WIKIPEDIA
The Free Encyclopedia

German tank problem

In the statistical theory of estimation, the **German tank problem** consists of estimating the maximum of a discrete uniform distribution from sampling without replacement. In simple terms, suppose there exists an unknown number of items which are sequentially numbered from 1 to *N*. A random sample of these items is taken and their sequence numbers observed; the problem is to estimate *N* from these observed numbers.

The problem can be approached using either frequentist inference or Bayesian inference, leading to different results. Estimating the population maximum based on a *single* sample yields divergent results, whereas estimation based on *multiple* samples is a practical estimation question whose answer is simple (especially in the frequentist setting) but not obvious (especially in the Bayesian setting).

The problem is named after its historical application by Allied forces in World War II to the estimation of the monthly rate of German tank production from very limited data. This exploited the manufacturing practice of assigning and attaching ascending sequences of serial numbers to tank components (chassis, gearbox, engine, wheels), with some of the tanks eventually being captured in battle by Allied forces.

During World War II, production of German tanks such as the Panther was accurately estimated by Allied intelligence using statistical methods

Suppositions

The adversary is presumed to have manufactured a series of tanks marked with consecutive whole numbers, beginning with serial number 1. Additionally, regardless of a tank's date of manufacture, history of service, or the serial number it bears, the distribution over serial numbers becoming revealed to analysis is uniform, up to the point in time when the analysis is conducted.

Example

Assuming tanks are assigned sequential serial numbers starting with 1, suppose that four tanks are captured and that they have the serial numbers: 19, 40, 42 and 60.

The *frequentist* approach predicts the total number of tanks produced will be:

$$N \approx 74$$

The *Bayesian* approach predicts that the median number of tanks produced will be very similar to the frequentist prediction:

$$N_{med} \approx 74.5$$

whereas the Bayesian mean predicts that the number of tanks produced would be:

$$N_{av} \approx 89$$

Let N equal the total number of tanks predicted to have been produced, m equal the highest serial number observed and k equal the number of tanks captured.

The frequentist prediction is calculated as:

$$N \approx m + \frac{m}{k} - 1 = 74$$

The Bayesian median is calculated as:

$$N_{med} \approx m + \frac{m \ln(2)}{k - 1} = 74.5$$

The Bayesian mean is calculated as:

$$N_{av} \approx (m - 1)\frac{k - 1}{k - 2} = 89$$

Both Bayesian computations are based on the following probability mass function:

$$\Pr(N=n) = \begin{cases} 0 & \text{if } n < m \\ \frac{k-1}{k}\frac{\binom{m-1}{k-1}}{\binom{n}{k}} & \text{if } n \geq m, \end{cases}$$

This distribution has a positive skewness, related to the fact that there are at least 60 tanks. Because of this skewness, the mean may not be the most meaningful estimate. The median in this example is 74.5, in close agreement with the frequentist formula. Using Stirling's approximation, the Bayesian probability function may be approximated as

$$\Pr(N=n) \approx \begin{cases} 0 & \text{if } n < m \\ (k-1)m^{k-1}n^{-k} & \text{if } n \geq m, \end{cases}$$

which results in the following approximation for the median:

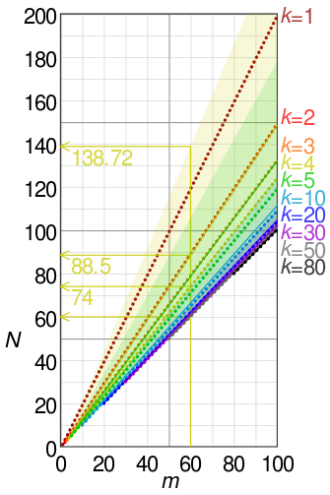
$$N_{med} \approx m + \frac{m\ln(2)}{k-1}$$

Finally, the average estimate by Bayesians, and its deviation, are computed as:

$$N \approx \mu \pm \sigma = 89 \pm 50,$$

$$\mu = (m-1)\frac{k-1}{k-2},$$

$$\sigma = \sqrt{\frac{(k-1)(m-1)(m-k+1)}{(k-3)(k-2)^2}}.$$



Estimated population size (N). The number of observations in the sample is *k*. The largest sample serial number is *m*. Frequentist analysis is shown with dotted lines. Bayesian analysis has solid yellow lines with mean and shading to show range from minimum possible value to mean plus 1 standard deviation). The example shows if four tanks are observed and the highest serial number is "60", frequentist analysis predicts 74, whereas Bayesian analysis predicts a mean of 88.5 and standard deviation of 138.72 − 88.5 = 50.22, and a minimum of 60 tanks. In the SVG file (https://upload.wikimedia.org/wikipedia/commons/3/3e/German_tank_problem_graphs.svg), hover over a graph to highlight it.

