 1, that there is an unknown number N produced (or in theater), that the numbers are consecutive, and that k values are observed, with the largest being m. Given this information, the goal is to find $\widehat{N}$, the best estimate of N. The formula that was derived is

$$\widehat{N} = m\left(1+\frac{1}{k}\right) - 1.$$

For the reader unfamiliar with this topic, we deliberately do not state here in the introduction how well this formula does versus what was accomplished by espionage.[4] While this story has been well told before (see, for example, [3–6]), our contribution is to extend the analysis to consider the more general case, namely what happens when we do not know the smallest value. To our knowledge, this result has not been isolated in the literature; we shall derive that if s is the spread between the smallest and the largest of the observed k serial numbers, then

$$\widehat{N} = s\left(1+\frac{2}{k-1}\right) - 1.$$

In the last section of our paper, "The German Tank Problem and Linear Regression," we use regression to show that these are reasonable formulas, and thus the German tank problem can also be used to introduce some problems and subtleties in regression analysis, as well as serve as an introduction to mathematical modeling.

Since it is rare to have a clean closed-form expression such as those above, we briefly remark on our fortune. The key observation is that we have a combinatorial problem in which certain binomial identities are available, and these lead to tremendous simplifications.

Derivation with a Known Minimum

In this section we prove that

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

when we observe k tanks, the largest labeled m, knowing that the smallest number is 1 and the tanks are consecutively numbered. Before proving this, as a smell test we look at some extreme cases. First, we never obtain an estimate that is less than the largest observed number. Second, if there are many tanks and we observe just one (so $k = 1$), then $\widehat{N}$ is approximately $2m$. This is very reasonable, and essentially just means that if we have only one data point, it is a good guess that it was in the middle. Further, as k increases, the amount by which we must inflate our observed maximum value decreases. For example, when $k = 2$, we inflate m by approximately a factor of $3/2$, or in other words, our observed maximum value is probably about two-thirds of the true value. Finally, if k equals the number of tanks N, then m must also equal N, and the formula simplifies to $\widehat{N} = N$.

We break the proof into two parts. While we are fortunate in that we are able to obtain a closed-form expression, if we have a good guess as to the relationship, we can use statistics to test its reasonableness; we do that in the last section. For the proof, we first determine the probability that the observed largest value is m. Next we compute the expected value, and then show how to pass from that to an estimate for N. We need two combinatorial results.

[3] There are advantages in consecutive labeling; for example, it simplifies some maintenance issues, since it is clear which of two tanks is older.

[4] That said, since this paper is appearing in a mathematics journal and not a bulletin of a spy agency, we invite the reader to conjecture which method did better.

The first is Pascal's identity:

$$\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}, \tag{1}$$

which can be written as

$$\binom{n+1}{r} = \binom{1}{0}\binom{n}{r} + \binom{1}{1}\binom{n}{r-1},$$

since $\binom{1}{0} = \binom{1}{1} = 1$. The proof is standard.[5]

The second identity involves sums of binomial coefficients and follows immediately by induction and Pascal's identity:

$$\sum_{m=k}^{N} \binom{m}{k} = \binom{N+1}{k+1}. \tag{2}$$

While this identity suffices for the original formulation of the German tank problem, when we do not know the starting serial number, the combinatorics become slightly more involved, and we need a straightforward generalization:

$$\sum_{\ell=a}^{b} \binom{\ell}{a} = \binom{b+1}{a+1}; \tag{3}$$

the proof follows similarly.

The Probability That the Sample Maximum Is *m*

Let M be the random variable for the maximum number observed, and let m be the maximum number that we see. Again, k is the number of tanks observed, and the total number of tanks is N. Note that there is zero probability of observing a value smaller than k or larger than N. We claim for $k \leq m \leq N$ that

$$\text{Prob}(M = m) = \frac{\binom{m}{k} - \binom{m-1}{k}}{\binom{N}{k}} = \frac{\binom{m-1}{k-1}}{\binom{N}{k}}.$$

We give two proofs. The first is to note that there are $\binom{N}{k}$ ways to choose k numbers from N when order does not matter. The probability that the largest observed value is exactly m equals the probability that the largest value is at most m minus the probability that the largest value is at most $m-1$. The first probability is just $\binom{m}{k}/\binom{N}{k}$, since if the largest value is at most m, then all k observed numbers must be taken from $\{1, 2, \ldots, m\}$. A similar argument gives that the second probability is $\binom{m-1}{k}/\binom{N}{k}$, and the claim now follows on using Pascal's identity to simplify the difference of the binomial coefficients.

We could also argue as follows. If the largest value is m, then we have to choose that serial number, and now we must choose $k-1$ tanks from the $m-1$ smaller values; thus we find that the probability is just $\binom{m-1}{k-1}/\binom{N}{k}$.

Remark 1. Interestingly, we can use the two equivalent arguments above to prove Pascal's identity.

