GENERALIZING THE GERMAN TANK PROBLEM

ANTHONY LEE AND STEVEN J. MILLER

ABSTRACT. The German Tank Problem dates back to World War II when the Allies used a statistical approach to estimate the number of enemy tanks produced or on the field from observed serial numbers after battles. Assuming that the tanks are labeled consecutively starting from 1, if we observe k tanks from a total of N tanks with the maximum observed tank being m, then the best estimate for N is $m(1 + 1/k) - 1$. We refer to an estimate as 'best' when the estimate is closest to the actual number of tanks. We explore many generalizations; first, we looked at the discrete and continuous one-dimensional case. We attempted to improve the original formula by using different estimators such as the second largest and L^{th} largest tank, and applied motivation from portfolio theory by seeing if a weighted average of different estimators would produce less variance; however, the original formula, using the largest tank proved to be the best; the continuous case was similar. Then, we looked at the discrete and continuous square and circle variants where we pick pairs instead of points, which were more complex as we dealt with problems in geometry and number theory, such as dealing with curvature issues in the circle, and the problem that not every number is representable as a sum of two squares. In some cases, when we could not derive precise formulas, we derived approximate formulas. For the discrete and continuous square, we tested various statistics, but found that the largest observed component of our pairs is the best statistic to look at; the scaling factor for both cases is $(2k + 1)/2k$. For the circle we used motivation from the equation of a circle; for the continuous case, we looked at $\sqrt{X^2 + Y^2}$ and for the discrete case, we looked at $X^2 + Y^2$ and took a square root at the end to estimate for r. Interestingly, the scaling factors, a number, generally a little greater than 1, that we multiplied to scale up to get our estimation, were different for the cases. Lastly, we generalized the problem into L dimensional squares and circles. The discrete and continuous square proved to be similar to the two-dimensional square problem. However, for the L^{th} dimensional circle, we had to use formulas for the volume of the L-ball, and had to approximate the number of lattice points inside it. The discrete circle formula was particularly interesting, as there was no L dependence in the formula.

Keywords: German Tank Problem, Uniform Distribution, Discrete setting, Continuous setting,

CONTENTS

Date: January 17, 2023.
2010 *Mathematics Subject Classification.* 60B10, 11B39, 11B05 (primary) 65Q30 (secondary).

1. INTRODUCTION

We study a problem where we are given some information about observations, and we have to estimate the number of objects. The motivation comes from the German Tank problem, a classic problem in probability. During World War II, Germans used powerful tanks to their advantage. To develop appropriate military strategies against the Germans, the Allies had to estimate the number of tanks on various battlefields. Initially, spies were used. However, using statistical estimates proved to be more accurate; see [Ka] for a history of the problem, and [Si] for additional details and a solution through a Bayesian approach. During battles, the Allies realized that the destroyed or captured tanks had serial numbers and that they could reverse engineer and use these to their advantage. Assuming that the serial numbers are consecutive[1] and started from 1, the statisticians came up with a formula using the largest observed tank m_k and the number of tanks observed k to make a estimate for N, which we denote $\widehat{N}$:

$$\widehat{N} \;=\; m_k\left(1+\frac{1}{k}\right) \;-\; 1. \tag{1.1}$$

This formula proved to be effective. A comparison between the actual and estimated production rate of tanks from [Al] shows how accurate this formula was: while the intelligence estimated that Germans were producing 1,400 tanks per month, using the formula, the statisticians estimated that Germans were producing 256 tanks per month, and indeed, 255 tanks were made! The chart below compares the estimation from the statistical method and intelligence to the actual number of German Tanks [Ru].

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342

FIGURE 1. Statistics vs Intelligence estimates.

We see that the statistical estimate is reasonably similar to the German records. However, the intelligence estimate is far off from the actual number of tanks. Had the Allies used the intelligence estimate, their strategy would have misled them as they would have overly prepared defense to minimize damage from tanks or assigned too many tanks.

[1]It is convenient to have the numbers in order; by looking at the serial number one can often tell when it was made and thus when it may need certain types of repairs. However, as the previous work shows, this opens one up to disclosing more information than one would like, and in many applications now companies use formulas to determine the serial numbers, masking information.

The German Tank problem is an excellent example of how statistical inference can be applied to real world problems. We attempt to generalize the well known German Tank Problem. Previously, Clark, Gonye, and Miller [CGM] derived a more general formula where the smallest serial number is not 1, but the tanks are still numbered consecutively. If the spread between the smallest and largest observed serial number is s, then their formula to estimate the number of tanks is

$$\widehat{N} \ = \ s\left(1+\frac{2}{k-1}\right) \ - \ 1. \tag{1.2}$$

We start by recalling the derivation of the original German tank problem, as we will be extending those calculations. We then attempt to improve the one-dimensional formula by looking at different estimators and using motivation from portfolio theory in financial mathematics: we look at the weighted sums of estimators to see if we can find a new statistic that has the same mean of its predictions but smaller variance; see Appendix A for a review of portfolio theory and the construction of such statistics. However, we find that the original formula does better than formulas from these two approaches.

Generalizing further, we looked at what would happen if we modified some of the assumptions of the problem. First, we modified the condition that all serial numbers are consecutive integers and we looked at the standard German Tank Problem in the continuous setting. The problem changed slightly as we sample k tanks from the range of 0 to N, but the tanks can be non-integer real numbers. To find the answer of the question of finding the best estimator for N, we tried various statistics, starting with some obvious choices such as the largest observed tank, the second largest observed tank, and the weighted sum. Interestingly, the continuous case turned out to be similar to the discrete case, as the best statistic to look at in the continuous case was the largest observed tank, as it produced a formula with the least variance. Furthermore, the scaling factor, a number, generally a little greater than 1, that we multiplied to scale up to get our estimation, for the discrete and continuous case were both $(k+1)/k$, which shows the similarity between the two. Even in the continuous case, the largest observed tank is the best statistic to study.

Moving on from the one-dimensional case, we generalized the German Tank problem further into two dimensions, selecting k pairs without replacement. Specifically, we looked at the square and circle, as we have explicit closed form and asymptotic expressions for the area and number of lattice points inside. The main strategy of deriving formulas was using the CDF method for the discrete and continuous case. (See Lemma 2.9 for details.) Starting from the two-dimensional case, we see geometry involved in the calculations. For the discrete square, we looked at the square with bottom left vertex at $(1,1)$ and upper right vertex at (N,N) and looked at the number of lattice points inside the square. From inspiration from the one-dimensional problem, we looked at the largest observed component. Unfortunately, because we get complex closed form expressions that are difficult to invert, especially in the discrete cases, we approximate by using the main term and sometimes the second order, which gives us a accurate approximation. The continuous square problem was easier as calculating integrals was easier than calculating sums. We set the bottom left vertex as $(0,0)$ and the upper right vertex as (N,N). We also looked at the largest observed component for this case too. However, instead of looking at the number of lattice points as we did in the discrete case, we looked at the area, which was easier. Interestingly, the scaling factor for both the discrete and continuous square was the same, and we state both formulas.

$$\textbf{Discrete Square} : \widehat{N} \ = \ \frac{2k+1}{2k}(m-1). \quad \textbf{Continuous Square} : \widehat{N} \ = \ \frac{2k+1}{2k}\cdot m. \tag{1.3}$$

After looking at the square problem in the discrete and continuous case, we looked at the circle problem. For both the discrete and continuous cases, we start with a circle with center at (0,0) and radius r. The

statistics we looked at the for the discrete and continuous cases were different. For the continuous case, we looked at the largest observed value of $m_2 = \sqrt{X^2 + Y^2}$. The motivation comes from the standard form equation of the circle, which is $(x - x_1)^2 + (y - y_1)^2 = r^2$ where (x, y) is the point on the circumference, r is the radius, and (x_1, y_1) are the coordinates of the center. We used the CDF method with areas to calculate the formula.

Before we look at the discrete case, we see that we need to know the number of lattice points inside a circle with center $(0, 0)$ and radius r. We visit the classic Gauss Circle problem and use the approximations and denote the error term using Big-O notation to denote the number of lattice points inside the circle. To solve the discrete case, we looked at various statistics. First, we looked at the largest observed component, as we were able to obtain nice formulas in the square problem using the largest observed component. However, this statistic was too complex as when we used a circle, we had to split the ranges of the side of the square to see if the square fit in a circle or not. Thus, we decided that this is not the best statistic to look at. We wanted to look at a statistic similar to $\sqrt{X^2 + Y^2}$, the formula for the radius of the circle, but we had to make sure that the statistic that we were studying gave only discrete values as outcomes. Therefore, we chose to look at $m = X^2 + Y^2$ as all values of m are integers. At the end, we have to take the square root of m because now, we are essentially estimating for m^2. Another complication that arises in this problem is that $X^2 + Y^2 \equiv 0, 1, 2 \pmod 4$. We explain more about the number theory complication when we derive the formula. Also, because we used an asymptotic formula for the number of lattice points inside the circle, we weren't able to produce an exact formula but one that still gives accurate estimations. We see that the discrete and continuous formulas are different; see Remark 6.4.

$$\textbf{Discrete Circle} : \widehat{r} = \sqrt{\frac{k+1}{k}(m-1)}. \quad \textbf{Continuous Circle} : \widehat{r} = \frac{2k+1}{2k} \cdot m. \tag{1.4}$$

Finally, we generalized the problem into L dimensions. We looked at the L^{th} dimensional square and the L^{th} dimensional sphere, also known as the L-ball, in the continuous and discrete setting. The problem changes slightly as we are selecting k tuples of length L without replacement. For the discrete and continuous L^{th} dimensional square cases, we looked at the largest observed component value from the tuples and derived the formulas. For the continuous L^{th} dimensional square, we obtained an exact formula, but for the discrete L^{th} dimensional square, we approximated the result using the main term. We see that the scaling factors for the discrete and continuous setting are the the same.

$$\textbf{Discrete L-dim Square} : \widehat{N} = \frac{Lk+1}{Lk}(m-1). \quad \textbf{Continuous L-dim Square} : \widehat{N} = \frac{Lk+1}{Lk} \cdot m. \tag{1.5}$$

The difficulty is calculating the number of lattice points inside the L-ball. We use the known formula about the volume of the L-ball, and use this formula to estimate how many lattice points are contained inside using Big-O notation. For our statistic in the discrete case, we looked at $m = X_1^2 + X_2^2 + \cdots + X_L^2$, as this statistic guarantees that all m values are integer-valued. Then after estimating, we took the square root of m to find the estimate for r. For the continuous case, we looked at $m = \sqrt{X_1^2 + X_2^2 + \cdots + X_L^2}$ as we are allowed to get real numbers for m values. The continuous L-ball problem was easier, as we could plug in the equation for the L-ball to get the volume instead of having to estimate the number of lattice points. We state the formulas we derived:

$$\textbf{Discrete L-ball} : \quad \widehat{r} = \sqrt{\frac{k+1}{k} \cdot (m-1)}. \quad \textbf{Continuous L-ball} : \widehat{r} = \frac{Lk+1}{Lk} \cdot m. \tag{1.6}$$

2. PRELIMINARIES

We review some needed results from probability, combinatorics, and integration.

2.1. **Probability Review.** We list a few standard results from probability; for proofs see for example [Mi, Sh].

