Queen Mary
University of London

Specific Learning Difficulties Cover Note

Student ID: 170815918

Advice for assessors and examiners

Guidelines for markers assessing coursework and examinations of students diagnosed with Specific Learning Difficulties (SpLDs) –
As far as the learning outcomes for the module allow, examiners are asked to mark exam scripts sympathetically, ignoring the types of errors that students with SpLDs make and to focus on content and the student's understanding of the subject. Specific learning difficulties such as Attention Deficit Disorders, dyslexia and or dyspraxia may affect student performance in the following ways:

- The candidate's spelling, grammar and punctuation may be less accurate than expected
- The candidate's organisation of ideas may be confused, affecting the overall structure of written work
- The candidate's proof reading may be weak with some errors undetected, particularly homophones and homonyms which can avoid spell checkers

Under examination conditions, these difficulties are likely to be exacerbated. Errors are likely to become more marked towards the end of scripts.

Useful approaches can include:

- Reading the passage quickly for content
- Including positive/constructive comments amongst the feedback so that students can work with specialist study skills tutors on developing new coping strategies
- Using clear English and when correcting; explain what is wrong and give examples
- Using non-red coloured pens for comments/corrections

Colleagues in schools are asked to ensure that students with specific learning difficulties access the support provided by the Disability and Dyslexia Service.

For more information regarding marking guidelines see DDS webpage http://www.dds.qmul.ac.uk/staffinfo/index.html and the Institutional Marking Practices for Dyslexic Students

Disability and Dyslexia Service
Student Services
Room 2.06 Francis Bancroft Building
www.dds.qmul.ac.uk
Tel: 020 7882 2756
Email: dds@qmul.ac.uk

THE GERMAN TANK PROBLEM

Third Year Project

Kamryn Harris

Queen Mary University of London | 2020

Chapter 1 – The Introduction

Section 1.1: Introduction

One of the most interesting applications of mathematics comes from the iconic population estimator problem that came about during WWII and has since been called "The German Tank Problem". This famous statistical problem and its approach has influenced history and has given further support to using mathematics creatively to solve complex, real-world problems.

A population size estimate problem is a statistical problem that uses statistical methods and instruments to figure out the total size of a population based on a sample from the population. "The German Tank Problem" was given its name because of the nature of the problem it addresses. In this historical case, during World War II the Allies used mathematics to estimate the *total number* of German tanks from a *small sample* (serial numbers obtained from captured, destroyed or observed tanks).

There are two statistical approaches for this type of problem, one using the frequentist approach and the other using the Bayesian approach. The difference between Bayesian and frequentist is that frequentists "use probability only to model certain processes broadly described as "sampling." Bayesians use probability more widely to model both sampling and other kinds of uncertainty." [1]. For this paper, I will focus on the frequentist approach. For a population size estimation problem such as this one, the frequentist approach, essentially, assumes that the total population size is a fixed number, N, in contrast to the Bayesian approach which assumes that N is a random variable.

In this paper I will cover the context of "The German Tank Problem" with its historical applications. I will then go into the motivation for the structure of the analytics of the problem. I will then discuss the frequentist approach in more detail and will construct the mathematical formulas used to solve the problem using this specific statistical approach. Lastly, I will go over a few simulations conducted in order to show how actual calculations work.

Note that I will use serial number and value to mean the same thing throughout the paper.

Section 1.2: The History – "The German Tank Problem"

During WWII the Allies had to be clever and find new advanced solutions to get an advantage over their opponent. The Allies were very shrewd in their methods to outwit the Germans and mathematics was a vital part of the Allies' strategy. It is widely known that mathematics was used by the Allies in cracking the Enigma code, but they also used mathematics to face another important question about German defence production. It was important for the Allies to know what they were up against in order to know what they needed to match or outdo the German side. So, the Allies wanted to know how many tanks the German army was producing each month. There were two different strategies used to get this information.

The first way was the traditional intelligence gathering route, in which, an estimate would be made solely based on intelligence gathered from spies.

The other way was more revolutionary. Traditional means of intelligence were more respected, and nothing had challenged it before. However, mathematics and statistics had made great leaps and were beginning to look like they could be used as well. The use of statistics in this way would later influence future problems of this nature. The approach to this problem was one of the greatest achievements of mathematicians during WWII.