The Best Guess for $\widehat{N}$

We now compute the best guess for N by first finding the expected value of M. Recall that the expected value of a random variable M is the sum of all the possible values of M times the probability of observing that value. We write $\mathbb{E}[M]$ for this quantity, and thus we must compute

$$\mathbb{E}[M] := \sum_{m=k}^{N} m \,\text{Prob}(M = m)$$

(note that we need to worry about m only in this range, since for all other m, the probability is zero and thus does not contribute). Once we find a formula for $\mathbb{E}[M]$, we will convert that to one for the expected number of tanks.

Our first step is to substitute in the probability that M equals m, obtaining

$$\mathbb{E}[M] = \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}}.$$

Fortunately, this sum can be simplified into a nice closed-form expression; it is this simplification that allows us to obtain a simple formula for $\widehat{N}$. We expand the binomial coefficients in the expression for $\mathbb{E}[M]$ and then use our second combinatorial identity, (2), to simplify the sum of $\binom{m}{k}$ that emerges. After some straightforward algebra, we obtain

$$\mathbb{E}[M] = \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}} = \frac{k(N+1)}{k+1}.$$

Since we have such a clean expression, it is trivial to solve for N in terms of k and $\mathbb{E}[M]$:

$$N = \mathbb{E}[M]\left(1 + \frac{1}{k}\right) - 1.$$

Thus if we substitute in m (our observed value for M) as our best guess for $\mathbb{E}[M]$, we obtain our estimate for the number of tanks produced:

$$\widehat{N} = m\left(1 + \frac{1}{k}\right) - 1,$$

completing the proof.

Remark 2. A more advanced analysis can prove additional results about our estimator, for example, whether it is unbiased.

Remark 3. There are many ways to see that this formula is reasonable. The first is to try extreme cases, such as $k = N$ (which forces m to be N and gives N as the answer), or to

[5] Interpret the sides as two different ways of counting how many ways we can choose r people from a group of $n+1$, where exactly n of the candidate people are in one set and exactly one person is in another set.

try $k=1$, in which case we expect our one observation to be around $N/2$, and thus a formula that gives doubling the single observation as the best guess is logical. We can also get close to this formula by trying to guess the functional form (for more details, see the last section of this paper). We know that our best guess must be at least m, so let us write it as $m+f(m,k)$. For a fixed k, we might expect our guess to increase as m increases, while for fixed m, we would expect a smaller boost as k increases. These heuristics suggest that $f(m, k)$ increases with m and decreases with k; the simplest such function is bm/k for some constant b. This leads to a guess of $m+bm/k$, and again looking at extreme cases, we get very close to the correct formula.

Derivation with an Unknown Minimum

Not surprisingly, when we do not know the lowest serial number, the resulting algebra becomes more involved; fortunately, though, with a bit of work we are still able to get nice closed-form expressions for the needed sums and obtain again a clean answer for the estimated number of tanks. We still assume that the tanks are numbered sequentially and focus on the spread (the difference between the largest and smallest observed values). Similar to the previous section, we derive a formula to inflate the observed spread to be a good estimate of the number of total tanks.

We first set some notation:

- the minimum tank serial number, N_1,
- the maximum tank serial number, N_2,
- the total number of tanks, N ($N = N_2 - N_1 + 1$),
- the observed minimum value, m_1 (with corresponding random variable M_1),
- the observed maximum value, m_2 (with corresponding random variable M_2),
- the observed spread s (with corresponding random variable S).

Since $s = m_2 - m_1$, in the arguments below we can focus on just s and S. We will prove that the best guess is $s\left(1+\frac{2}{k-1}\right)-1$.

Remark 4. There are two differences between this formula and that obtained in the case that the smallest serial number is known. The first is that we divide by $k-1$ and not k; however, since we cannot estimate a spread with one observation, this is reasonable. Note the similarity here with the sample standard deviation, where we divide by one less than the number of observations; while one point suffices to estimate a mean, we need at least two for the variance. The second difference is that we have a factor of 2, which can be interpreted as movement in both directions.

The Probability That the Spread Is s

We claim that if we observe k tanks, then for $k-1 \le s \le N_2 - N_1$, we have

$$\begin{aligned}\mathrm{Prob}(S=s) &= \frac{\sum_{m=N_1}^{N_2-s}\binom{s-1}{k-2}}{\binom{N_2-N_1+1}{k}} = \frac{(N_2-N_1+1-s)\binom{s-1}{k-2}}{\binom{N_2-N_1+1}{k}}\\ &= \frac{(N-s)\binom{s-1}{k-2}}{\binom{N}{k}},\end{aligned}$$

and for all other s, the probability is zero.

To see this, note that the spread s must be at least $k-1$ (since we have k observations) and cannot be larger than $N_2 - N_1$. If we want a spread of s, then if the smallest observed value is m, the largest is $m+s$. We must choose exactly $k-2$ of the $s-1$ numbers in the set $\{m+1, m+2, \ldots, m+s-1\}$; there are $\binom{s-1}{k-2}$ ways to do so. This proves the first equality, the sum over m. Since all the summands are the same, we get the second equality, and the third follows from our definition of N.