Historical example of the problem

During the course of the Second World War, the Western Allies made sustained efforts to determine the extent of German production and approached this in two major ways: conventional intelligence gathering and statistical estimation. In many cases, statistical analysis substantially improved on conventional intelligence. In some cases, conventional intelligence was used in conjunction with statistical methods, as was the case in estimation of Panther tank production just prior to D-Day.

The allied command structure had thought the Panzer V (Panther) tanks seen in Italy, with their high velocity, long-barreled 75 mm/L70 guns, were unusual heavy tanks and would only be seen in northern France in small numbers, much the same way as the Tiger I was seen in Tunisia. The US Army was confident that the Sherman tank would continue to perform well, as it had versus the Panzer III and Panzer IV tanks in North Africa and Sicily.[a] Shortly before D-Day, rumors indicated that large numbers of Panzer V tanks were being used.

Panther tanks are loaded for transport to frontline units, 1943

To determine whether this was true, the Allies attempted to estimate the number of tanks being produced. To do this, they used the serial numbers on captured or destroyed tanks. The principal numbers used were gearbox numbers, as these fell in two unbroken sequences. Chassis and engine numbers were also used, though their use was more complicated. Various other components were used to cross-check the analysis. Similar analyses were done on wheels, which were observed to be sequentially numbered (i.e., 1, 2, 3, ..., *N*).[2][b][3][4]

The analysis of tank wheels yielded an estimate for the number of wheel molds that were in use. A discussion with British road wheel makers then estimated the number of wheels that could be produced from this many molds, which yielded the number of tanks that were being produced each month. Analysis of wheels from two tanks (32 road wheels each, 64 road wheels total) yielded an estimate of 270 tanks produced in February 1944, substantially more than had previously been suspected.[5]

German records after the war showed production for the month of February 1944 was 276.[6][c] The statistical approach proved to be far more accurate than conventional intelligence methods, and the phrase "German tank problem" became accepted as a descriptor for this type of statistical analysis.

Estimating production was not the only use of this serial-number analysis. It was also used to understand German production more generally, including number of factories, relative importance of factories, length of supply chain (based on lag between production and use), changes in production, and use of resources such as rubber.

Specific data

According to conventional Allied intelligence estimates, the Germans were producing around 1,400 tanks a month between June 1940

and September 1942. Applying the formula below to the serial numbers of captured tanks, the number was calculated to be 246 a month. After the war, captured German production figures from the ministry of Albert Speer showed the actual number to be 245.[3]

Estimates for some specific months are given as:[7]

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

Similar analyses

Similar serial-number analysis was used for other military equipment during World War II, most successfully for the V-2 rocket.[8]

V-2 rocket production was accurately estimated by statistical methods

Factory markings on Soviet military equipment were analyzed during the Korean War, and by German intelligence during World War II.[9]

In the 1980s, some Americans were given access to the production line of Israel's Merkava tanks. The production numbers were classified, but the tanks had serial numbers, allowing estimation of production.[10]

The formula has been used in non-military contexts, for example to estimate the number of Commodore 64 computers built, where the result (12.5 million) matches the low-end estimates.[11]

Countermeasures

To confound serial-number analysis, serial numbers can be excluded, or usable auxiliary information reduced. Alternatively, serial numbers that resist cryptanalysis can be used, most effectively by randomly choosing numbers without replacement from a list that is much larger than the number of objects produced, or by producing random numbers and checking them against the list of already assigned numbers; collisions are likely to occur unless the number of digits possible is more than twice the number of digits in the number of objects produced (where the serial number can be in any base); see birthday problem.[d] For this, a cryptographically secure pseudorandom number generator may be used. All these methods require a lookup table (or breaking the cypher) to back out from serial number to production order, which complicates use of serial numbers: a range of serial numbers cannot be recalled, for instance, but each must be looked up individually, or a list generated.