Definition 2.1. *(Pascal's Identity) We have*

$$\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}. \tag{2.1}$$

Definition 2.2. *(Hockey Stick Identity) We have*

$$\sum_{i=r}^{n} \binom{i}{r} = \binom{n+1}{r+1}. \tag{2.2}$$

Definition 2.3. *The variance for a random variable X is the average of the squared difference from the mean, $\mathbb{E}[X]$:*

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{2.3}$$

Lemma 2.4. *The variance can be computed by*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{2.4}$$

Definition 2.5. *For two jointly distributed real valued random variables X and Y, the covariance is defined as the expected value of the product of their deviations from their individual expected values:*

$$\mathrm{Cov}(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\right]. \tag{2.5}$$

Lemma 2.6. *The covariance can be computed by*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{2.6}$$

Theorem 2.7. *(Linearity of expectation) Let $X_1, \dots, X_n$ be random variables, and let $g_1, \dots, g_n$ be functions such that $\mathbb{E}[|g_i(X_i)|]$ exists and is finite, and let $a_1, \dots, a_n$ be any real numbers; note the random variables do not have to be independent. Then*

$$\mathbb{E}\left[a_1 g_1(X_1) + \cdots + a_n g_n(X_n)\right] = a_1\mathbb{E}[g_1(X_1)] + \cdots + a_n\mathbb{E}[g_n(X_n)]. \tag{2.7}$$

Theorem 2.8. *(Joint Probability Density Function) Let $X_1, X_2, \dots, X_n$ be continuous random variables with densities $f_{X_1}, f_{X_2}, \dots, f_{X_n}$ defined on $\mathbb{R}$. The joint density function of the tuple $(X_1, \dots, X_n)$ is a non-negative integrable function $f_{X_1,X_2,\dots,X_n}$ such that for every set $S \subset \mathbb{R}^n$ we have*

$$\mathrm{Prob}((X_1, \dots, X_n) \in S) = \int \cdots \int_S f_{X_1,X_2,\dots,X_n}(x_1, \dots, x_n)\, dx_1 \cdots dx_n, \tag{2.8}$$

and

$$f_{X_i}(x_i) = \int_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_n=-\infty}^{\infty} f_{X_1,X_2,\dots,X_n}(x_1, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n. \tag{2.9}$$

For discrete random variables, we can replace integrals with sums.

Definition 2.9. *The cumulative distribution function (CDF) of a random variable X with density f, denoted F, is given by*

$$F(x) := \mathrm{Prob}(X \le x) = \int_{-\infty}^{x} f(t)\, dt, \text{ for any } x \in \mathbb{R}. \tag{2.10}$$

In many situations it is easy to compute the CDF, and thus by using the Fundamental Theorem of Calculus we can determine the probability density function. In the continuous case it is the derivative of the CDF, in the discrete case when the values of our random variable are non-negative integers, then the probability of m is $F(m) - F(m-1)$.

2.2. **Analysis Review.** Because many of the terms we encounter are complex and un-invertible, we estimate using the main term and sometimes the second order term. The following Theorem frequently provides a good, easily computed bound; see for example [YK].

Theorem 2.10. *(Euler-Maclaurin formula) For p a positive integer and a function $f(x)$ that is p times continuously differentiable on the interval $[a, b]$, we have*

$$\sum_{i=a}^{b} f(i) \quad = \quad \int_a^b f(x)\,dx + \frac{f(a)+f(b)}{2} \; + \sum_{q=1}^{\lfloor \frac{p}{2} \rfloor} \frac{B_{2q}}{(2q)!}(f^{2q-1}(b) - f^{2q-1}(a)) + R_p, \quad (2.11)$$

and

$$|R_p| \; \leq \; \frac{2\zeta(p)}{(2\pi)^p} \int_m^n |f^{(p)}(x)|\,dx. \quad (2.12)$$

The estimation below is needed in computing upper and lower bounds in applications of the Euler-Maclaurin formula later.

Lemma 2.11. *For $m, L \geq 0$ and $k \geq 1$,*

$$\left(m^{Lk} - m^{Lk-L}\left(\frac{k(k-1)}{2}\right)\right) \; \leq \; m^L(m^L-1)\cdots(m^L-(k-1)) \leq m^{Lk}. \quad (2.13)$$

Proof. The upper bound follow trivially as

$$m^L(m^L-1)\cdots(m^L-(k-1)) \; \leq \; (m^L)(m^L)\cdots(m^L). \quad (2.14)$$

To justify the lower bound, we want to prove the inequality

$$m^L(m^L-1)\cdots(m^L-(r-1)) \; \geq \; m^{Lr} - \frac{r(r-1)}{2}m^{Lr-L}. \quad (2.15)$$

The base case is satisfied when $r = 1$ as $m^2 \; \geq \; m^2$. We assume that the case when $r = k$ is true:

$$m^L(m^L-1)\cdots(m^L-(k-1)) \; \geq \; m^{Lk} - \frac{k(k-1)}{2}m^{Lk-L}. \quad (2.16)$$

We must show it holds when $r = k+1$; namely we must show

$$m^L(m^L-1)\cdots(m^L-k) \; \geq \; m^{Lk+L} \; - \; \frac{(k+1)k}{2}m^{Lk}. \quad (2.17)$$

This follows immediately by substitution and expansion:

$$\begin{aligned} m^L(m^L-1)\cdots(m^L-(k-1))(m^L-k) \quad &\geq \quad \left(m^{Lk} - \frac{k(k-1)}{2}m^{Lk-L}\right)(m^L-k) \\ &\geq \quad m^{Lk+L} \; - \; \frac{k(k+1)}{2}m^{Lk} \; + \; \frac{k^2(k-2)}{2}m^{Lk-L} \\ &\geq \quad m^{Lk+L} \; - \; \frac{k(k+1)}{2}m^{Lk}. \end{aligned} \quad (2.18)$$

□

Finally, in many of our generalizations it is impossible to obtain simple closed form expressions such as those in the original problem. We thus investigate the large N limit, and Big-O notation is useful in isolating the main term from lower order terms which have negligible effect.

Definition 2.12. *(Big-O Notation) Suppose $f(x)$ and $g(x)$ are two functions defined on the real numbers. We write $f(x) = O(g(x))$ (read "f is Big-O of g") if there exists a positive constant C such that $|f(x)| \leq Cg(x)$ is satisfied for all sufficiently large x.*

2.3. **Combinatorial Results.** We list four useful identities that are needed in later calculations. These are generalizations of the famous hockey stick identity from Pascal's triangle; that identity is responsible for the closed form expression in the original problem, and these identities play a similar role in our generalizations. The identities can be proved using the hockey stick numerous times; to see the full proofs, refer to Appendix B.

Identity I: For all $k \geq 0$,

$$\sum_{i=0}^{k} \binom{a+i}{a}\binom{b+k-i}{b} \;=\; \binom{a+b+k+1}{a+b+1}. \tag{2.19}$$

Identity II: For all $N \geq k$,

$$\sum_{m=k}^{N} \binom{m-b}{k-c} \;=\; \binom{N-b+1}{k-c+1} \;-\; \binom{k-b}{k-c+1}. \tag{2.20}$$

Identity III: For all $N \geq k$,

$$\sum_{m=k-a+1}^{N-a+1} m\frac{\binom{m-1}{k-a}\binom{N-m}{a-1}}{\binom{N}{k}} \;=\; \frac{(N+1)(k-a+1)}{k+1}. \tag{2.21}$$

Identity IV: For all $N \geq k$,

$$\begin{aligned}\sum_{m=k-a+1}^{N-a+1} m^2\frac{\binom{m-1}{k-a}\binom{N-m}{a-1}}{\binom{N}{k}} \;&=\; \frac{(k-a+1)(k-a+2)(N+2)(N+1)}{(k+2)(k+1)} \\ &\quad - \frac{(N+1)(k-a+1)}{k+1}.\end{aligned} \tag{2.22}$$

3. Derivation of Original German Tank Problem

We derive the formula for the original problem where m_k represents the largest tank observed, k the number of tanks observed, and N the number of tanks (numbered consecutively from 1 to N). We use $\widehat{N}$ to represent our estimate for the total number of tanks, N. It is useful to run through this argument before we generalize the inference problem, and we follow the exposition in [CGM]. We prove

$$\widehat{N} \;=\; m_k\left(1+\frac{1}{k}\right) \;-\; 1. \tag{3.1}$$

We check some extreme cases to see if this formula is reasonable. When we observe just one tank, the estimation formula is $2m-1$. We expect the m_k to be $N/2$, so multiplying by 2 is reasonable. When we observe all N tanks (so $k = N$), the estimation $m = N$, which is also reasonable because we know the number all the tanks. We use Pascal's identity and the hockey stick identity frequently in calculations.

3.0.1. *The PDF of the Largest Observed Tank.* Let M_k be the random variable for the largest tank observed, and let m_k be its observed value. Thus m_k is the largest tank serial number that we observe. Our goal is to find a formula to estimate N from m_k and k. We first compute the PDF for M_k.

Lemma 3.1. *For $k \leq m_k \leq n$,*

$$\text{Prob}(M_k = m_k) \;=\; \frac{\binom{m_k-1}{k-1}}{\binom{N}{k}}. \tag{3.2}$$

Proof. We know that the total number of ways to select k tanks from N possible tanks is $\binom{N}{k}$. If the largest tank we observe is m, then we have to choose $k-1$ tanks from $m_k - 1$ possibilities. Therefore, the probability that the largest tank we observe is m_k is $\binom{m_k-1}{k-1}/\binom{N}{k}$, thus proving the claim. □

Remark 3.2. *It is worth remarking that the process of deriving the formula is different from the application. In the application, we use tanks we observe, $m_1, m_2, \ldots, m_k$ and estimate for N by applying the formula. However, when we are deriving the formula, we use N and k to compute the expected value of M_k, and then we invert the equation to find the formula for N.*

3.0.2. *Derivation.* Now we calculate the expected value of M_K in order to find an equation with N and k. To calculate the expected value, we multiply each value of the random variable by its probability and add the products. We expand the binomials, and regroup so that we can use identity III. We find

$$\mathbb{E}[M_k] \;=\; \sum_{m_k=k}^{N} m_k \text{Prob}(M_k = m_k) \;=\; \sum_{m_k=k}^{N} m_k \frac{\binom{m_k-1}{k-1}}{\binom{N}{k}} \;=\; \frac{k(N+1)}{k+1}. \tag{3.3}$$

Using our formula for M_k as a function of N and k, we invert:

$$N \;=\; \mathbb{E}[M_k]\left(\frac{k+1}{k}\right) \;-\; 1. \tag{3.4}$$

To obtain our estimate $\widehat{N}$ we substitute the observed m_k for $\mathbb{E}[M_k]$:

$$\widehat{N} \;=\; m_k\left(\frac{k+1}{k}\right) \;-\; 1. \tag{3.5}$$

We see how well the formula does with simulations that gives mean and variance when we input N and k. We see difference in variance in the two runs; this is because on the left run, we saw 15 percent of total tanks whereas on the right run, we saw 10 percent of total tanks, which caused the difference in variance.

```
germantankcombinedestimatenolist[400, 60, 1 000 000]
German Tank Problem Calculations: N = 400, k = 60.
Check of Means / Variances.
Mean(X1) = 400.002; Var(X1) = 36.6004.
Mean(X2) = 399.997; Var(X2) = 74.5369.
Mean(X) = 400.002; Var(X) = 36.6004.
```

```
germantankcombinedestimatenolist[100, 10, 1 000 000]
German Tank Problem Calculations: N = 100, k = 10.
Check of Means / Variances.
Mean(X1) = 99.9985; Var(X1) = 75.7603.
Mean(X2) = 100.001; Var(X2) = 168.491.
Mean(X) = 99.9985; Var(X) = 75.7603.
```

FIGURE 2. Results for formula using largest tank.

4. ESTIMATING WITH MORE TANKS

There are two natural approaches to try to improve our prediction for the original problem. One is to find another estimator; perhaps the mean or median might do better. We pursue this and investigate statistics derived from the L^{th} largest tank.