The Best Guess for $\widehat{N}$

We argue similarly as in the previous section, again using our second binomial identity, (2). Relabeling the parameters yields

$$\sum_{\ell=a}^{b}\binom{\ell}{a} = \binom{b+1}{a+1}. \tag{4}$$

We begin by computing the expected value of the spread. We include all the details of the algebra; the idea is to manipulate the expressions and pull out terms that are independent of the summation variable, and then rewrite expressions so that we can identify binomial coefficients and then apply our combinatorial results. We have

$$\begin{aligned}\mathbb{E}[S] &= \sum_{s=k-1}^{N-1} s\mathrm{Prob}(S=s)\\ &= \sum_{s=k-1}^{N-1} s\frac{(N-s)\binom{s-1}{k-2}}{\binom{N}{k}}\\ &= \binom{N}{k}^{-1} N \sum_{s=k-1}^{N-1}\frac{s!(k-1)}{(s-k+1)!(k-1)!}\\ &\quad - \binom{N}{k}^{-1}\sum_{s=k-1}^{N-1}\frac{ss!(k-1)}{(s-k+1)!(k-1)!}\\ &=: T_1 - T_2.\end{aligned}$$

We first simplify T_1. Below, we always try to multiply by 1 in such a way that we can combine ratios of factorials into binomial coefficients:

$$\begin{aligned}T_1 &= \binom{N}{k}^{-1} N \sum_{s=k-1}^{N-1} \frac{s!(k-1)}{(s-k+1)!(k-1)!}\\ &= \binom{N}{k}^{-1} N(k-1) \sum_{s=k-1}^{N-1} \binom{s}{k-1}\\ &= \binom{N}{k}^{-1} N(k-1)\binom{N}{k} \;=\; N(k-1),\end{aligned}$$

where we used (4) with $a = k-1$ and $b = N-1$.

Turning to T_2, we argue similarly, replacing s with $(s-1)+1$ in the second line to assist in collecting factors into a binomial coefficient:

$$\begin{aligned}T_2 &= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{ss!(k-1)}{(s-k+1)!(k-1)!}\\ &= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{(s+1-1)s!(k-1)}{(s-(k-1))!(k-1)!}\\ &= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} k(k-1)\binom{s+1}{k}\\ &\quad -\binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} (k-1)\binom{s}{k-1}\\ &=: T_{21} + T_{22}.\end{aligned}$$

We can immediately evaluate T_{22} using (3) with $a = k-1$ and $b = N-1$, and we obtain

$$T_{22} = \binom{N}{k}^{-1}(k-1)\binom{N}{k} = k-1.$$

Thus all that remains is to analyze T_{21}:

$$T_{21} = \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \binom{s+1}{k} k(k-1).$$

We pull $k(k-1)$ outside the sum, and letting $w = s+1$, we see that

$$T_{21} = \binom{N}{k}^{-1} k(k-1) \sum_{w=k}^{N} \binom{w}{k},$$

and then from (3) with $a = k$ and $b = N$, we obtain

$$T_{21} = \binom{N}{k}^{-1} k(k-1) \sum_{w=k}^{N} \binom{w}{k} = \binom{N}{k}^{-1} k(k-1)\binom{N+1}{k+1}.$$

Substituting everything back yields

$$\mathbb{E}[S] = N(k-1) + (k-1) - \binom{N}{k}^{-1} k(k-1)\binom{N+1}{k+1}.$$

Simplifying the right-hand side yields

$$(N+1)(k-1) - k(k-1)\frac{\frac{(N+1)!}{(N-k)!(k+1)!}}{\frac{N!}{(N-k)!k!}} = \frac{k-1}{k+1},$$

and thus we obtain

$$\mathbb{E}[S] = (N+1)\frac{k-1}{k+1}.$$

The analysis is completed as before, where we pass from our observation of s for S to a prediction $\widehat{N}$ for N:

$$\widehat{N} = \frac{k+1}{k-1}s - 1 = s\left(1 + \frac{2}{k-1}\right) - 1,$$

where the final equality is due to rewriting the algebra to mirror more closely the formula from the case in which the first tank is numbered 1. Note that this formula passes the same smell checks the other did; for example, $s\frac{2}{k-1} - 1$ is always at least 1 (recall that k is at least 2), and thus the lowest estimate we can get for the number of tanks is $s+1$.

Comparison of Approaches

So, which did better, statistics or spies? Once the Allies had won the war, they could look into the records of Albert Speer, the Nazi minister of armaments, to see the exact number of tanks produced each month; see Table 1.