Alternatively, sequential serial numbers can be encrypted with a simple substitution cipher, which allows easy decoding, but is also easily broken by frequency analysis: even if starting from an arbitrary point, the plaintext has a pattern (namely, numbers are in sequence). One example is given in Ken Follett's novel *Code to Zero*, where the encryption of the Jupiter-C rocket serial numbers is given by:

H	U	N	T	S	V	I	L	E	X
1	2	3	4	5	6	7	8	9	0

The code word here is Huntsville (with repeated letters omitted) to get a 10-letter key.[12] The rocket number 13 was therefore "HN", and the rocket number 24 was "UT".

Frequentist analysis

Minimum-variance unbiased estimator

For point estimation (estimating a single value for the total, $\widehat{N}$), the minimum-variance unbiased estimator (MVUE, or UMVU estimator) is given by:[e]

$$\widehat{N} = m(1 + k^{-1}) - 1,$$

where m is the largest serial number observed (sample maximum) and k is the number of tanks observed (sample size).[10][13] Note that once a serial number has been observed, it is no longer in the pool and will not be observed again.

This has a variance[10]

$$\operatorname{var}\left(\widehat{N}\right) = \frac{1}{k}\frac{(N-k)(N+1)}{(k+2)} \approx \frac{N^2}{k^2} \text{ for small samples } k \ll N,$$

so the standard deviation is approximately N/k, the expected size of the gap between sorted observations in the sample.

The formula may be understood intuitively as the sample maximum plus the average gap between observations in the sample, the sample maximum being chosen as the initial estimator, due to being the maximum likelihood estimator,[f] with the gap being added to compensate for the negative bias of the sample maximum as an estimator for the population maximum,[g] and written as

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

This can be visualized by imagining that the observations in the sample are evenly spaced throughout the range, with additional observations just outside the range at 0 and $N + 1$. If starting with an initial gap between 0 and the lowest observation in the sample (the sample minimum), the average gap between consecutive observations in the sample is $(m-k)/k$; the $-k$ being because the observations themselves are not counted in computing the gap between observations.[h]. A derivation of the expected value and the variance of the sample maximum are shown in the page of the discrete uniform distribution.

This philosophy is formalized and generalized in the method of maximum spacing estimation; a similar heuristic is used for plotting position in a Q–Q plot, plotting sample points at $k / (n + 1)$, which is evenly on the uniform distribution, with a gap at the end.

Confidence intervals

Instead of, or in addition to, *point* estimation, *interval* estimation can be carried out, such as confidence intervals. These are easily computed, based on the observation that the probability that k observations in the sample will fall in an interval covering p of the range ($0 \le p \le 1$) is p^k (assuming in this section that draws are *with* replacement, to simplify computations; if draws are without replacement, this overstates the likelihood, and intervals will be overly conservative).

Thus the sampling distribution of the quantile of the sample maximum is the graph $x^{1/k}$ from 0 to 1: the p-th to q-th quantile of the sample maximum m are the interval $[p^{1/k}N, q^{1/k}N]$. Inverting this yields the corresponding confidence interval for the population maximum of $[m/q^{1/k}, m/p^{1/k}]$.

For example, taking the symmetric 95% interval $p = 2.5\%$ and $q = 97.5\%$ for $k = 5$ yields $0.025^{1/5} \approx 0.48$, $0.975^{1/5} \approx 0.995$, so the confidence interval is approximately $[1.005m, 2.08m]$. The lower bound is very close to m, thus more informative is the asymmetric confidence interval from $p = 5\%$ to 100%; for $k = 5$ this yields $0.05^{1/5} \approx 0.55$ and the interval $[m, 1.82m]$.

More generally, the (downward biased) 95% confidence interval is $[m, m/0.05^{1/k}] = [m, m{\cdot}20^{1/k}]$. For a range of k values, with the UMVU point estimator (plus 1 for legibility) for reference, this yields:

k	Point estimate	Confidence interval
1	$2m$	$[m, 20m]$
2	$1.5m$	$[m, 4.5m]$
5	$1.2m$	$[m, 1.82m]$
10	$1.1m$	$[m, 1.35m]$
20	$1.05m$	$[m, 1.16m]$

Immediate observations are:

- For small sample sizes, the confidence interval is very wide, reflecting great uncertainty in the estimate.
- The range shrinks rapidly, reflecting the exponentially decaying probability that *all* observations in the sample will be significantly below the maximum.
- The confidence interval exhibits positive skew, as N can never be below the sample maximum, but can potentially be arbitrarily high above it.