Another approach is to use motivation from portfolio theory; see for example [H]. Imagine we have two independent stocks with the same expected return. By looking at a linear combination, the combined portfolio will still have the same expected return, but if the weights are chosen properly a smaller variance; we provide details in Appendix A. The idea of weighing two random variables to create a combined one with less variance is a common method, and we apply that to our problem to see if we can improve the quality of the estimator for the German tank problem by looking at a combination of two tanks. The computation is a bit more involved than the two stock example, as the values of the two tanks are not independent; if we know the largest tank value is m_k then clearly $m_{k-1} < m_k$. We found that the quality of the estimator is *not* improved by incorporating both values.

4.1. **Estimation from Various Tanks.** Let M_{k-1} be the random variable for the value of the second largest tank observed and let m_{k-1} be the value we observe. We find the probability that the second largest tank is m_{k-1}. We claim that for $k-1 \le m_{k-1} \le n-1$, the probability that $M_{k-1} = m_{k-1}$ is

$$\text{Prob}(M_{k-1} = m_{k-1}) \;=\; \frac{\binom{m_{k-1}-1}{k-2}\binom{N-m_{k-1}}{1}}{\binom{N}{k}}. \tag{4.1}$$

Clearly the probability is zero for tanks outside this range. There are $\binom{N}{k}$ ways to choose k tanks from N. We need $k-2$ tanks to be smaller than $m_{k-1}-1$; there are $\binom{m_{k-1}-1}{k-2}$ ways for that to happen. We then have to choose tank m_k, which has to be larger than tank m_{k-1}. Thus, the range from m_k goes from $m_{k-1}+1$ to N, and there are $\binom{N-m_{k-1}}{1}$ ways to do this. □

We can now calculate the expected value of M_{k-1} using Identity III:

$$\begin{aligned}\mathbb{E}[M_{k-1}] \;&=\; \sum_{m_{k-1}=k-1}^{N-1} m_{k-1}\text{Prob}(M_{k-1} \;=\; m_{k-1}) \\ &=\; \sum_{m_{k-1}=k-1}^{N-1} m_{k-1}\frac{\binom{m_{k-1}-1}{k-2}(N-m_{k-1})}{\binom{N}{k}} \;=\; \frac{(N+1)(k-1)}{k+1}.\end{aligned} \tag{4.2}$$

We attach results of simulations to see the mean and variances of the formula using m_{k-1}.

```
Timing[germantankcombinedestimatenolist2[400, 60, 1 000 000] ]
German Tank Problem Calculations: N = 400, k = 60.
Check of Means / Variances.
Mean(X2) = 399.999; Var(X2) = 74.6674.
{7.0493, Null}
```

```
In[21]:= Timing[germantankcombinedestimatenolist2[100, 10, 1 000 000]]
German Tank Problem Calculations: N = 100, k = 10.
Check of Means / Variances.
Mean(X2) = 100.002; Var(X2) = 168.349.
Out[21]= {6.03502, Null}
```

FIGURE 3. Results for formula using second largest tank.

Remark 4.1. *Comparing this result to the variance of the formula using the largest tank, we see that the formula using the second largest tank has higher variance, which we prove later.*

Calculating similarly, we obtain the formula for the L^{th} largest tank:

$$\widehat{N} \;=\; m_{k-L+1}\frac{k+1}{k-L+1} \;-\; 1. \tag{4.3}$$

We calculate the variances for statistics arising from the different tank (from largest observed to smallest); we multiply each by a multiplicative factor so that the mean is N, and see which has the least variance. We let X_k be the estimation from using the largest tank, X_{k-1} from using the second largest tank, and so on. To calculate the variances of these formulas, we calculate the variances of M_k, the largest tank, M_{k-1}, the second largest tank, and so on; we then multiply by a multiplicative factor to get the variances we want.

4.1.1. *Variance of M_k.* : We use Lemma 2.4, to compute the variance. To calculate the $\mathbb{E}[M_k^2]$ term, we use Identity IV. We find

$$\begin{aligned}\mathrm{Var}(M_k) &= \mathbb{E}[M_k^2] - \mathbb{E}[M_k]^2 = \sum_{m_k=k}^{N} m_k^2 \mathrm{Prob}(\mathrm{M_k = m_k}) - \left[\sum_{m_k=k}^{N} m_k \mathrm{Prob}(\mathrm{M_k = m_k})\right]^2 \\ &= \sum_{m_k=k}^{N} m_k^2 \frac{\binom{m_k-1}{k-1}}{\binom{N}{k}} - \left[\frac{k(N+1)}{k+1}\right]^2 = \frac{(k)(N-k)(N+1)}{(k+1)^2(k+2)}. \end{aligned} \tag{4.4}$$

Now that we have the variance for M_k, we easily scale the formula and obtain the variance of $X_k = m_k(k+1)/k$, as the variance of a times X is a^2 times the variance of X:

$$\mathrm{Var}(X_k) = \mathrm{Var}(M_k) \cdot \frac{(k+1)^2}{k^2} = \frac{(N-k)(N+1)}{(k)(k+2)}. \tag{4.5}$$

Through similar calculations, we obtain the variance of X_{k-1}:

$$\mathrm{Var}(X_{k-1}) = \frac{2(N-k)(N+1)}{(k+2)(k-1)}. \tag{4.6}$$

We see that the estimator using m_{k-1}, and more generally using m_{k-L+1}, is worse than using m_k, as the variance is larger. We can plug in some values of N and k and compare the variances. Also, we see that the variance for X_{k-1} is roughly two times the variance of X_k, which shows that X_k is a better statistic. We can also compare variances easily by writing some code and numerically exploring. We made a simulation where we sample the k tanks from 1 to N when all tanks are equally likely to be seen. Thus, we conclude that if we are only going to use one observed value, it is best to use the largest. This leads to our second question: can we do better if we create a statistic combining two or more observed values?

4.2. **Weighted Statistic.** Previously, we have only considered using one tank to estimate N. We create a statistic using the largest and second largest tank values and show that this weighted statistic is not better. Before we state the weighted formula, we set our notation.

- X_k = Statistic to estimate N using M_k: $M_k(\frac{k+1}{k})$ - 1.
- X_{k-1} = Statistics to estimate N using M_{k-1}: $M_{k-1}(\frac{k+1}{k-1})$ - 1.

Let $\alpha \in [0,1]$ and define the weighted statistic X_α by

$$X_\alpha := \alpha X_k + (1-\alpha) X_{k-1}. \tag{4.7}$$

From the formula, we see that when α is 1, this collapses to the formula for N using m_k, and similarly, when α is 0, we get the m_{k-1} formula. In order to see if there is a better estimation, we find the optimal α value, the value that minimizes the variance of X_α. If the optimal α value is 1, then we can conclude that the m_k formula is the best we can do to estimate for N. We compute the variance of X:

$$\mathrm{Var}\ \mathrm{X}_\alpha = \alpha^2 \mathrm{Var}\ \mathrm{X_k} + (1-\alpha)^2 \mathrm{Var}\ \mathrm{X_{k-1}} + 2\,\alpha(1-\alpha)\mathrm{Cov}(X_k, X_{k-1}). \tag{4.8}$$

As we have already calculated the variances of X_k and X_{k-1}, we have to calculate the covariance term.

4.2.1. ***Covariance term***. We calculate the term, $\mathrm{Cov}[X_k, X_{k-1}]$, separately first. By Lemma 2.6, we have

$$\mathrm{Cov}[X_k, X_{k-1}] \;=\; \mathbb{E}[X_k \cdot X_{k-1}] \;-\; \mathbb{E}[X_k] \cdot \mathbb{E}[X_{k-1}]. \tag{4.9}$$

Recall that

$$X_k \;=\; m_k\left(\frac{k+1}{k}\right) - 1\,, \quad X_{k-1} \;=\; m_{k-1}\left(\frac{k+1}{k-1}\right) - 1. \tag{4.10}$$

We know that the second term of the covariance, $\mathbb{E}[X_k]\cdot\mathbb{E}[X_{k-1}]$, is N^2 because they are both estimation formulas. Thus, we only have to calculate the first term.

We use linearity of expectation to expand. By Theorem 2.7, we have

$$\begin{aligned}\mathbb{E}[X_k \cdot X_{k-1}] &= \mathbb{E}\left[\left(M_k\left(\frac{k+1}{k}\right) - 1\right) \cdot \left(M_{k-1}\left(\frac{k+1}{k-1}\right) - 1\right)\right] \\ &= \frac{(k+1)^2}{k(k-1)}\mathbb{E}[M_k \cdot M_{k-1}] \;-\; \frac{k+1}{k}\mathbb{E}[M_k] \;-\; \frac{k+1}{k-1}\mathbb{E}[M_{k-1}] \;+\; \mathbb{E}[1].\end{aligned} \tag{4.11}$$

We calculate term by term. We use the joint PDF, recall Theorem 2.8, to calculate $\mathbb{E}[M_k \cdot M_{k-1}]$.

$$\begin{aligned}\mathbb{E}[M_k \cdot M_{k-1}] &= \sum_{m_k=k}^{N}\sum_{m_{k-1}=k-1}^{m_k-1} m_k m_{k-1}\mathrm{Prob}(M_{k-1} = m_{k-1}, M_k = m_k) \\ &= \sum_{m_k=k}^{N}\sum_{m_{k-1}=k-1}^{m_k-1} m_k m_{k-1}\frac{\binom{m_{k-1}-1}{k-2}}{\binom{N}{k}}.\end{aligned} \tag{4.12}$$

The probability is the following because, after selecting m_k and m_{k-1}, we have to select $(k-2)$ more tanks from the possible $(m_{k-1}-1)$ tanks. The range of m_{k-1} is dependent on the value of m_k because $m_{k-1} \;\leq\; m_k$ is always satisfied. We use Identity III two times to calculate the following term.

$$\mathbb{E}[M_k \cdot M_{k-1}] \;=\; \sum_{m_k=k}^{N} m_k \sum_{m_{k-1}=k-1}^{m_k-1} m_{k-1}\frac{\binom{m_{k-1}-1}{k-2}}{\binom{N}{k}} \;=\; \frac{(Nk+N+k)(N+1)(k-1)}{(k+2)(k+1)}. \tag{4.13}$$

Now that we have calculated the joint PDF, we know all the terms and find

$$\begin{aligned}\mathbb{E}[X_k \cdot X_{k-1}] &= \frac{(k+1)^2}{k(k-1)}\left[\frac{(Nk+N+k)(N+1)(k-1)}{(k+2)(k+1)}\right] \\ &\quad - \frac{k+1}{k}\frac{(N+1)k}{k+1} \;-\; \frac{k+1}{k-1}\frac{(N+1)(k-1)}{k+1} \;+\; 1,\end{aligned} \tag{4.14}$$

and from our values after scaling, we have

$$\begin{aligned}\mathbb{E}[X_k \cdot X_{k-1}] - \mathbb{E}[X_k] \cdot \mathbb{E}[X_{k-1}] &= \frac{(k+1)^2}{k}\frac{(N+2)(N+1)}{(k+2)} \;-\; \frac{(k+1)}{k}(N+1) \;-(N+1)^2 \\ &= \frac{(N+1)(N-k)}{k(k+2)}.\end{aligned} \tag{4.15}$$