The meticulous German record-keeping comes in handy for the vindication of the statisticians; their estimates were astoundingly more accurate. While certainly not perfect (an underestimation of 30 tanks could have dire consequences when the high command is allocating resources), the statistical analysis was vastly superior to the intelligence estimates, which were off by factors of five or more. We mentioned earlier the lessons to be learned from McClellan's caution. He was the first general of the Army of the Potomac (which was the Union army headquartered near Washington), and he repeatedly missed opportunities to deliver a debilitating blow to General Robert E. Lee's Army of Northern Virginia, most famously during Lee's retreat from Antietam. Despite vastly outnumbering Lee in men and supplies, McClellan chronically overestimated Lee's forces, causing him to be overly cautious and far too timid a commander. Ultimately, the Civil War would drag on for four years and cost over 650,000 American lives. One wonders how different the outcome would have been had McClellan been more willing to take the field.

We encourage the reader to write some simple code to simulate both problems discussed here (or see [5]), namely when we know and when we don't know the lowest tank serial number. These problems provide a valuable warning on how easy it is to accidentally convey information. In many situations today, numbers are randomly generated to

Table 1. Comparison of estimates from statistics and espionage to the true values. (Courtesy of [6].)

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1000	122
June 1941	244	1550	271
August 1942	327	1550	342

prevent such an analysis. Alternatively, sometimes numbers are deliberately initialized at a high value to fool an observer into thinking that more is present than actually is (examples frequently seen are the counting done during a workout, putting money in the tip jar at the start of a shift to encourage future patrons to be generous, or checkbooks with the first check numbered 100 or higher so that the recipient will believe that it is not from a new account).

The German Tank Problem and Linear Regression

The German tank problem is frequently used in classes in probability or discrete math, since it illustrates the power of those two disciplines to use binomial identities to great advantage. It is also seen in statistics classes in discussing how to find good estimators of population values. Focusing on these examples, however, neglects another setting in which it may be used effectively: as an application of the power of linear regression (or the method of least squares). We quickly review how these methods yield the best-fit line or hyperplane, and then generalize to certain nonlinear relationships. We show how simulations can be used to provide support for formulas. This is extremely important, since often we are unable to prove conjectured relationships. Returning to World War II, the Allies could run trials (say drawing numbered pieces of paper from a bag) to model the real-world problem and use the gathered data to sniff out the relationship among m, k, and N.

Additionally, we use this section as an opportunity to discuss some of the issues that can arise in implementing the method of least squares to find the best-fit line. While these do not occur in most applications, it is worthwhile knowing that they can occur and seeing some solutions.

Theory of Regression

Suppose we believe that there are choices of a and b such that given an input x, we should observe $y = ax + b$, but we don't know what these values are. We could observe a large number of pairs of data $\{x_i, y_i\}_{i=1}^{I}$ and use these to find the values of a and b that minimize the sum of the squares of the errors[6] between the observed and predicted values:

$$E(a,b) = \sum_{i=1}^{I} (y_i - (ax_i + b))^2.$$

Setting

$$\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0,$$

we find after some algebra[7] that the best-fit values are

$$\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{I} x_i^2 & \sum_{i=1}^{I} x_i \\ \sum_{i=1}^{I} x_i & \sum_{i=1}^{I} 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{I} x_i y_i \\ \sum_{i=1}^{I} y_i \end{pmatrix}.$$

See, for example, the supplemental material online for [2]. What matters is that the relation is linear in the unknown parameters a and b (or more generally $a_1, \ldots, a_\ell$); similar formulas hold for

$$y = a_1 f_1(x) + \cdots + a_\ell f(x_\ell).$$

For a linear-algebraic approach to regression, see, for example, [4].

Regression is a rich subject; we wish to try to find the best-fit parameters to relate N to m and k; however, we shall shortly see that our initial guess at a relationship is nonlinear. Fortunately, by taking logarithms, we can convert many nonlinear relations to linear ones, and thus the formulas above are available again. The idea is that by doing extensive simulations, we can gather enough data to make a good conjecture on the relationship. Sometimes, as is the case with the German tank problem, we are able to do a phenomenal job in predicting the functional form and coefficients, while at other times, we can get only some values with confidence.

To highlight these features, we first quickly review a well-known problem, the birthday paradox (see, for example, [2]). The standard formulation assumes that we have a year with D days and asks how many people we need in a room to have a greater than 50% probability that at least two share a birthday, under the assumption that the birthdays are independent and uniformly distributed from 1 to D. A straightforward analysis shows that the answer is approximately $D^{1/2}\sqrt{\log 4}$. We now consider the closely related but less well known problem of the expected number of people P we need in a room before there is a match.[8] Based on the first problem, it is reasonable to expect the answer also to be of order $D^{1/2}$, but what is the constant factor? We can try a relation of the form $P = BD^a$,

[6] We cannot just add the errors, since then a positive error could cancel a negative error. We could take the sum of the absolute values, but the absolute value function is not differentiable; it is in order to have calculus available that we measure errors by sums of squares.