Note that m/k cannot be used naively (or rather $(m + m/k - 1)/k$) as an estimate of the standard error SE, as the standard error of an estimator is based on the *population* maximum (a parameter), and using an estimate to estimate the error in that very estimate is circular reasoning.

Bayesian analysis

The Bayesian approach to the German tank problem is to consider the credibility $(N = n \mid M = m, K = k)$ that the number of enemy tanks N is equal to the number n, when the number of observed tanks, K is equal to the number k, and the maximum observed serial number M is equal to the number m. The answer to this problem depends on the choice of prior for N. One can proceed using a proper

prior, e.g., the Poisson or Negative Binomial distribution, where closed formula for the posterior mean and posterior variance can be obtained.[14] An alternative is to proceed using direct calculations as shown below.

For brevity, in what follows, $(N = n \mid M = m, K = k)$ is written $(n \mid m, k)$

Conditional probability

The rule for conditional probability gives

$$(n \mid m, k)(m \mid k) = (m \mid n, k)(n \mid k) = (m, n \mid k)$$

Probability of *M* knowing *N* and *K*

The expression

$$(m \mid n, k) = (M = m \mid N = n, K = k)$$

is the conditional probability that the maximum serial number observed, M, is equal to m, when the number of enemy tanks, N, is known to be equal to n, and the number of enemy tanks observed, K, is known to be equal to k.

It is

$$(m \mid n, k) = \binom{m-1}{k-1}\binom{n}{k}^{-1}[k \le m][m \le n]$$

where $\binom{n}{k}$ is a binomial coefficient and $[k \le n]$ is an Iverson bracket.

The expression can be derived as follows: $(m \mid n, k)$ answers the question: "What is the probability of a specific serial number m being the highest number observed in a sample of k tanks, given there are n tanks in total?"

One can think of the sample of size k to be the result of k individual draws. Assume m is observed on draw number d. The probability of this occurring is:

$$\underbrace{\frac{m-1}{n}\cdot\frac{m-2}{n-1}\cdot\frac{m-3}{n-2}\cdots\frac{m-d+1}{n-d+2}}_{\text{d-1 - times}}\cdot\underbrace{\frac{1}{n-d+1}}_{\text{draw no. d}}\cdot\underbrace{\frac{m-d}{n-d}\cdot\frac{m-d-1}{n-d-1}\cdots\frac{m-d-(k-d-1)}{n-d-(k-d-1)}}_{k-d-times} = \frac{(n-k)!}{n!}\cdot\frac{(m-1)!}{(m-k)!}.$$

As can be seen from the right-hand side, this expression is independent of d and therefore the same for each $d \le k$. As m can be drawn on k different draws, the probability of any specific m being the largest one observed is k times the above probability:

$$(m \mid n, k) = k \cdot \frac{(n-k)!}{n!}\cdot\frac{(m-1)!}{(m-k)!} = \binom{m-1}{k-1}\binom{n}{k}^{-1}.$$

Probability of *M* knowing only *K*

The expression $(m \mid k) = (M = m \mid K = k)$ is the probability that the maximum serial number is equal to m once k tanks have been observed but before the serial numbers have actually been observed.

The expression $(m \mid k)$ can be re-written in terms of the other quantities by marginalizing over all possible n.

$$\begin{aligned}(m \mid k) &= (m \mid k)\cdot 1 \\ &= (m \mid k)\sum_{n=0}^{\infty}(n \mid m, k) \\ &= (m \mid k)\sum_{n=0}^{\infty}(m \mid n, k)\frac{(n \mid k)}{(m \mid k)} \\ &= \sum_{n=0}^{\infty}(m \mid n, k)(n \mid k)\end{aligned}$$

Credibility of *N* knowing only *K*

The expression

$$(n \mid k) = (N = n \mid K = k)$$

is the credibility that the total number of tanks, N, is equal to n when the number K tanks observed is known to be k, but before the serial numbers have been observed. Assume that it is some discrete uniform distribution

$$(n \mid k) = (\Omega - k)^{-1}[k \leq n][n < \Omega]$$

The upper limit Ω must be finite, because the function

$$f(n) = \lim_{\Omega \to \infty} (\Omega - k)^{-1}[k \leq n][n < \Omega] = 0$$

is not a mass distribution function.