4.2.2. *Finding optimal alpha.* Now that we've calculated $\mathrm{Var}(X_k)$, $\mathrm{Var}(X_{k-1})$, and $\mathrm{Cov}(X_k, X_{k-1})$, we find the optimal α value (that minimizes the variance of X_α) by taking the derivative. Let $\alpha_{k,k-1}$

denote the specific value of α that minimizes the variance. We have

$$\begin{aligned}\text{Var}(X_\alpha) &= \alpha^2\text{Var}(X_k) + (1-\alpha)^2\text{Var}(X_{k-1}) + 2\alpha(1-\alpha)\text{Cov}(X_k, X_{k-1}) \\ &= \alpha^2(\text{Var}(X_k) + \text{Var}(X_{k-1}) - 2\,\text{Cov}(X_k, X_{k-1})) \\ &\quad + 2\alpha(\text{Cov}(X_k, X_{k-1}) - \text{Var}(X_{k-1})) + \text{Var}(X_{k-1}). \end{aligned} \tag{4.16}$$

Taking the derivative with respect to α yields

$$\begin{aligned}\text{Var}(X_\alpha)' &= 2\alpha(\text{Var}(X_k) + \text{Var}(X_{k-1}) - 2\,\text{Cov}(X_k, X_{k-1}) \\ &\quad + 2\,(\text{Cov}(X_k, X_{k-1}) - \text{Var}(X_{k-1})). \end{aligned} \tag{4.17}$$

Because we want to find the optimal α value, we solve for $\text{Var}(X)' = 0$ (and of course also check the endpoints of $\alpha = 0$ or 1). After substituting in the formulas and doing some algebra, we see that the optimal alpha value is 1, which is in fact one of our endpoints and corresponds to only using the largest tank. Thus if we use both the largest and second largest values observed we do worse than just using the largest:

$$\alpha_{k,k-1} = \frac{\text{Var}(X_{k-1}) - \text{Cov}(X_k, X_{k-1})}{\text{Var}(X_k) + \text{Var}(X_{k-1}) - 2\,\text{Cov}(X_k, X_{k-1})} = 1. \tag{4.18}$$

Equation 4.18 is obtained by reorganizing the terms so that we get an expression for $a_{k,k-1}$ and substituting the expressions for $\text{Var}(X_{k-1}), \text{Var}(X_k), \text{Cov}(X_k, X_{k-1})$. Thus, after simplifying, we see that when α is 1, the variance of X_α is minimized. Similar calculations hold for other weighted combinations. Therefore, in the discrete one-dimensional German Tank problem where we sample without replacement, the formula using the largest tank is best.

5. Continuous one-dimensional Problem

We now explore our first generalization and consider a continuous one-dimensional analogue. In the original formulation, the serial numbers were integers drawn from 1 to N. We now consider a continuous version, where we select k tanks from the interval $[0, N]$ with N unknown. Our goal is to find a statistic to estimate N. We discuss the effectiveness of various statistics and compare the continuous formulas to the discrete ones; the scaling factor is the same in the continuous and discrete one-dimensional problems.

5.1. Formulas Using Largest and Second Largest Observations.

We begin by estimating using the largest and second largest tanks by using the CDF method to find the PDF. In the continuous case, to find the PDF, we take the derivative of the CDF.

5.1.1. *Formula from Largest.*

We first find the CDF by computing the probability that all are at most m_k:

$$\text{Prob}(M_k \le m_k) = \left(\frac{m_k}{N}\right)^k; \tag{5.1}$$

this is because in the continuous case we can view all k observations as independent, and the probability any is at most m_k is just m_k/N. Taking the derivative gives the PDF:

$$f(m_k) = PDF_{M_k}(m_k) = CDF_{M_k}(m_k)' = \frac{km_k^{k-1}}{N^k}. \tag{5.2}$$

Now that we have the PDF, we calculate the expected value of m_k:

$$\mathbb{E}[M_k] = \int_0^N m_k f(m_k)\,dm_k = \int_0^N m_k \frac{km_k^{k-1}}{N^k}\,dm_k = \frac{k}{k+1}\cdot N. \tag{5.3}$$

Thus we obtain

$$\widehat{N} \;=\; m_k \cdot \left(1 + \frac{1}{k}\right). \tag{5.4}$$

Remark 5.1. *Comparing the continuous formula to the discrete formula in equation 3.1, the only difference is that there is a (-1) in the discrete formula. This difference occurs because in the discrete case we sampled from tanks numbered from 1 to N while in the continuous case we sampled from the interval $[0, N]$.*

5.1.2. *Formula using Second Largest.* We first find the CDF by computing the probability that the second largest tank is at most m_{k-1}. There are two possibilities – all the tanks are less than m_{k-1}, or one tank (and there is $\binom{k}{1}$ ways to choose which of the k tanks that is) is larger than m_{k-1} and the rest are m_{k-1} or smaller. Thus

$$\begin{aligned} CDF_M(m_{k-1}) &= \text{Prob}(M_{k-1} \leq x) \\ &= \left(\frac{x}{N}\right)^k + \binom{k}{1}\left(\frac{x}{N}\right)^{k-1}\frac{N-x}{N}. \end{aligned} \tag{5.5}$$

We take the derivative of the CDF in order to find the PDF. After standard integration, we find

$$\widehat{N} \;=\; m_{k-1} \cdot \frac{k+1}{k-1}. \tag{5.6}$$

Remark 5.2. *As we saw in the discrete case, because the continuous and discrete formula are essentially the same formulas, the formula using the largest tank has the least variance.*

5.2. **Continuous weighted formula.** As we did in the discrete one-dimensional case, we see if constructing a statistic that is a linear combination of m_k and m_{k-1} does a better job estimating N. Similar to before, the best value is again when $\alpha = 1$, meaning that the formula using only m_k gives the least variance and there is no benefit to including m_{k-1}. The calculation is omitted as it is quite similar to the discrete weighted formula, except the continuous cases uses integrals instead of sums.

6. TWO-DIMENSIONAL DISCRETE GENERALIZATIONS

We now generalize to two dimensions, after which it will be easy to extend to higher dimensions. We look at the discrete and continuous cases in the square and the circle. We find for each problem which statistic gives the best estimate for N, and compare the formulas of the discrete two-dimensional and discrete one-dimensional square. If we use all the terms in the calculation, we would not get a closed form so we approximate using the main term and get formulas for fixed k and N tending to infinity.

6.1. **Square Problem.** We consider the case of the square from $(1, 1)$ to (N, N) as the natural generalization of the one-dimensional set $\{1, \ldots, N\}$. There are N^2 pairs, we select k of them without replacement. We call the two components the X and the Y list and use the pairs to find the best estimate for N. We look at two statistics: the largest number from the two lists, and a recursive method where we start with an estimate of N and use the largest L to estimate for N again until the value of our estimate for N stabilizes.

6.1.1. *Maximum from Lists:* The motivation of looking at the largest observed component in the two-dimensional square comes from the one-dimensional problem, where looking at the largest tank gave the most accurate estimation. To calculate the formula, we use the CDF method. The CDF method in

the discrete case is slightly different from the one in the continuous case. The statistic we look at is the largest observed component of the X list and Y list, which we denote by m. To calculate the probability that the largest observed component is exactly m, we calculate

$$\text{Prob}(M \leq m) \; - \; \text{Prob}(M \leq m-1) \tag{6.1}$$

to get the $\text{Prob}(\text{Max} = m)$. The PDF is computed similarly as in previous arguments, giving

$$\begin{aligned} \text{PDF}_M(m) &= \text{Prob}(M \leq m) \; - \; \text{Prob}(M \leq m-1) \\ &= \frac{\binom{m^2}{k}}{\binom{N^2}{k}} \; - \; \frac{\binom{(m-1)^2}{k}}{\binom{N^2}{k}}. \end{aligned} \tag{6.2}$$

Equation 6.2 is true because there are m^2 numbers of pairs in a m by m square and out of the m^2 choices, we are choosing k of them. Same applies to choosing $(m-1)^2$ numbers of pairs from a $m-1$ by $m-1$ square. Thus the expected value of M is

$$\begin{aligned} \mathbb{E}[M] &= \sum_{m=\lceil\sqrt{k}\rceil}^{N} m \cdot \text{PDF}_M(M=m) \; = \; \sum_{m=\lceil\sqrt{k}\rceil}^{N} \frac{m\binom{m^2}{k} \; - \; m\binom{(m-1)^2}{k}}{\binom{N^2}{k}} \\ &= \sum_{m=\lceil\sqrt{k}\rceil}^{N} \frac{m\binom{m^2}{k} \; - \; (m-1)\binom{(m-1)^2}{k}}{\binom{N^2}{k}} \; - \; \sum_{m=\lceil\sqrt{k}\rceil}^{N} \frac{\binom{(m-1)^2}{k}}{\binom{N^2}{k}} \\ &= \frac{N\binom{N^2}{k} \; - \; (\lceil\sqrt{k}\rceil - 1)\binom{(\lceil\sqrt{k}\rceil-1)^2}{k}}{\binom{N^2}{k}} \; - \; \sum_{m=\lceil\sqrt{k}\rceil}^{N} \frac{\binom{(m-1)^2}{k}}{\binom{N^2}{k}}. \end{aligned} \tag{6.3}$$

In Equation 6.3, we simplified the summation in the first term by telescoping. The value of m ranges from $\lceil\sqrt{k}\rceil$ to N because the minumum number of points we can have in the square is k. Thus, m has to be at least greater than $\lceil\sqrt{k}\rceil$ to have k points inside the square. We now determine the second term above; it is

$$\sum_{m=\lceil\sqrt{k}\rceil}^{N} \frac{\binom{(m-1)^2}{k}}{\binom{N^2}{k}} \;=\; \frac{1}{\binom{N^2}{k}} \sum_{m=\lceil\sqrt{k}\rceil}^{N-1} \binom{m^2}{k} \;=\; \frac{1}{k!\binom{N^2}{k}} \sum_{m=\lceil\sqrt{k}\rceil}^{N-1} m^2(m^2-1)\cdots(m^2-(k-1)) \tag{6.4}$$

Note that the range for m starts at $\lceil\sqrt{k}\rceil$ because if m is $\lceil\sqrt{k}\rceil - 1$, we get 0. We use Lemma 2.11 to provide upper and lower bounds for the sum:

$$\sum_{m=\lceil\sqrt{k}\rceil}^{N-1} \left(m^{2k} - m^{2k-2}\left(\frac{k(k-1)}{2}\right)\right) \leq \sum_{m=\lceil\sqrt{k}\rceil}^{N-1} m^2(m^2-1)\cdots(m^2-(k-1) \leq \sum_{m=\lceil\sqrt{k}\rceil}^{N-1} m^{2k}, \tag{6.5}$$

and we now use the Euler-Maclaurin formula, Lemma 2.10, to approximate the sums with integrals, and bound the error of the approximation in terms of the derivative of the function at the boundary points. We take $p = 2$, as this gives an excellent bound and an expression that is easy to work with.

Remark 6.1. *When we calculate the upper and lower bounds, we see that main term of the upper and lower are the same, which is why we can set these bounds.*

Now that we have applied Euler-Maclaurin on both the upper and lower bounds, we use the main term (which as $N \to \infty$ dominates the lower order terms) to find an equation of m in terms of N and k:

$$\begin{aligned} \sum_{m=\lceil\sqrt{k}\rceil}^{N-1} m^{2k} &\approx \frac{(N-1)^{2k+1} - (\lceil\sqrt{k}\rceil)^{2k+1}}{2k+1} \\ &\approx \frac{(N-1)^{2k+1}}{2k+1}. \end{aligned} \tag{6.6}$$

Because we assumed that k is fixed, if N is very large the other terms are negligible. We plug this estimation back into the formula for $\mathbb{E}[M]$:

$$\begin{aligned} \mathbb{E}[M] &= \frac{N\binom{N^2}{k} - (\lceil\sqrt{k}\rceil - 1)\binom{(\lceil\sqrt{k}\rceil-1)^2}{k}}{\binom{N^2}{k}} - \sum_{m=\lceil\sqrt{k}\rceil}^{N} \frac{\binom{(m-1)^2}{k}}{\binom{N^2}{k}} \\ &\approx N - \frac{\frac{(N-1)^{2k+1}}{2k+1}}{\binom{N^2}{k}\cdot k!} \approx N\left[1 - \frac{1}{2k+1}\left(1 - \frac{1}{N}\right)^{2k+1}\right] \approx N\left[\frac{2k}{2k+1}\right] + 1. \end{aligned} \tag{6.7}$$

We use the same argument as above to say that the term with k is negligible. We plug in our estimation for the second term and expand out the first two terms to estimate for $(1-1/N)^{2k+1}$. If we use more terms, the accuracy slightly increases, but the equation will not be invertible. Thus, we only write the first two terms. We obtain a good estimate of $(1-1/N)^{2k+1}$ by using two terms. Inverting the equation, we get

$$\widehat{N} = \frac{2k+1}{2k}(m-1). \tag{6.8}$$

Now that we have a estimation formula for $\widehat{N}$, we run some simulations to see how accurate it is. We see that the two-dimensional formula does well as the variance is small. In the next subsubsection, we compare the one-dimensional to the two-dimensional formula and see which one does better.