The method worked like this: The Allies were lucky in that the Germans were, by nature, organised even in war. Consequently, the Germans had thoroughly labelled tanks and their parts in consecutive order from 1 to some unknown value. The Allies used the serial numbers on captured or destroyed tanks to get a sample and, from the sample they collected, they made a statistical inference about how many tanks the Germans were making each month.

This was essential to their strategy for winning battles and even helped the Allies better understand their opponent. “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.” [3]

In the end, the estimations made using the statistical methods proved to be more accurate than the conventional intelligence. Evident in the following chart. [3]

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

“The German Tank Problem” was revolutionary in that it helped provide evidence that using statistics in problems such as this one was more effective than using traditional means of intelligence gathering.

This method was used for other military equipment during WWII as well. Similar methods have since been used in getting an edge in wars, including the Korean War.

It is important to note, however, that since, there have been efforts made to make it more difficult to gather the original sample data from serial numbers that have been replaced using cryptography.

Section 1.3: Motivation

To get an understanding for the problem and gain some intuition around it, I will motivate the mathematics I will eventually use.

For this example, assume the Allies have observed and recorded serial numbers from a population of wrecked tanks, labelled with positive integer valued serial numbers in consecutive order. (Positive integer values because they were labelled starting with serial number 1 and increasing in value from there).

Suppose that an Ally scientist is given a sample of serial numbers, let's say,

$$\{23,99,56,131\}$$

Intuitively, most people that make a guess will most likely make the assumption that the sample is most likely somewhat equally distributed and so the total value must be a similar distance away from the maximal value in the set. From that assumption, the naïve approach might be to assume the formula for finding N, the total number of tanks produced in a particular month, looks something like:

$$N = x_{max} + x_{max}\frac{1}{sample\ size}$$

(Where x_{max} is the value of the maximum serial number from the sample)

There isn't any formal mathematical support for this formula, but it somehow intuitively makes sense. So for the given example, I see there are 4 integers in the sample, I assume they are all equally distributed so that must mean that the maximum value of the sample, 131, is 80% of the total population, N, which would mean I need to add 20% more to that maximum value to find the N.

I will later refer back to this 'intuition formula'.

Chapter 2 - Overview of Frequentist Approach for "The German Tank Problem"

Section 2.1: Introduction to the frequentist approach

The frequentist approach to this type of population size estimator problem assumes the value for the total number of tanks produced in a month, N, is a fixed value. "In a frequentist approach to inference, unknown parameters are often […] treated as having fixed but unknown values that are not capable of being treated as random variates in any sense, and hence there is no way that probabilities can be associated with them." [4]

The frequentist approach is useful because it can be more intuitive to work with yet still yield reliable results. It is a less abstract way of working with statistical instruments.

This approach will use statistical instruments such as binomial coefficients, a probability mass function and expectation in order to get a precise estimate for the total number of tanks, N, produced in a particular month.

With this approach, we are considering samples of certain sizes, m, so we will *count* the number of serial numbers (*values*) that we can *choose* from the total population using a binomial coefficient to mathematically represent this process. From there, it will become evident that we are interested in whether the probability of the maximal value of the sample is equal to some particular value. This is where the need for a probability mass function comes in and will make a contribution to the results. The probability mass function will be used to calculate the expectation of what the maximal value is. From the expression for the expectation (of the maximal value), we will be able to derive an expression for N after introducing another random variable, Y, dependent on the random variable of the maximal value.

The most important serial number is the maximal value of the sample. Since the maximal value implies the value for N is definitely greater than or equal to the maximal value of the sample. N is clearly greater than the smaller values so theses simply provide information on the size of the sample. The maximal value in the sample will therefore play a key role in the subsequent calculations.

Section 2.1.1: The Analytics

In this section I will discuss and layout the necessary tools I will use for this population size estimate using the Frequentist approach.

Notation:
Let N denote the parameter for the actual size of the population, i.e., total number of tanks
Let m denote the sample size.
Let x_m denote the last value in the sequence of value in the sample
Let X_i denote the random variable for the values in position for $i = 1,2,\dots,m$
Let X_M denote the random variable for maximal value, i.e., $X_M = max\{X_1, X_2, X_3, \dots, X_m\}$

Remarks:
In order to find the value for N, I will start by laying out the random variable X_M which I will eventually use to estimate N.