[7] The resulting matrix is invertible, and hence there is a unique solution as long as at least two of the x_i differ. One can see this through some algebra, where the determinant of the matrix is essentially the variance of the x_i; if they are not all equal, then the variance is positive. If the x_i are all equal, then the inverse does not exist, but in such a case we should not be able to predict how y varies with x, since we are not varying x!

[8] As a nice exercise, use linearity of expectation to show that we expect at least two people to share a birthday when $P = D^{1/2}\sqrt{2} + 1$.

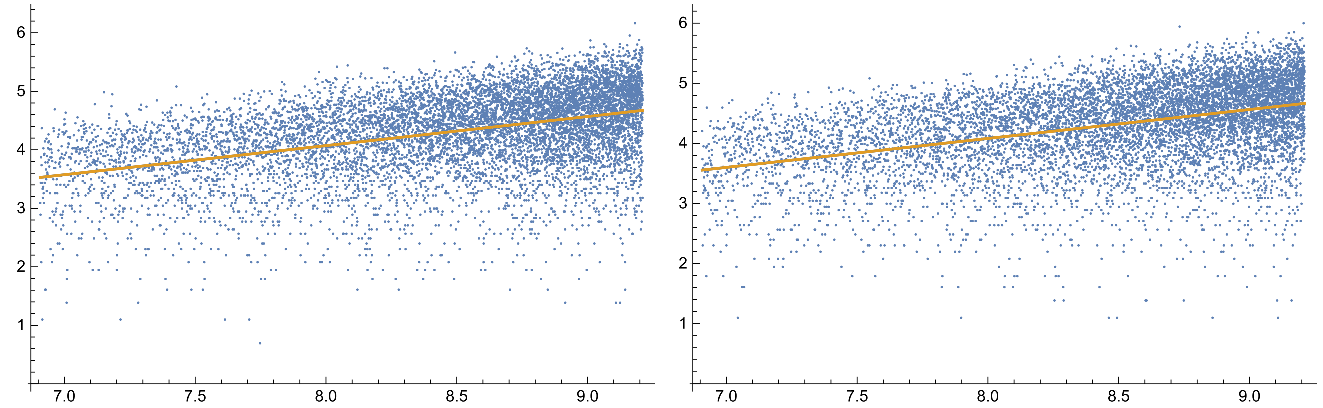

Figure 2. Plot of best-fit line for P as a function of D. We twice ran 10 000 simulations with D chosen from 10 000 to 100 000. Best-fit values were $a \approx 0.506167$, $b \approx -0.0110081$ (left), and $a \approx 0.48141$, $b \approx 0.230735$ (right).

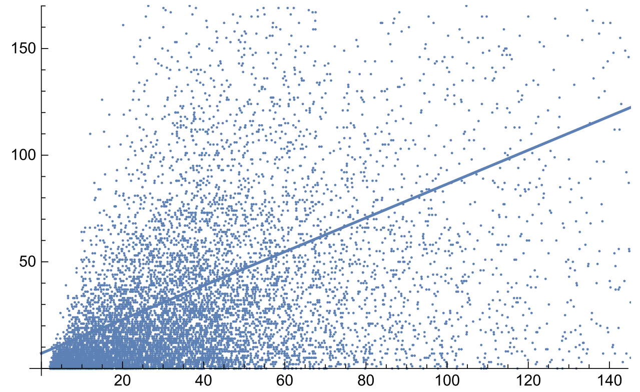

Figure 3. Plot of best-fit line for $N - m$ as a function of m/k. We ran 10 000 simulations with N chosen from [100, 2000] and k from [10, 50]. Best-fit values for $N - m = a(m/k) + b$ for this simulation were $a \approx 0.793716$, $b \approx 7.10602$.

and then taking logarithms (and setting $b = \log B$), we would get $\log P = a \log D + b$. See Figure 2.

Both simulations have similar values for a, both of them consistent with an exponent of 1/2. Unfortunately, the values for b differ wildly, though of the two parameters, we care more about a, since it tells us how our answer changes with the number of days. There is an important lesson here: data analysis can often suggest much of the answer, but it is not always the full story, and there is a role for theory in supplementing such analysis.

Issues in Applying the German Tank Problem

Building on this lesson, we return to the German tank problem. What is a reasonable choice for N as a function of m and k? Clearly, N is at least m, so we try $N = m + f(m, k)$, which transfers the problem to estimating $f(m, k)$. We expect that as m increases, this should increase, and as k increases, it should decrease. Looking at extreme cases is useful: if $k = N$, then $f(N, N)$ should vanish, since then m must equal N. The simplest function that fits this is $f(m, k) = b \cdot m/k$ with b as our free parameter, and we are led to conjecture a relationship of the form

$$N = m + b\frac{m}{k} = m\left(1 + \frac{b}{k}\right).$$

Note that this guess is quite close to the correct answer, but because the observed quantities m and k appear as they do, it is not a standard regression problem. We could try to fix this by looking at $N - m$, the number of tanks we need to add to our observed largest value to get the true number. We could then try to write this as a linear function of the ratio m/k:

$$N - m = a\frac{m}{k} + b,$$

where we allowed ourselves a constant term to increase our flexibility of what we can model. While for $a = -b = 1$ this reproduces the correct formula, finding the best-fit values leads to a terrible fit, as evidenced in Figure 3.