6.1.2. *Comparing Formulas.* We compare the one-dimensional formula and the discrete square formula to see which one does better. We have to make sure that we are making correct comparisons (apples to apples), because a pair gives two data points whereas a point gives one. Also, we want to make sure that both formulas estimate for N. For the N by N square, we pick k pairs, which gives us $2k$ components of N^2 pairs. For the one-dimensional case, we pick $2k$ tanks from N^2 possible tanks. This will give us a estimate for N^2, and we take the square root to find the estimate for N. By comparing these two quantities, we make sure that we observe $2k$ data points. In Figure 4, for the left trial, we set $k = 20$ and for the right trial, we set $k = 2$.

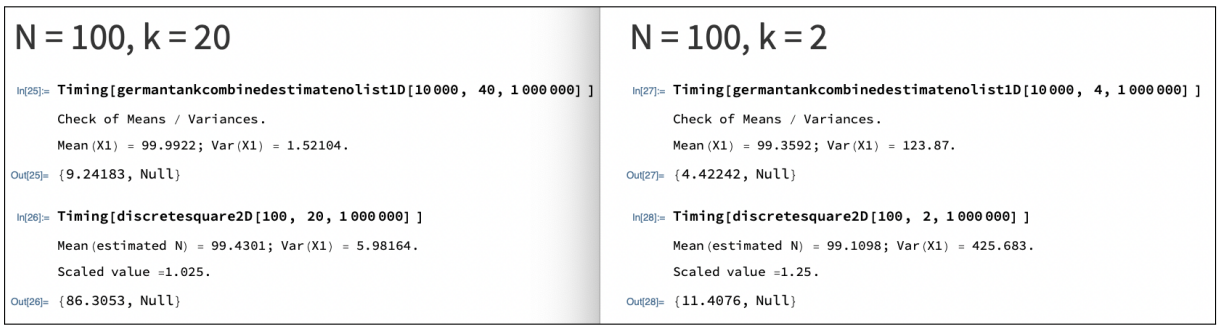

FIGURE 4. Comparing results from square formula to one-dimensional formula

Remark 6.2. *From the simulation, we see that the 1-dimensional case does a better job than the two-dimensional case, as the one-dimensional case has lower variance. The difference is clearly visible when k is a very small value such as 2. We have seen from Figure 4 that the 1-dimensional formula does better, and we confirm this by theory.*

6.1.3. *Recursive Argument:* For another approach, we start with an initial estimate for N, which we call N_0. We construct a formula to recursively generate new estimates of N from previous; thus let N_1 be our next guess. We investigate if the values of N converge, and if they converges to a more accurate estimate. We first transform the N^2 pairs into a list of numbers from 1 to N^2. For each pair (X,Y), we write the tanks as $(X-1) + N(Y-1) + 1$. This way, we can represent all tanks from 1 to N^2, as this maps the N^2 pairs uniquely to the integers from 1 to N^2. Unfortunately, when we use observed tank values to estimate, we do not know the value of N so we cannot immediately use this formula. Instead we replace N with our estimate. With that estimate, we express a new N with the largest observed component. As before, let M denote the maximum of the two lists. Then

$$\text{Prob}(M = L) \;=\; \frac{\binom{L-1}{k-1}}{\binom{N^2}{k}}. \tag{6.9}$$

Thus

$$\begin{aligned} \mathbb{E}[L] \;&=\; \sum_{L=k}^{N^2} L \cdot \frac{\binom{L-1}{k-1}}{\binom{N^2}{k}} \\ &\quad \text{We calculate and get} \\ &=\; k \cdot \frac{N^2+1}{k+1}. \end{aligned} \tag{6.10}$$

Thus

$$\widehat{N} \;=\; \sqrt{\mathbb{E}[L] \cdot \frac{k+1}{k} - 1}. \tag{6.11}$$

In order to create a iterative process, we plug in $\mathbb{E}[L] \;=\; \text{Max}X + (\text{current estimation for N})(\text{Max}Y - 1)$. We rewrite our estimation for $\widehat{N}$

$$\widehat{N} \;=\; \sqrt{\Big[\text{Max}X + (\text{current estimation for N})(\text{Max}Y - 1)\Big] \cdot \frac{k+1}{k} - 1}. \tag{6.12}$$

Now, we've got a recursive function of N. We can use the preliminary estimation of N to get another estimate for N. Therefore, by starting with a value of N, we can continue to produce estimates of N. The hope is that by producing more values of N, the values converge to the actual number N. To see how well this process does, we attached the code for simulation in the Appendix C. Using the simulation, we plugged in different values of Max(X) and Max(Y). However, though the results converge, they do not do a better job as often the value it converges to is off from the actual N.

6.1.4. ***Continuous Square Problem***. Unlike the discrete square problem where we could pick points from a discrete setting from $(1,1)$ to (N,N), in the continuous square problem, we can pick points from a continuous setting from $(0,0)$ to (N,N). To calculate the PDF, we use the CDF method by calculating the CDF and taking the derivative. Let m denote the largest component observed:

$$CDF_M(m) \;=\; \text{Prob}(M \le m) \;=\; \left(\frac{m^2}{N^2}\right)^k \;=\; \frac{m^{2k}}{N^{2k}}, \tag{6.13}$$

and thus

$$\mathrm{PDF}_M(m) \;=\; \mathrm{CDF_M}'(m) \;=\; \frac{2k\cdot m^{2k-1}}{N^{2m}}. \tag{6.14}$$

$$\text{Therefore, } \mathbb{E}[M] \;=\; \int_0^N m\cdot\frac{2k\cdot m^{2k-1}}{N^{2k}}\,dm \;=\; \frac{2k}{2k+1}\cdot N. \text{ Solving yields } \widehat{N} \;=\; m\cdot\frac{2k+1}{2k} \tag{6.15}$$

Remark 6.3. *We see that the scaling factor for the continuous case is the same as in the discrete case. The scaling factor of* $(2k+1)/2k$ *is reasonable, and we see this by comparing this formula to the one-dimensional formula. In the one-dimensional case, the scaling factor was* $(k+1)/k$*, which is larger than* $(2k+1)/2k$*. Because in the two-dimensional case we are looking at the largest of both components, we have more data points and therefore we will likely get a larger* M *value. Thus, in the two-dimensional case, we would have to scale by a value smaller than* $(k+1)/k$*, and scaling by* $(2k+1)/2k$ *makes sense.*

6.2. **Circle Problem.**

6.2.1. ***Discrete Circle Problem***. The goal of the discrete circle problem is to find a formula that estimates the radius. We assume the circle is centered at $(0,0)$ with radius r and we select k different lattice points contained in the circle without replacement. We look at X^2+Y^2 as our statistic, because the resulting values are integers, and we can then take a square-root at the end.

We let $m_1 = X^2+Y^2$. However, some elementary number theory enters in two dimensions and not all values of m_1 are attainable. Notice that X^2 and Y^2 are each 0 or 1 modulo 4, so any attainable m_1 is either 0, 1 or 2 modulo 4.

The number of lattice points inside a circle with radius r and center $(0,0)$ is the well studied Gauss Circle problem [Co]. Let $P(r)$ be the number of lattice points inside such a circle, i.e.,

$$P(r) \;:=\; \text{Number of } ((q,n)\in\mathbb{Z}^2 | q^2+n^2\le r^2). \tag{6.16}$$

The number of lattice points inside the circle is well estimated by the area of the circle, πr^2; the challenge is determining the size of the error. We have

$$P(r) \;=\; \pi r^2 \;+\; E(r). \tag{6.17}$$

We do not need the best known results, so we write $E(r)$ as $O(r^\delta)$ where $(0\le\delta\le 1)$, using Big-O notation (see Definition 2.12); the sharpest known bounds has $.5 < \delta < .63$ (see [Co]). Note $P(m) \;=\; \mathrm{CDF}_{M_1}(M\le m_1)$, where M is the random variable that is the value of X^2+Y^2. We see that $\mathrm{PDF}_M(m)$ is $P(m_1)-P(m_1-1)$. Therefore, if we have $m_1 \equiv 3 \pmod 4$, then $\mathrm{PDF}_M(m_1)=0$, and thus we do not have to worry about this case, though for completeness we do include it below.

We calculate the expected value of M:

$$\begin{aligned}
\mathbb{E}[M] &= \sum_{m_1=0}^{r^2} m_1\cdot\mathrm{Prob}(M=m_1)\\
&= [P(1)-P(0)]+[2P(2)-2P(1)]+\cdots+r^2[P(r^2)-r^2P(r^2-1)]\\
&= r^2P(r^2)-[P(1)+P(2)+\cdots+P(r^2-1)\\
&= r^2 \;-\; \frac{1}{\binom{\pi r^2+O(r^\delta)}{k}}\sum_{m_1=0}^{r^2}\binom{\pi(m-1)+O((m-1)^\delta)}{k}.
\end{aligned} \tag{6.18}$$

We estimate the second term of the equation above by using Lemma 2.11 and the main term. After applying Euler-Maclaurin on both sides, and using the main term to calculate the second term, we get:

$$\mathbb{E}[M] \approx r^2 \cdot \frac{k}{k+1} + 1, \text{ and thus } \widehat{r} = \sqrt{\frac{k+1}{k}(m_1 - 1)}. \tag{6.19}$$

Remark 6.4. *We analyze the formula from the discrete circle problem. The formula is quite interesting, because we have a square root involved, and unlike other cases, the continuous and discrete setting have different scaling factors. First, we have the square root of $\mathbb{E}[M]$ which is a value similar to $\sqrt{X^2+Y^2}$, and is similar to r. Also, the scaling factor for the discrete case is $\sqrt{(k+1)/k}$, which is similar to $(2k+1)/2k$ as we take the square root because taking the square root decreases the value by a little bit. Though the formula doesn't completely align with the continuous circle, this formula makes a lot of sense, and the difference likely results from the different statistics that we looked at for the discrete and continuous circle.*

6.2.2. ***Continuous Circle Problem***. The continuous circle problem has similar conditions as the discrete circle problem, but we can select any points contained in the circle; the points don't necessarily have to be lattice points. We approach the continuous circle problem similarly as the continuous square problem. We look at $m_2 = \sqrt{X^2+Y^2}$ because that is the formula for the radius. Let m_2 be the largest observed statistic, and M the corresponding random variable. Then

$$\text{Prob}(\sqrt{\text{X}^2+\text{Y}^2} \leq \text{m}_2) = \frac{(m_2^2\pi)^k}{(r^2\pi)^k} = \frac{m_2^{2k}}{r^{2k}}. \tag{6.20}$$

$$\text{Thus, } \text{PDF}_M(m_2) = \text{CDF}'_M(m_2) = \frac{2k \cdot m_2^{2k-1}}{r^{2k}}. \tag{6.21}$$

The calculation for the expected value is the same as the continuous square, so we omit it, and we find

$$\widehat{r} = m_2 \cdot \frac{2k+1}{2k}. \tag{6.22}$$

Remark 6.5. *We compare the continuous circle formula to the discrete circle formula. They don't look similar, as the discrete formula has a square root. Note that $m_1 = m_2^2$, by the values of statistics we look at. If we Taylor expand $\sqrt{1+1/k}$, we get $1 + 1/2k + \cdots$. In the discrete circle, as the value of k gets very large, $\sqrt{1+1/k}$ looks like $1 + 1/2k$, which is the formula for the continuous circle. Thus, though the formulas look different, if we take the limit as k gets large, we see how similar these are.*

We run some code to check the formula and see how well it does.

```
Timing[discretecircle[50, 15, 100000] ]
Mean(estimated radius) = 49.9825; Var(estimated radius) = 2.57532.
Actual Radius = 50 .
Scaled value =1.03333.
{17.0498, Null}
```

```
In[34]:= Timing[discretecircle[50, 5, 100000] ]
         Mean(estimated radius) = 49.9617; Var(estimated radius) = 20.9235.
         Actual Radius = 50 .
         Scaled value =1.1.
Out[34]= {7.00978, Null}
```

FIGURE 5. Code for discrete circle.