Credibility of *N* knowing *M* and *K*

$$(n \mid m, k) = (m \mid n, k)\left(\sum_{n=m}^{\Omega-1}(m \mid n, k)\right)^{-1}[m \leq n][n < \Omega]$$

If $k \geq 2$, then $\sum_{n=m}^{\infty}(m \mid n, k) < \infty$, and the unwelcome variable Ω disappears from the expression.

$$(n \mid m, k) = (m \mid n, k)\left(\sum_{n=m}^{\infty}(m \mid n, k)\right)^{-1}[m \leq n]$$

For $k \geq 1$ the mode of the distribution of the number of enemy tanks is m.

For $k \geq 2$, the credibility that the number of enemy tanks is *equal to* n, is

$$(N = n \mid m, k) = (k-1)\binom{m-1}{k-1}k^{-1}\binom{n}{k}^{-1}[m \leq n]$$

The credibility that the number of enemy tanks, N, is *greater than* n, is

$$(N > n \mid m, k) = \begin{cases} 1 & \text{if } n < m \\ \frac{\binom{m-1}{k-1}}{\binom{n}{k-1}} & \text{if } n \geq m \end{cases}$$

Mean value and standard deviation

For $k \geq 3$, N has the finite mean value:

$$(m-1)(k-1)(k-2)^{-1}$$

For $k \geq 4$, N has the finite standard deviation:

$$(k-1)^{1/2}(k-2)^{-1}(k-3)^{-1/2}(m-1)^{1/2}(m+1-k)^{1/2}$$

These formulas are derived below.

Summation formula

The following binomial coefficient identity is used below for simplifying series relating to the German Tank Problem.

$$\sum_{n=m}^{\infty}\frac{1}{\binom{n}{k}} = \frac{k}{k-1}\frac{1}{\binom{m-1}{k-1}}$$

This sum formula is somewhat analogous to the integral formula

$$\int_{n=m}^{\infty} \frac{dn}{n^k} = \frac{1}{k-1}\frac{1}{m^{k-1}}$$

These formulas apply for $k > 1$.

One tank

Observing one tank randomly out of a population of n tanks gives the serial number m with probability $1/n$ for $m \le n$, and zero probability for $m > n$. Using Iverson bracket notation this is written

$$(M = m \mid N = n, K = 1) = (m \mid n) = \frac{[m \le n]}{n}$$

This is the conditional probability mass distribution function of m.

When considered a function of n for fixed m this is a likelihood function.

$$\mathcal{L}(n) = \frac{[n \ge m]}{n}$$

The maximum likelihood estimate for the total number of tanks is $N_0 = m$, clearly a biased estimate since the true number can be more than this, potentially many more, but cannot be fewer.

The marginal likelihood (i.e. marginalized over all models) is infinite, being a tail of the harmonic series.

$$\sum_n \mathcal{L}(n) = \sum_{n=m}^{\infty} \frac{1}{n} = \infty$$

but

$$\begin{aligned}\sum_n \mathcal{L}(n)[n < \Omega] &= \sum_{n=m}^{\Omega-1} \frac{1}{n} \\ &= H_{\Omega-1} - H_{m-1}\end{aligned}$$

where H_n is the harmonic number.

The credibility mass distribution function depends on the prior limit Ω:

$$\begin{aligned}&(N = n \mid M = m, K = 1) \\ &= (n \mid m) = \frac{[m \le n]}{n}\frac{[n < \Omega]}{H_{\Omega-1} - H_{m-1}}\end{aligned}$$

The mean value of N is

$$\begin{aligned}\sum_n n \cdot (n \mid m) &= \sum_{n=m}^{\Omega-1} \frac{1}{H_{\Omega-1} - H_{m-1}} \\ &= \frac{\Omega - m}{H_{\Omega-1} - H_{m-1}} \\ &\approx \frac{\Omega - m}{\log\left(\frac{\Omega-1}{m-1}\right)}\end{aligned}$$