Why is the agreement so poor, given that proper choices exist? The problem is the way m and k interact, and in the setup above, we have the observed quantity m both as an input variable and as an output in the relation. We thus need a way to separate m and k, keeping both on the input side. As remarked, we can do this through logarithms; we discuss another approach in the next subsection.

Resolving Implementation Issues

We look at our best-fit line for two choices of k. The left side of Figure 4 plots N for $k = 1$, while Figure 5 does so for $k = 5$. Both of these show a terrible fit of N as a linear function of m (for fixed k). In particular, when $k = 1$, we expect N to be $2m - 1$, but our best-fit line is about $0.784m + 2875$. This is absurd, since for large m, we predict N to be less than m. Note, however, that the situation is completely different if instead, we plot m against N (the right-hand side of those figures). Clearly, if N depends linearly on m, then m depends linearly on N. When we do the fits this way, the results are excellent.

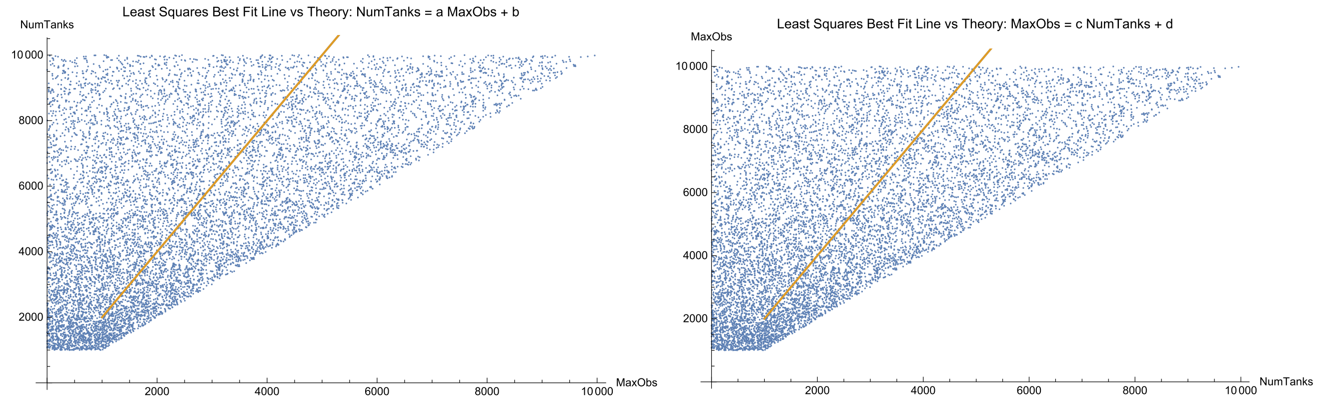

Figure 4. Left: Plot of N vs maximum observed tank m for fixed $k = 1$. Theory: $N = 2m - 1$, best fit $N = 0.784m + 2875$. Right: Plot of maximum observed tank m vs N for fixed $k = 1$. Theory: $m = 0.5N + 0.5$, best fit $m = 0.496N + 10.5$.

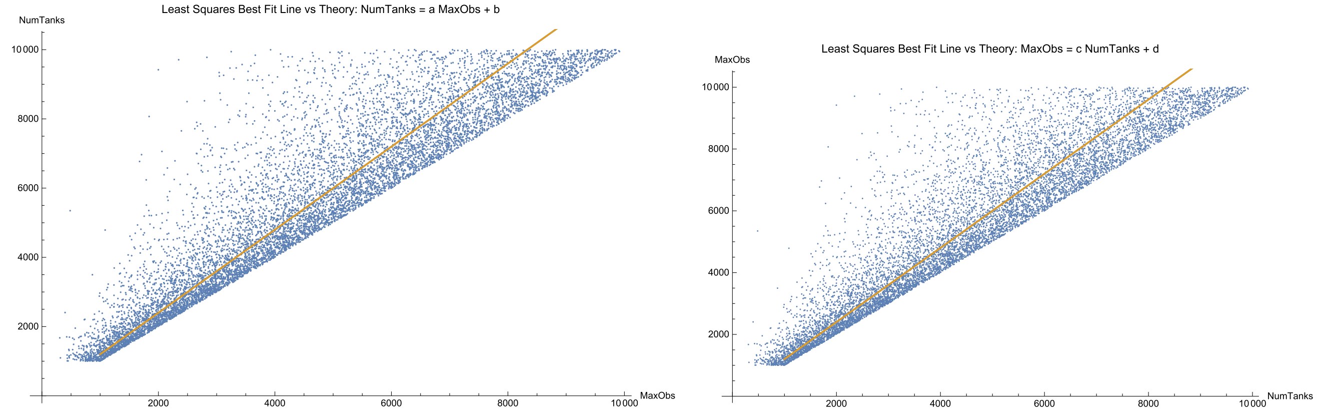

Figure 5. Left: Plot of N vs maximum observed tank m for fixed $k = 5$. Theory: $N = 1.2m - 1$, best fit $N = 1.037m + 749$. Right: Plot of maximum observed tank m vs N for fixed $k = 5$. Theory: $m = 0.883N + 0.883$, best fit $m = 0.828N + 25.8$.