7. HIGHER DIMENSION VERSION

7.1. **Generalized Square Problem.**

7.1.1. *Discrete Square Problem:* To calculate the formula for the discrete L-dimensional square, we use similar strategies as the two-dimensional square. We let M be the the largest observed coordinate. Thus

$$\text{Prob}(M = m) \;=\; F(m) - F(m-1) \;=\; \frac{\binom{m^L}{k}}{\binom{N^L}{k}} \;-\; \frac{\binom{(m-1)^L}{k}}{\binom{N^L}{k}}. \tag{7.1}$$

The expected value is thus

$$\begin{aligned}
\mathbb{E}[M] \;&=\; \sum_{n=\lceil \sqrt[L]{k} \rceil}^{N} m \cdot \text{PDF}_M(m) \;=\; \sum_{n=\lceil \sqrt[L]{k} \rceil}^{N} \frac{m\binom{m^L}{k} - m\binom{(m-1)^L}{k}}{\binom{N^L}{k}} \\
&=\; \sum_{n=\lceil \sqrt[L]{k} \rceil}^{N} \frac{m\binom{m^L}{k} - (m-1)\binom{(m-1)^L}{k}}{\binom{N^L}{k}} \;-\; \sum_{n=\lceil \sqrt[L]{k} \rceil}^{N} \frac{\binom{(m-1)^L}{k}}{\binom{N^L}{k}} \\
&=\; \frac{1}{\binom{N^L}{k}} \left[N\binom{N^L}{k} - (\lceil \sqrt[L]{k} \rceil)\binom{\lceil \sqrt[L]{k} \rceil)^L}{k} \right] \;-\; \sum_{n=\lceil \sqrt[L]{k} \rceil}^{N} \frac{\binom{(m-1)^L}{k}}{\binom{N^L}{k}}.
\end{aligned} \tag{7.2}$$

To calculate the second term, we apply Lemma 2.11 to bound the sums, apply Euler-Maclaurin, and find the main terms to estimate. Though we are in higher dimensions, because this process is similar to the discrete two-dimensional square calculation, we only state the results.

$$\mathbb{E}[M] \;=\; N\left[\frac{Lk}{Lk+1}\right] \;+\; 1. \text{ Inverting, we get, } \widehat{N} \;=\; \frac{Lk+1}{Lk}(m-1). \tag{7.3}$$

7.1.2. *Continuous Square Problem:* The continuous problem is easily generalized to L dimensions. We select k tuples of length L and let M be the largest observed component:

$$\text{Prob}(M \le m) \;=\; \left(\frac{m^L}{N^L}\right)^k \;=\; \frac{m^{Lk}}{N^{Lk}}. \tag{7.4}$$

$$\text{PDF}_M(m) \;=\; \text{CDF}'_M(m) \;=\; \frac{Lk \cdot m^{Lk-1}}{N^{Lk}}. \tag{7.5}$$

Thus

$$\mathbb{E}[M] \;=\; \int_0^N m \cdot \frac{Lk \cdot m^{Lk-1}}{N^{Lk}}\, dm. \text{ Solving, we get } \widehat{N} \;=\; m \cdot \frac{Lk+1}{Lk}. \tag{7.6}$$

Remark 7.1. *We see that the scaling factors for the discrete L dimensional square continuous L dimensional square are the same. This scaling factor makes sense because if we have L dimensions, we have many more components to choose from. Thus by scaling by $(Lk+1)/Lk$, which is a value smaller than $(2k+1)/2k$, and is close to 1, we get a good estimate for N.*

7.2. **Generalized Circle Problem.**

7.2.1. *Discrete Circle Problem.* : We look at the generalized L dimensional circle problem. We study a statistic similar to the two-dimensional circle:

$$X_1^2 + X_2^2 + \cdots + X_L^2 \;=\; m_1. \tag{7.7}$$

We see that all values of m_1 are integers. Using this statistic, we estimate for r^2; after getting the formula, we take the square root of m_1 to estimate for r. Let $P(r)$ be the number of lattice points inside an L-ball, a L-dimensional sphere with radius r. We use the volume of the L-ball to find an approximate

value for the number of lattice points inside the L-ball (see [Gi]).

$$V(n) \;=\; \frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L. \tag{7.8}$$

We use this formula to find $P(r)$. We denote the bounds with Big-O notation:

$$P(r) \;=\; \frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L \;+\; O(r^\delta). \tag{7.9}$$

We now calculate the density using the CDF method. We calculate the expected value

$$\begin{aligned}
\mathbb{E}[M] \;&=\; \sum_{m_1=0}^{r^2} m_1 \cdot \mathrm{Prob}(M = m_1) \\
&=\; \frac{1}{\binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L + O(r^\delta)}{k}} \sum_{m_1=0}^{r^2} \left[m_1 \binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} m_1 \;+\; O(m_1^\delta)}{k} \right. \\
&\quad \left. - (m_1-1)\binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)}(m_1-1) \;+\; O((m_1-1)^\delta)}{k} \right] \\
&\quad - \frac{1}{\binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L + O(r^\delta)}{k}} \sum_{m_1=0}^{r^2} \binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)}(m_1-1) \;+\; O((m_1-1)^\delta)}{k} \\
&\quad \text{We telescope and get} \\
&=\; r^2 \;-\; \frac{1}{\binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L + O(r^\delta)}{k}} \sum_{m_1=0}^{r^2} \binom{\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)}(m_1-1) \;+\; O((m_1-1)^\delta)}{k}.
\end{aligned} \tag{7.10}$$

We calculate the summation part, using the bounds lemma and the main term to estimate. We have

$$\mathbb{E}[M] \;\approx\; r^2 \cdot \frac{k}{k+1} + 1. \text{ Inverting, we get } \widehat{r} \;=\; \sqrt{(m_1-1)\cdot\frac{k+1}{k}}. \tag{7.11}$$

Remark 7.2. *Notice that the formula for discrete L-dimensional circle problem is not dependent on N.*

7.2.2. ***Continuous L-dimensional circle problem.*** We select k tuples of length L where each component is contained in the L-dimensional circle. Let m be the largest observed component. We look at $m_2 = \sqrt{X_1^2 \;+\; X_2^2 \;+ \cdots + X_L^2}$ as our statistic. We use the volume of the L dimensional circle to calculate the PDF. Recall the formula of the L-ball. We see that

$$V(n) \;=\; \frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L. \tag{7.12}$$

We calculate the CDF:

$$\mathrm{Prob}\left(\sqrt{X_1^2 \;+\; X_2^2 \;+ \cdots + X_L^2} \le m_2\right) \;=\; \frac{(\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} m_2^L)^k}{(\frac{\pi^{\frac{L}{2}}}{\Gamma(\frac{L}{2}+1)} r^L)^k} \;=\; \frac{m_2^{Lk}}{r^{Lk}}. \tag{7.13}$$

The rest of the calculation is identical as the L-dimensional square problem, yielding

$$\hat{r} \;=\; m_2 \cdot \frac{Lk+1}{Lk}. \tag{7.14}$$

Remark 7.3. *Notice that the L dependence in the formula is small as the formula is* $1 + 1/Lk$ *and if L and k are reasonably large, then the scaling factor is close to 1.*

8. Conclusion and Further Direction

Through this research, we generalized the German Tank Problem into different extensions of the problem. First, we attempted to improve the original one−dimensional discrete problem, and concluded that the original formula using the largest tank produces the most accurate estimations with least variance. Then, we generalized the one-dimensional problem into the continuous case where we found that using the largest observed tank produces the most accurate formula. We then generalized into the two dimensional case where we looked at the discrete and continuous square and circle, and derived formulas for each case. Lastly, we generalized into the $L-$dimensional discrete and continuous square and circle and derived formulas. A possible further step in this research is to look at different shapes such as a hemisphere or an ellipse and select points inside different shapes.

9. Acknowledgements

We thank those who attended of the 2022 AISC conference and the PUMP undergraduate referee for many helpful comments.

Appendix A. Portfolio Theory

Consider two stocks X_1, X_2 with the same mean return and standard deviations σ_1 and σ_2; the variances are not necessarily equal. For simplicity we assume the two stocks' performances are independent, though in general we need to consider covariances. We construct a weighted portfolio $X_\alpha := \alpha X_1 + (1-\alpha)X_2$, with $\alpha \in [0,1]$. It is easy to see that the expected value of X_α is that of the two stocks; our goal is to find α that minimizes the variance of X_α and thus gives us the most certainty in knowing our future performance. Of course, such a strategy decreases the possibility of getting a larger than expected return, but it also minimizing the possibility of having a significantly smaller return.

Let's say we have two options. The first is we are guaranteed \$500,000. The second is we have a 50% chance of getting \$1,000,000, and a 50% chance of getting \$0 dollars. The expected value for both is \$500,000 dollars. Though it may depend on each person's financial situation, we see that taking the \$500,000 dollars has no risk. For some, this may be life changing (and the marginal utility of the second \$500,000 is almost surely less than the first).

The hypothetical situation above is a simple example of modern portfolio theory. This theory was pioneered by Markowitz; see [In]. A key idea of this theory is diversification. Because most investments are either high risk and high return or low risk and low return, Markowitz argued that perhaps investors could achieve best profit with acceptable risk by choosing an optimal mix of the investments.

We return to the simple case of two stocks both with mean μ, standard deviations σ_i (we may assume $0 \le \sigma_2 \le \sigma_1$) and we assume the two stocks are independent. Then if $X_\alpha = \alpha X_1 + (1-\alpha)X_2$ we have

$$\mathrm{Var}(X_\alpha) \;=\; \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2. \tag{A.1}$$

To find the minimum variance we check the endpoints ($\alpha = 0$ or 1) and the critical points from the derivative of the variance is zero: that happens when

$$2\alpha\sigma_1^2 - 2(1-\alpha)\sigma_2^2 \;=\; 0, \text{ which gives a critical value of } \alpha_* \;=\; \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}. \tag{A.2}$$

We now see which value of alpha gives the minimum variance.