Two tanks

If two tanks rather than one are observed, then the probability that the larger of the observed two serial numbers is equal to m, is

$$(M = m \mid N = n, K = 2) = (m \mid n) = [m \le n]\frac{m-1}{\binom{n}{2}}$$

When considered a function of n for fixed m this is a likelihood function

$$\mathcal{L}(n) = [n \geq m]\frac{m-1}{\binom{n}{2}}$$

The total likelihood is

$$\begin{aligned}\sum_n \mathcal{L}(n) &= \frac{m-1}{1}\sum_{n=m}^{\infty}\frac{1}{\binom{n}{2}} \\ &= \frac{m-1}{1}\cdot\frac{2}{2-1}\cdot\frac{1}{\binom{m-1}{2-1}} \\ &= 2\end{aligned}$$

and the credibility mass distribution function is

$$\begin{aligned}&(N = n \mid M = m, K = 2) \\ &= (n \mid m) \\ &= \frac{\mathcal{L}(n)}{\sum_n \mathcal{L}(n)} \\ &= [n \geq m]\frac{m-1}{n(n-1)}\end{aligned}$$

The median $\tilde{N}$ satisfies

$$\sum_n [n \geq \tilde{N}](n \mid m) = \frac{1}{2}$$

so

$$\frac{m-1}{\tilde{N}-1} = \frac{1}{2}$$

and so the median is

$$\tilde{N} = 2m - 1$$

but the mean value of N is infinite

$$\mu = \sum_n n \cdot (n \mid m) = \frac{m-1}{1}\sum_{n=m}^{\infty}\frac{1}{n-1} = \infty$$

Many tanks

Credibility mass distribution function

The conditional probability that the largest of k observations taken from the serial numbers $\{1,\ldots,n\}$, is equal to m, is

$$\begin{aligned}&(M = m \mid N = n, K = k \geq 2) \\ &= (m \mid n, k) \\ &= [m \leq n]\frac{\binom{m-1}{k-1}}{\binom{n}{k}}\end{aligned}$$

The likelihood function of n is the same expression

$$\mathcal{L}(n) = [n \geq m]\frac{\binom{m-1}{k-1}}{\binom{n}{k}}$$

The total likelihood is finite for $k \geq 2$:

$$\begin{aligned}\sum_n \mathcal{L}(n) &= \frac{\binom{m-1}{k-1}}{1} \sum_{n=m}^{\infty} \frac{1}{\binom{n}{k}} \\ &= \frac{\binom{m-1}{k-1}}{1} \cdot \frac{k}{k-1} \cdot \frac{1}{\binom{m-1}{k-1}} \\ &= \frac{k}{k-1}\end{aligned}$$

The credibility mass distribution function is

$$\begin{aligned}&(N = n \mid M = m, K = k \geq 2) = (n \mid m, k) \\ &= \frac{\mathcal{L}(n)}{\sum_n \mathcal{L}(n)} \\ &= [n \geq m] \frac{k-1}{k} \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \\ &= [n \geq m] \frac{m-1}{n} \frac{\binom{m-2}{k-2}}{\binom{n-1}{k-1}} \\ &= [n \geq m] \frac{m-1}{n} \frac{m-2}{n-1} \frac{k-1}{k-2} \frac{\binom{m-3}{k-3}}{\binom{n-2}{k-2}}\end{aligned}$$

The complementary cumulative distribution function is the credibility that $N > x$

$$\begin{aligned}&(N > x \mid M = m, K = k) \\ &= \begin{cases} 1 & \text{if } x < m \\ \sum_{n=x+1}^{\infty} (n \mid m, k) & \text{if } x \geq m \end{cases} \\ &= [x < m] + [x \geq m] \sum_{n=x+1}^{\infty} \frac{k-1}{k} \frac{\binom{m-1}{k-1}}{\binom{N}{k}} \\ &= [x < m] + [x \geq m] \frac{k-1}{k} \frac{\binom{m-1}{k-1}}{1} \sum_{n=x+1}^{\infty} \frac{1}{\binom{n}{k}} \\ &= [x < m] + [x \geq m] \frac{k-1}{k} \frac{\binom{m-1}{k-1}}{1} \cdot \frac{k}{k-1} \frac{1}{\binom{x}{k-1}} \\ &= [x < m] + [x \geq m] \frac{\binom{m-1}{k-1}}{\binom{x}{k-1}}\end{aligned}$$