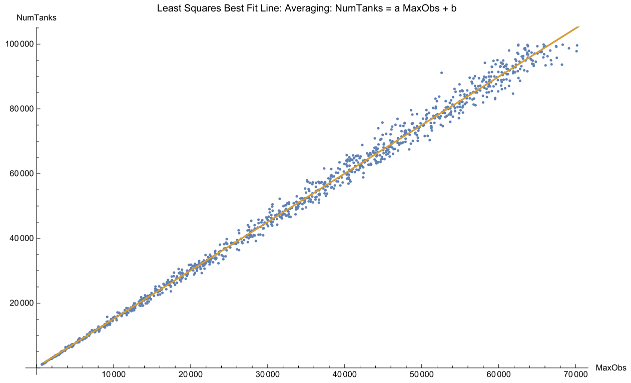

Figure 6. Plot of N vs maximum observed tank m for fixed $k = 1$. Theory: $N = 1.5m - 1$, best fit $N = 1.496m + 171.2$.

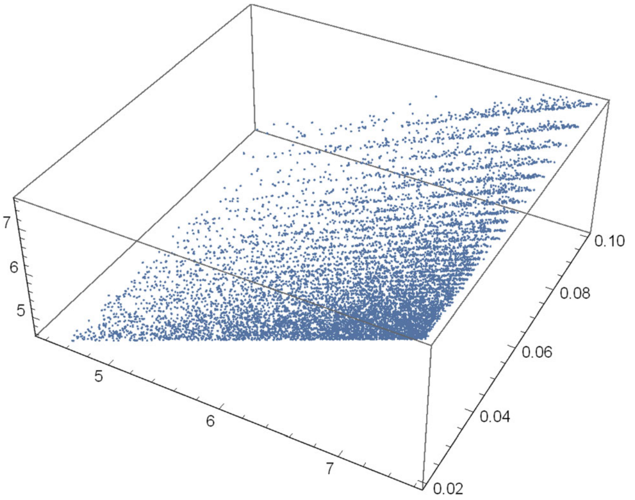

Figure 7. Plot of $\log N$ vs $\log m$ and $1/k$. We ran 10 000 simulations with N chosen from [100, 2000] and k from [10, 50]. The data are well approximated by a plane (which we do not draw, in order to prevent our image from being too cluttered).

Note that from the point of view of an experiment, it makes more sense to plot m as the dependent variable and N as the independent input variable. The reason is the way

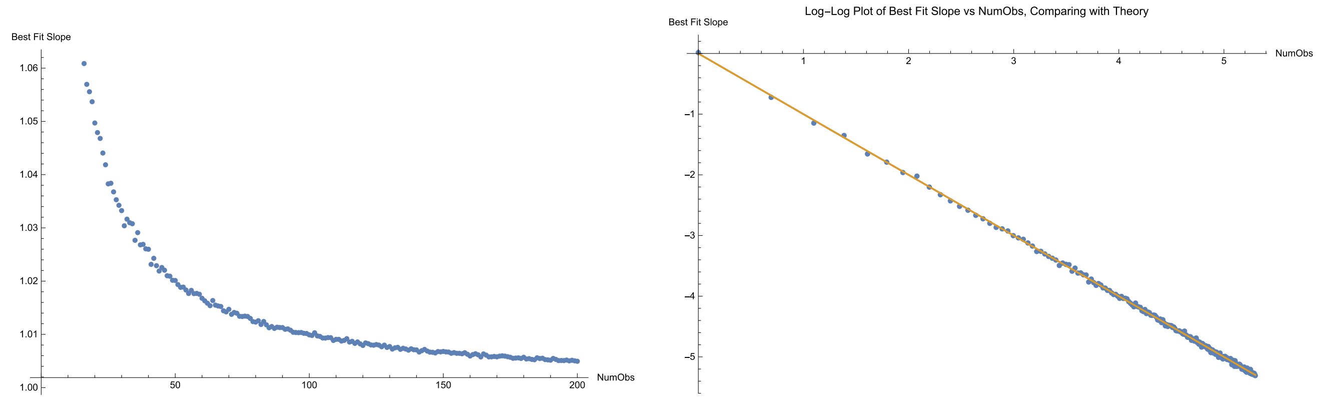

Figure 8. Left: Plot of $a(k)$, the slope in $N = a(k)m + b$, versus k. Right: Log-log plot of $a(k) - 1$ versus k. In $\log(a(k) - 1)$ versus $\log k$, the theory is $\log(a(k) - 1) = -1\log(k)$, while the best-fit line is $\log(a(k) - 1) = -0.999\log(k) - 0.007$.

we simulate; we fix k and N and then choose k distinct numbers uniformly from $\{1, \ldots, N\}$.