- When $\alpha = 0$, $\mathrm{Var}(X_\alpha) = \sigma_2^2$.
- When $\alpha = 1$, $\mathrm{Var}(X_\alpha) = \sigma_1^2$.
- When $\alpha = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}$,

$$\mathrm{Var}(X_\alpha) = \left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2 \sigma_1^2 + \left(1-\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2 \sigma_2^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}. \tag{A.3}$$

From our assumption that $0 < \sigma_2 \leq \sigma_1$, we see that $\sigma_2^2 \leq \sigma_1^2$. Now, we compare $\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ and σ_2^2. Straightforward algebra shows

$$\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \quad < \quad \sigma_2^2; \tag{A.4}$$

to see this multiply both sides by $\sigma_1^2+\sigma_2^2$, and subtract $\sigma_1^2\sigma_2^2$ and obtain $0 < \sigma_2^4$. Thus the variance of the weighted quantity is always less than the variance of the smaller one! If the two variances are equal, the new variance is half of that.

Appendix B. Proof of Identities

Proof. **Identity I.**

$$\begin{aligned}
\sum_{m_k=k}^{N} \binom{m-b}{k-c} &= \binom{k-b}{k-b} + \binom{k-b+1}{k-b} + \cdots + \binom{N-b}{k-c} \\
&= \binom{k-c}{k-c} + \binom{k-c+1}{k-c} + \cdots + \binom{N-b}{k-c} \\
&\quad - \left[\binom{k-c}{k-c} + \binom{k-c+1}{k-c} + \cdots + \binom{k-b-1}{k-c}\right] \\
&= \binom{N-b+1}{k-c+1} - \binom{k-b}{k-c+1}.
\end{aligned} \tag{B.1}$$

□

Proof. **Identity II.**

$$\begin{aligned}
\sum_{m=k-a+1}^{N-a+1} m\frac{\binom{m-1}{k-a}\binom{N-m}{a-1}}{\binom{N}{k}} &= \frac{1}{\binom{N}{k}} \sum_{m=k-a+1}^{N-a+1} m \cdot \binom{m-1}{k-a}\binom{N-m}{a-1} \\
&= \frac{1}{\binom{N}{k}} \sum_{m=k-a+1}^{N-a+1} \frac{m!}{(k-a)!(m-k+a-1)!} \cdot \frac{(N-m)!}{(a-1)!(N-m-a+1)!} \\
&= \frac{k-a+1}{\binom{N}{k}} \cdot \sum_{m=k-a+1}^{N-a+1} \binom{m}{k-a+1} \cdot \binom{N-m}{a-1} \\
&\quad \text{We use Identity IV} \\
&= \frac{(N+1)(k-a+1)}{k+1}
\end{aligned} \tag{B.2}$$

□

Proof. **Identity III.**

$$\begin{aligned}
\sum_{m=k-a+1}^{N-a+1} m^2 \frac{\binom{m-1}{k-a}\binom{N-m}{a-1}}{\binom{N}{k}} &= \frac{1}{\binom{N}{k}} \sum_{m=k-a+1}^{N-a+1} (m+1)m\binom{m-1}{k-a}\binom{N-m}{a-1} \\
&\quad - \frac{1}{\binom{N}{k}} \sum_{m=k-a+1}^{N-a+1} m\binom{m-1}{k-a}\binom{N-m}{a-1} \\
&= \frac{1}{\binom{N}{k}} \sum_{m=k-a+1}^{N-a+1} \frac{(m+1)!}{(k-a)!(m-k+a-1)!}\binom{N-m}{a-1} \\
&\quad - \frac{(N+1)(k-a+1)}{k+1} \\
&= \frac{(k-a+1)(k-a+2)(N+2)(N+1)}{(k+2)(k+1)} \\
&\quad - \frac{(N+1)(k-a+1)}{k+1} \qquad \text{(B.3)}
\end{aligned}$$

□

Proof. **Identity IV.**

$$\binom{a+b+k+1}{a+b+1} = \sum_{i=0}^{k} \binom{a+i}{a}\binom{b+k-i}{b}. \qquad \text{(B.4)}$$

We use proof by strong induction to prove the identity. Case where $k = 0$:

$$\binom{a+b+1}{a+b+1} = \binom{a}{a}\binom{b}{b} = 1 \qquad \text{(B.5)}$$

Case where k=1:

$$\binom{a+b+2}{a+b+1} = \binom{a}{a}\binom{b+1}{b} + \binom{a+1}{a}\binom{b}{b} = a+b+2. \qquad \text{(B.6)}$$

We continue through all k until $k = k$ and we assume that the case where $k = k$ is true. We use this assumption to prove that the case works for $(k+1)$. We add all the equations when $k = k$ to when $k = 1$. We write the sum of equations as the following.

$$\binom{a}{a}\binom{b+k}{b} + \binom{a+1}{a}\binom{b+k-1}{b} + \cdots + \binom{a+k}{a}\binom{b}{b} = \binom{a+b+k+1}{a+b+1}. \qquad \text{(B.7)}$$

$$\binom{a}{a}\binom{b+k-1}{b} + \binom{a+1}{a}\binom{b+k-2}{b} + \cdots + \binom{a+k-1}{a}\binom{b}{b} = \binom{a+b+k}{a+b+1}. \qquad \text{(B.8)}$$

$$\binom{a}{a}\binom{b+k-2}{b} + \binom{a+1}{a}\binom{b+k-3}{b} + \cdots + \binom{a+k-2}{a}\binom{b}{b} = \binom{a+b+k-1}{a+b+1}. \qquad \text{(B.9)}$$

$$\binom{a}{a}\binom{b+1}{b} + \binom{a+1}{a}\binom{b}{b} = \binom{a+b+2}{a+b+1}. \tag{B.10}$$

$$\binom{a}{a}\binom{b}{b} = \binom{a+b+1}{a+b+1}. \tag{B.11}$$

We add all the columns.

$$\begin{aligned} & \binom{a}{a}\binom{b+k}{b} + \binom{a+1}{a}\binom{b+k-1}{b} \\ & + \cdots + \binom{a+k}{a}\binom{b}{b} \\ = \quad & \binom{a+b+k+2}{a+b+2}. \end{aligned} \tag{B.12}$$

The resulting equation is the equation in the $(k+1)$ case. Since we assumed by strong induction that cases when $k = 1$ to $k = k$ is all true, we are able to prove that the $(k+1)$ case is true. Therefore, the identity is proved. $\square$

APPENDIX C. MATHEMATICA CODE

One Dimensional German Tank Problem code

```
germantankcombinedestimatenolist[NumTanks_, k_, numdo_] := Module[{},
  tanks = {}; (* store list of tanks here *)
  For[i = 1, i ≤ NumTanks, i++, tanks = AppendTo[tanks, i]];
  sumX1 = 0; (* save prediction from largest tank here *)
  sumX1sq = 0; (* saves sum of squares *)
  alpha = 1; (* add formula here - function of k and NumTanks*)
  For[n = 1, n ≤ numdo, n++,
   {observedtanks = RandomSample[tanks, k]; (* uniformly at random chooses k tanks from 1 to NumTanks*)
    observedtanks = Sort[observedtanks]; (* sorts list*)
    largest = observedtanks[[-1]]; (*largest tank*)
    X1 = largest * ((k + 1.0) / k) - 1;
    sumX1 = sumX1 + X1;
    sumX1sq = sumX1sq + X1^2;
   }]; (* end of n loop*)
  Print["German Tank Problem Calculations: N = ", NumTanks, ", k = ", k, "."];
  Print["Check of Means / Variances."];
  Print["Mean(X1) = ", sumX1 1.0 / numdo, "; Var(X1) = ", (sumX1sq - sumX1^2 / numdo) / (numdo - 1.0), "."];
 ] (* end of module*)


Timing[germantankcombinedestimatenolist[400, 60, 1 000 000] ]

German Tank Problem Calculations: N = 400, k = 60.

Check of Means / Variances.

Mean(X1) = 399.989; Var(X1) = 36.781.

{5.499, Null}

Timing[germantankcombinedestimatenolist[4000, 60, 1 000 000] ]

German Tank Problem Calculations: N = 4000, k = 60.

Check of Means / Variances.

Mean(X1) = 3999.99; Var(X1) = 4245.42.

{8.69669, Null}

Timing[germantankcombinedestimatenolist[100, 2, 1 000 000] ]

German Tank Problem Calculations: N = 100, k = 2.

Check of Means / Variances.

Mean(X1) = 100.052; Var(X1) = 1234.92.

{3.5706, Null}

Timing[germantankcombinedestimatenolist[10 000, 2, 1 000 000] ]

German Tank Problem Calculations: N = 10 000, k = 2.

Check of Means / Variances.

Mean(X1) = 10 000.6; Var(X1) = 1.25123 × 10^7.

{3.89604, Null}
```

FIGURE 6. Simulation for one dimensional discrete case.

Two Dimensional Circle Code

```
discretecircle[radius_, k_, numdo_] := Module[{},
  sumestimatedradius = 0; (* save prediction from largest tank here *)
  sumestimatedradiussq = 0; (* saves sum of squares *)
iterations = 0;
  For[n = 1, n ≤ numdo, n++,
   count = 0;
R = radius;
list = {};
iterations = 0;
   {While[count < k, {iterations = iterations + 1;
      x = RandomInteger[{-R, R}]; (*Select pairs*)
      y = RandomInteger[{-R, R}]; (*Select pairs*)
      If[x^2 + y^2 ≤ R^2 && MemberQ[list, {x, y}] == False, {list = AppendTo[list, {x, y}];
        count = count + 1;}];(*end of if statement*)}];
    findingmax = Max[Table[Total[list[[i]]^2], {i, 1, Length[list]}]];
    largestpair = Flatten[Select[list, #[[1]]^2 + #[[2]]^2 == findingmax &]]; (*Finding the largest pair*)
    beforescaled = 1.0 Sqrt[largestpair[[1]]^2 + largestpair[[2]]^2];
    estimatedradius = 1.0 Sqrt[largestpair[[1]]^2 + largestpair[[2]]^2] (2 k + 1) / (2 k);
    (*Making estimates for the radius*)
    sumestimatedradius = sumestimatedradius + estimatedradius;
    sumestimatedradiussq = sumestimatedradiussq + estimatedradius^2;
   }]; (* end of n loop*)

  Print["Mean(estimated radius) = ", sumestimatedradius 1.0 / numdo, "; Var(X1) = ",
   (sumestimatedradiussq - sumestimatedradius^2 / numdo) / (numdo - 1.0), "."];
  Print["Actual Radius = ", radius "."];
  Print["Scaled value =", 1.0 × (2 k + 1) / (2 k), "."];

 ] (* end of module*)
discretecircle[50, 5, 1000]
Mean(estimated radius) = 49.8962; Var(X1) = 22.5574.
Actual Radius = 50 .
Scaled value =1.1.
discretecircle[100, 15, 1000]
Mean(estimated radius) = 99.9577; Var(X1) = 10.8644.
Actual Radius = 100 .
Scaled value =1.03333.
discretecircle[100, 5, 1000]
Mean(estimated radius) = 99.6659; Var(X1) = 85.0634.
Actual Radius = 100 .
Scaled value =1.1.
```

FIGURE 7. Simulation for two dimensional discrete circle.