The cumulative distribution function is the credibility that $N \leq x$

$$\begin{aligned}&(N \leq x \mid M = m, K = k) \\ &= 1 - (N > x \mid M = m, K = k) \\ &= [x \geq m] \left(1 - \frac{\binom{m-1}{k-1}}{\binom{x}{k-1}}\right)\end{aligned}$$

Order of magnitude

The order of magnitude of the number of enemy tanks is

$$\begin{aligned}\mu &= \sum_n n \cdot (N = n \mid M = m, K = k) \\ &= \sum_n n[n \geq m] \frac{m-1}{n} \frac{\binom{m-2}{k-2}}{\binom{n-1}{k-1}} \\ &= \frac{m-1}{1} \frac{\binom{m-2}{k-2}}{1} \sum_{n=m}^{\infty} \frac{1}{\binom{n-1}{k-1}} \\ &= \frac{m-1}{1} \frac{\binom{m-2}{k-2}}{1} \cdot \frac{k-1}{k-2} \frac{1}{\binom{m-2}{k-2}} \\ &= \frac{m-1}{1} \frac{k-1}{k-2}\end{aligned}$$

Statistical uncertainty

The statistical uncertainty is the standard deviation σ, satisfying the equation

$$\sigma^2 + \mu^2 = \sum_n n^2 \cdot (N = n \mid M = m, K = k)$$

So

$$\begin{aligned}\sigma^2 + \mu^2 - \mu &= \sum_n n(n-1) \cdot (N = n \mid M = m, K = k) \\ &= \sum_{n=m}^{\infty} n(n-1) \frac{m-1}{n} \frac{m-2}{n-1} \frac{k-1}{k-2} \frac{\binom{m-3}{k-3}}{\binom{n-2}{k-2}} \\ &= \frac{m-1}{1} \frac{m-2}{1} \frac{k-1}{k-2} \cdot \frac{\binom{m-3}{k-3}}{1} \sum_{n=m}^{\infty} \frac{1}{\binom{n-2}{k-2}} \\ &= \frac{m-1}{1} \frac{m-2}{1} \frac{k-1}{k-2} \frac{\binom{m-3}{k-3}}{1} \frac{k-2}{k-3} \frac{1}{\binom{m-3}{k-3}} \\ &= \frac{m-1}{1} \frac{m-2}{1} \frac{k-1}{k-3}\end{aligned}$$

and

$$\begin{aligned}\sigma &= \sqrt{\frac{m-1}{1} \frac{m-2}{1} \frac{k-1}{k-3} + \mu - \mu^2} \\ &= \sqrt{\frac{(k-1)(m-1)(m-k+1)}{(k-3)(k-2)^2}}\end{aligned}$$

The variance-to-mean ratio is simply

$$\frac{\sigma^2}{\mu} = \frac{m-k+1}{(k-3)(k-2)}$$

See also

- Mark and recapture, other method of estimating population size
- Maximum spacing estimation, which generalizes the intuition of "assume uniformly distributed"
- Copernican principle and Lindy effect, analogous predictions of lifetime assuming just one observation in the sample (current age).
  - The Doomsday argument, application to estimate expected survival time of the human race.
- Generalized extreme value distribution, possible limit distributions of sample maximum (opposite question).
- Maximum likelihood

- Bias of an estimator
- Likelihood function

Further reading

- Goodman, L. A. (1954). "Some Practical Techniques in Serial Number Analysis". *Journal of the American Statistical Association*. American Statistical Association. **49** (265): 97–112. doi:10.2307/2281038 (https://doi.org/10.2307%2F2281038). JSTOR 2281038 (https://www.jstor.org/stable/2281038).

Notes

a. An Armored Ground Forces policy statement of November 1943 concluded: "The recommendation of a limited proportion of tanks carrying a 90 mm gun is not concurred in for the following reasons: The M4 tank has been hailed widely as the best tank of the battlefield today. ... There appears to be no fear on the part of our forces of the German Mark VI (Tiger) tank. There can be no basis for the T26 tank other than the conception of a tank-vs.-tank duel – which is believed to be unsound and unnecessary."[1]

b. The lower bound was unknown, but to simplify the discussion, this detail is generally omitted, taking the lower bound as known to be 1.

c. Ruggles & Brodie is largely a practical analysis and summary, not a mathematical one – the estimation problem is only mentioned in footnote 3 on page 82, where they estimate the maximum as "sample maximum + average gap".

d. As discussed in birthday attack, one can expect a collision after $1.25\sqrt{H}$ numbers, if choosing from H possible outputs. This square root corresponds to half the digits. For example, in any base, the square root of a number with 100 digits is approximately a number with 50 digits.

e. In a continuous distribution, there is no −1 term.

f. Given a particular set of observations, this set is most likely to occur if the population maximum is the sample maximum, not a higher value (it cannot be lower).

g. The sample maximum is never more than the population maximum, but can be less, hence it is a biased estimator: it will tend to *underestimate* the population maximum.

h. For example, the gap between 2 and 7 is (7 − 2) − 1 = 4, consisting of 3, 4, 5, and 6.