We end with another approach that works well and allows us to view N as a function of m. Instead of plotting each pair (m, N) for a fixed k, we instead fix k, choose N, and then do 100 trials. For each trial, we record the largest serial number m, and then we average these and plot $(\overline{m}, N)$, where $\overline{m}$ is the average. This greatly decreases the variability, and we now obtain a nearly perfect straight line and fit; see Figure 6.

Determining the Functional Form

We consider the more general relation

$$N = Cm^a\left(1 + \frac{b}{k}\right),$$

where we expect $C = a = b = 1$; note that this will not have the -1 summand we know should be there, but for large m, that should have negligible impact. Letting $C = e^c$ for notational convenience, we obtain

$$\log N = c + a\log(m) + \log\left(1 + \frac{b}{k}\right).$$

If x is large, then $\log(1 + 1/x) \approx 1/x$, so we try the approximation

$$\log N \approx c + a\log(m) + b\frac{1}{k}.$$

Figure 7 shows the analysis when $C = 1$ (so $c = 0$), since that analysis then reduces to the usual case with two unknown parameters. We chose to take $C = 1$ from the lesson we learned in the analysis of the birthday problem.

The best-fit values of the parameters are $a = 0.999911$ and $b = 0.961167$, which are reasonably close to $a = b = 1$. Thus these numerics strongly support our conjectured relation $N = m(1 + 1/k)$, showing the power of statistics. While we were able to see the arguments needed to prove that this relation is essentially correct, imagine that we could not prove it but still have our heuristic arguments and analysis of extreme cases that suggest it. By simulating data and running the regression, we see that our formula does a stupendous job of explaining our observations, and we thereby gain confidence to use it in the field.

We end with one last approach. Let us guess a relationship of the form $N = a(k)m + b(k)$, where $a(k) = 1 + f(k)$ (we write $a(k)$ as $1 + f(k)$, since we know that there have to be at least m tanks). We can fix k and find the best-fit values of $a(k)$ and $b(k)$. In Figure 8, we plot the best-fit slope $a(k)$ versus k, as well as a log-log plot. For the log-log plot, we look at $a(k) - 1$, subtracting the known component. We see a beautiful linear relation, and thus even if we did not know that it should be m plus a constant times m/k, the data suggest this beautifully. Specifically, we found the best-fit line to be $\log(a(k) - 1) = -0.999\log(k) - 0.007$, suggesting that $a(k) = 1 + 1/k$; we obtain the correct functional form just by running simulations.

ACKNOWLEDGMENTS

We thank our colleagues at Williams College, audiences at several talks over the years, and the referee for valuable feedback that improved the exposition. We thank Jason Zhou for pointing out some typos in an earlier draft. The third-named author was partially supported by NSF Grant DMS1561945. Parts of this paper were given by the first two named authors as their senior colloquium; we thank our colleagues in the Williams Mathematics Department for comments. The third author has presented this topic at several conferences and programs, and he thanks the participants for many suggestions; for video see https://www.youtube.com/watch?v=quV-MCB8Ozs or https://youtu.be/I3ngtIYjw3w.

George Clark
Department of Mathematics and Statistics
Williams College
Williamstown, MA 01267
USA
e-mail: gpclark96@gmail.com

Alex Gonye
Department of Mathematics and Statistics
Williams College
Williamstown, MA 01267
USA
e-mail: aag2@williams.edu

Steven J. Miller
Department of Mathematics and Statistics
Williams College
Williamstown, MA 01267
USA
e-mail: sjm1@williams.edu

REFERENCES

[1] M. Cozzens and S. J. Miller. *The Mathematics of Encryption: An Elementary Introduction*, AMS Mathematical World Series 29. American Mathematical Society, 2013.

[2] S. J. Miller. *The Probability Lifesaver*. Princeton University Press, 2018.

[3] Probability and Statistics Blog. How many tanks? MC testing the GTP. Available at https://statisticsblog.com/2010/05/25/how-many-tanks-gtp-gets-put-to-the-test/.

[4] G. Strang. *Introduction to Linear Algebra*, fifth edition. Wellesley-Cambridge Press, LL 2016.

[5] Statistical Consultants Ltd. The German tank problem. Available at https://www.statisticalconsultants.co.nz/blog/the-german-tank-problem.html.

[6] WikiEducator. Point estimation—German tank problem. Available at https://wikieducator.org/Point_estimation_-_German_tank_problem.