Two Dimensional Discrete Square Simulation

```
discretesquare2D[NumTanks_, k_, numdo_] := Module[{},
  sumestimatednvalue = 0; (* save prediction from largest tank here *)
  sumestimatednvaluesq = 0; (* saves sum of squares *)
iterations = 0;
  For[n = 1, n ≤ numdo, n++,
   count = 0;
list = {};
iterations = 0;
   {While[count < k, {iterations = iterations + 1;
      x = RandomInteger[{0, NumTanks}]; (*Select pairs*)
      y = RandomInteger[{0, NumTanks}]; (*Select pairs*)
      If[MemberQ[list, {x, y}] == False, {list = AppendTo[list, {x, y}];
        count = count + 1;}];(*end of if statement*)}];
    maxofcomponents = Max[list];
    estimatednvalue = 1.0 (maxofcomponents - 1) (2 k + 1) / (2 k);
    sumestimatednvalue = sumestimatednvalue + estimatednvalue;
    sumestimatednvaluesq = sumestimatednvaluesq + estimatednvalue^2;
   }]; (* end of n loop*)

  Print["Mean(estimated N) = ", sumestimatednvalue 1.0 / numdo, "; Var(X1) = ",
   (sumestimatednvaluesq - sumestimatednvalue^2 / numdo) / (numdo - 1.0), "."];
  Print["Scaled value =", 1.0 × (2 k + 1) / (2 k), "."];

 ] (* end of module*)

Timing[discretesquare2D[100, 10, 100 000] ]

Mean(estimated N) = 99.4054; Var(X1) = 23.1241.

Scaled value =1.05.

{4.29191, Null}

Timing[discretesquare2D[100, 3, 100 000] ]

Mean(estimated N) = 99.2409; Var(X1) = 212.373.

Scaled value =1.16667.

{1.53073, Null}

Timing[discretesquare2D[400, 60, 100 000] ]

Mean(estimated N) = 399.44; Var(X1) = 11.0866.

Scaled value =1.00833.

{30.4782, Null}
```

FIGURE 8. Simulation for Two Dimensional Discrete Square.

Recursive Method

```
n = 100;
maxX = 60;
maxY = 65;
k = 10;
nextN[maxX_, maxY_, currN_, k_] := Sqrt[(maxX + currN (maxY - 1)) (k + 1) / k - 1]
currN = 1.0 Max[maxX, maxY] * (k + 1) / k - 1;
Print["Initial guess is ", currN];
For[i = 1, i ≤ 10, i ++, { newNis = 1.0 nextN[maxX, maxY, currN, k];
  currN = newNis;
  Print["Guess ", i, " is ", currN];}]
Initial guess is 70.5
Guess 1 is 70.9098
Guess 2 is 71.1129
Guess 3 is 71.2134
Guess 4 is 71.2631
Guess 5 is 71.2876
Guess 6 is 71.2997
Guess 7 is 71.3057
Guess 8 is 71.3086
Guess 9 is 71.3101
Guess 10 is 71.3108
```

FIGURE 9. Simulation for recursion method for square.

```
n = 100;
maxX = 40;
maxY = 80;
k = 10;
nextN[maxX_, maxY_, currN_, k_] := Sqrt[(maxX + currN (maxY - 1)) (k + 1) / k - 1]
currN = 1.0 Max[maxX, maxY] * (k + 1) / k - 1;
Print["Initial guess is ", currN];
For[i = 1, i ≤ 10, i ++, { newNis = 1.0 nextN[maxX, maxY, currN, k];
  currN = newNis;
  Print["Guess ", i, " is ", currN];}]
Initial guess is 87.
Guess 1 is 87.1969
Guess 2 is 87.295
Guess 3 is 87.3438
Guess 4 is 87.368
Guess 5 is 87.3801
Guess 6 is 87.3861
Guess 7 is 87.3891
Guess 8 is 87.3906
Guess 9 is 87.3913
Guess 10 is 87.3917
```

FIGURE 10. Simulation for two dimensional discrete circle.

Code for Comparing Formulas for one and two dimensions

One Dimensional

```
germantankcombinedestimatenolist1D[NumTanks_, k_, numdo_] := Module[{},
  tanks = {}; (* store list of tanks here *)
  For[i = 1, i ≤ NumTanks, i++, tanks = AppendTo[tanks, i]];
  sumX1 = 0; (* save prediction from largest tank here *)
  sumX1sq = 0; (* saves sum of squares *)
  alpha = 1; (* add formula here - function of k and NumTanks*)
  For[n = 1, n ≤ numdo, n++,
   {observedtanks = RandomSample[tanks, k]; (* uniformly at random chooses k tanks from 1 to NumTanks*)
    observedtanks = Sort[observedtanks]; (* sorts list*)
    largest = observedtanks[[-1]]; (*largest tank*)
    X1 = Sqrt[largest * ((k + 1.0) / k) - 1]; (*We take the sqrt to estimate for N instead of N^2*)
    sumX1 = sumX1 + X1;
    sumX1sq = sumX1sq + X1^2;
   }]; (* end of n loop*)
  Print["Check of Means / Variances."];
  Print["Mean(X1) = ", sumX1 1.0 / numdo, "; Var(X1) = ", (sumX1sq - sumX1^2 / numdo) / (numdo - 1.0), "."];
 ] (* end of module*)
```

Two Dimensional

```
In[35]:= discretesquare2D[NumTanks_, k_, numdo_] := Module[{},
  sumestimatednvalue = 0; (* save prediction from largest tank here *)
  sumestimatednvaluesq = 0; (* saves sum of squares *)
iterations = 0;
  For[n = 1, n ≤ numdo, n++,
   count = 0;
list = {};
iterations = 0;
   {While[count < k, {iterations = iterations + 1;
      x = RandomInteger[{0, NumTanks}]; (*Select pairs*)
      y = RandomInteger[{0, NumTanks}]; (*Select pairs*)
      If[MemberQ[list, {x, y}] == False, {list = AppendTo[list, {x, y}];
        count = count + 1;}];(*end of if statement*)}];
    maxofcomponents = Max[list];
    estimatednvalue = 1.0 (maxofcomponents - 1) (2 k + 1) / (2 k);
    sumestimatednvalue = sumestimatednvalue + estimatednvalue;
    sumestimatednvaluesq = sumestimatednvaluesq + estimatednvalue^2;
   }]; (* end of n loop*)

  Print["Mean(estimated N) = ", sumestimatednvalue 1.0 / numdo, "; Var(X1) = ",
   (sumestimatednvaluesq - sumestimatednvalue^2 / numdo) / (numdo - 1.0), "."];
  Print["Scaled value =", 1.0 × (2 k + 1) / (2 k), "."];

 ] (* end of module*)
```

FIGURE 11. Code for Comparing Formulas for one and two dimensions.

Results of comparison

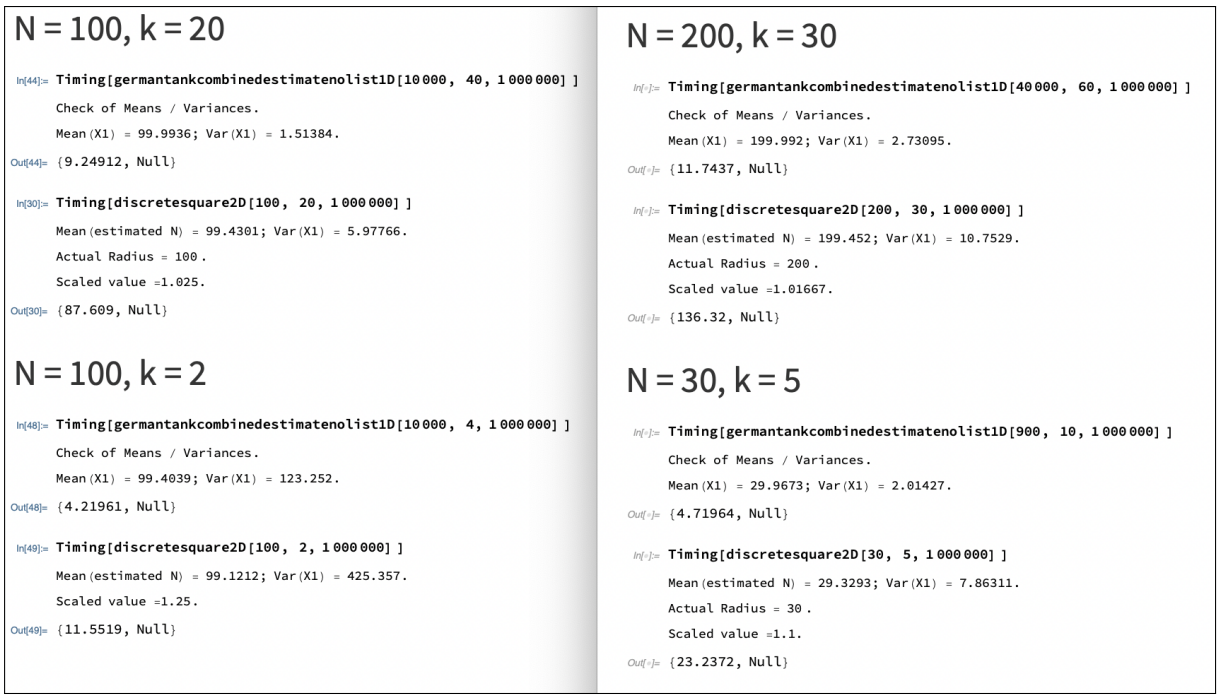

FIGURE 12. Results of comparison.

REFERENCES

[Al] The Albert Team. (2022, March 1). The German Tank Problem Explained: AP® Statistics Review. Albert Resources. Retrieved September 2, 2022, from https://www.albert.io/blog/german-tank-problem-explained-ap-statistics-review/

[CGM] G. Clark, A. Gonye, and S. J. Miller, *Lessons From the German Tank Problem*. https://arxiv.org/pdf/2101.08162.pdf.

[Co] P. Constantinescu (n.d.). The Gauss Circle problem - github pages. The Gauss circle problem. Retrieved September 4, 2022, from https://petruconstantinescu.github.io/The_Gauss_circle_problem.pdf.

[Gi] J. Gipple, *The Volume of n-balls*, Rose-Hulman Undergraduate Mathematics Journal **15** (2014), Issue 1, Article 14.

[H] A. Hayes, *Portfolio variance*, Investopedia (2022, February 8), retrieved September 2, 2022 from https://www.investopedia.com/terms/p/portfolio-variance.asp.

[In] The Investment Team, *Modern portfolio theory (MPT)*, Investopedia (2022, June 14), retrieved September 4, 2022 from https://www.investopedia.com/terms/m/modernportfoliotheory.asp.

[Ka] M. C. Kalu, *How the allies guessed the number of German tanks using serial numbers*, warhistoryonline (2019, September 18), retrieved September 2, 2022 from https://www.warhistoryonline.com/instant-articles/the-german-tank-problem.html.

[Ku] K. Kuter, *Chapter 4.1: Probability density functions (pdfs) and cumulative distribution functions (cdfs) for continuous random variables*, Statistics LibreTexts (2021, March 9), retrieved September 4, 2022 from https://tinyurl.com/2dchr2v8.

[Mi] S. J. Miller, *The Probability Lifesaver: All the Tools You Need to Understand Chance*, Princeton University Press (2017), https://doi.org/10.2307/j.ctvc7767n.

[Ru] R. Ruggles and H. Brodie, *An Empirical Approach to Economic Intelligence in World War II*, Journal of the American Statistical Association **42** (1947), no. 237, 72-–91. https://doi.org/10.2307/2280189.

[Sh] S. Sheffield, S. *18.600: Lecture 24 .1in covariance and some conditional expectation ...*, retrieved September 2, 2022 from https://math.mit.edu/~sheffield/2017600/Lecture24.pdf.

[Si] C. M. Simon, *A Bayesian treatment of the German tank problem*, preprint 2023, https://arxiv.org/pdf/2301.00046.pdf.

[YK] V. Kac and K. Yessenov, *18.704 Seminar in Algebra and Number Theory: Euler-Maclaurin Formula* (Fall 2015), retrieved September 1, 2022 from https://people.csail.mit.edu/kuat/courses/euler-maclaurin.pdf.

Email address: anthony_lee24@milton.edu

MILTON ACADEMY, MILTON, MA 02186

Email address: sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267