References

1. AGF policy statement. Chief of staff AGF. November 1943. MHI
2. Ruggles & Brodie 1947, p. ?.
3. "Gavyn Davies does the maths – How a statistical formula won the war" (https://www.theguardian.com/world/2006/jul/20/secondworldwar.tvandradio). *The Guardian*. 20 July 2006. Retrieved 6 July 2014.
4. Matthews, Robert (23 May 1998), "Data sleuths go to war, sidebar in feature "Hidden truths" " (https://web.archive.org/web/20010418025817/http://www.newscientist.com/ns/980523/features.html#data), *New Scientist*, archived from the original (https://www.newscientist.com/article/mg15821355.000-hidden-truths.html) on 18 April 2001
5. Bob Carruthers (1 March 2012). *Panther V in Combat* (https://books.google.com/books?id=99JRxKz4Da4C&pg=PT94). Coda Books. pp. 94–. ISBN 978-1-908538-15-4.
6. Ruggles & Brodie 1947, pp. 82–83.
7. Ruggles & Brodie 1947, p. 89.
8. Ruggles & Brodie 1947, pp. 90–91.
9. Volz 2008.
10. Johnson 1994.
11. "How many Commodore 64 computers were really sold?" (https://web.archive.org/web/20160306232450/http://www.pagetable.com/?p=547). *pagetable.com*. 1 February 2011. Archived from the original (http://www.pagetable.com/?p=547) on 6 March 2016. Retrieved 6 July 2014.
12. "Rockets and Missiles" (https://www.spaceline.org/rocketsum/jupiter-c.html). *www.spaceline.org*.
13. Joyce, Smart. "German Tank Problem" (https://web.archive.org/web/20120424231135/http://www.lhs.logan.k12.ut.us/~jsmart/tank.htm). Logan High School. Archived from the original (http://www.lhs.logan.k12.ut.us/~jsmart/tank.htm) on 24 April 2012. Retrieved 8 July 2014.
14. Höhle, M.; Held, L. (2006). "Bayesian Estimation of the Size of a Population" (https://epub.ub.uni-muenchen.de/2094/1/paper_499.pdf) (PDF). *Technical Report SFB 386, No. 399, Department of Statistics, University of Munich*. Retrieved 17 April 2016.

Works cited

- Johnson, R. W. (Summer 1994). "Estimating the Size of a Population" (https://web.archive.org/web/20140223104835/http://www.rsscse-edu.org.uk/tsj/wp-content/uploads/2011/03/johnson.pdf) (PDF). *Teaching Statistics*. **16** (2): 50–52. doi:10.1111/j.1467-9639.1994.tb00688.x (https://doi.org/10.1111%2Fj.1467-9639.1994.tb00688.x). Archived from the original (http://www.rsscse-edu.org.uk/tsj/wp-content/uploads/2011/03/johnson.pdf) (PDF) on 23 February 2014.
- Ruggles, R.; Brodie, H. (1947). "An Empirical Approach to Economic Intelligence in World War II". *Journal of the American*

Statistical Association. **42** (237): 72. doi:10.1080/01621459.1947.10501915 (https://doi.org/10.1080%2F01621459.1947.10501915). JSTOR 2280189 (https://www.jstor.org/stable/2280189).

- Volz, A. G. (July 2008). "A Soviet Estimate of German Tank Production". *The Journal of Slavic Military Studies*. **21** (3): 588–590. doi:10.1080/13518040802313902 (https://doi.org/10.1080%2F13518040802313902). S2CID 144483708 (https://api.semanticscholar.org/CorpusID:144483708).

Retrieved from "https://en.wikipedia.org/w/index.php?title=German_tank_problem&oldid=1146127588"