As a frequentist, I will assume that the total number of tanks being produced in a particular month, N, is a fixed value, as previously discussed. **For this section I will assume that the fixed value of N is known.**

Given sample with size m, I have a collection of random variables to represent the randomness of the value the sample can take at each position. X_1 will be the corresponding random variable for the values in the first position of the set, x_1, and so on until X_m is the random variable in the last position of the set. Since the list is not necessarily in increasing order, note that X_m is not necessarily the largest value in the set so I designated this as X_M under the 'Notation' heading

above. In other words, $X_M = max\{X_1, X_2, X_3, \dots, X_m\}$ while X_m refers to the random variable that represents the value in the last position of the set/sample).

The Random Variables: $X_1, X_2, X_3, \dots, X_m$

The corresponding values in the sample: $\{x_1, x_2, x_3, \dots, x_{m,}\}$

For example, here I have several random samples with sample size 4 (i.e. $m = 4$)

i. $\{2, 55, 123, 134\}$

ii. $\{15, 25, 99, 27\}$

iii. $\{2, 134, 123, 55\}$

The values are as follows:

i. $x_1 = 2,\ x_2 = 55, x_3 = 123,\ x_{4,} = 134$

Here $m = 4$, which implies $x_{m,} = x_4 = X_m = X_4 \quad x_4 = X_4 = 134$ and $X_M = 134$

ii. $x_1 = 15,\ x_2 = 25, x_3 = 99,\ x_4 = 27$

Here $m = 4$, which implies (again) $x_{m,} = x_4 = X_m = X_4\ so\ x_4 = X_4 = 27$ and $X_M = 99 = x_3$ $X_M = X_3$ since the largest in the x_3 position

iii. $x_1 = 2,\ x_2 = 134, x_3 = 123,\ x_4 = 55$

Here $m = 4$, which implies (again) $x_{m,} = x_4 = X_m = X_4\ so\ x_4 = X_4 = 27$ and $X_M = 99 = x_2$ $X_M = X_2$ since the largest in the x_2 position

Section 2.1.2 – Binomial Coefficients and The Probability Mass Function

The probability mass function, pmf, is defined by $P(X_M = \ell)$, i.e. the probability that the random variable, X_M, which represents the maximum value of the sample, will be some value ℓ. For any value of ℓ I can therefore work out the probability the maximum is equal to this value, ℓ. By inputting some value for ℓ, my output will be a probability.

It is clear that the probability mass function will have a 0 probability for any value less than m or for any value greater than N. I know this because it is impossible to get a value less than m considering I have m values in my sample. It is also impossible to get a value greater than N because I cannot get a value larger than the total population (I have assumed N is known). To further illustrate this point, say I am given a sample with size $m = 5$. I know $P(X_M = 1) = 0$

because if $m = 5$ it is impossible for the maximum of a sample to be the integer 1 by the previous assumptions I have made (that the tanks are labelled with integers, greater than zero, in increasing order) . I also know that $\mathrm{P}(X_M = N + 1) = 0$, by similar logic. Therefore, I can deduce there is a range of values of ℓ for which I will get a nonzero probability, $\ell = m, m + 1, \dots, N$.

From the setup of the problem I can deduce what I already know in order to help me further set the stage. I know from the basis of my problem that since I get a sample of serial numbers from N total tanks, I also know this sample is definitely selected at random since my sample is coming from randomly destroyed or captured tanks. To translate this into mathematical terms, I choose m numbers at random from a population with size N, therefore I know I must have the binomial coefficient $\binom{N}{m}$, "N choose m", somewhere in my pmf. Since a pmf is giving me the probability of something happening, it will be a fraction of something and also recall that probability is always some value between zero and one. It is apparent that the binomial coefficient $\binom{N}{m}$ gives me the number of all possibilities so it is clear that this is my denominator.

I am trying to work out the number of choices we can have in which we have a set of m serial numbers that are chosen from N total serial numbers and the set has the maximal serial number is ℓ. In order to understand a bit more and see how to set up the pmf, I will go into an example.

Example
Let $m = 5$ (i.e. samples with size 5)
Suppose $\ell = 143$, (i.e. the maximal value of the set is 143) then I take all my random samples and choose the samples with the maximal value 143 in sets with $m = 5$:

$$\{15, 25, 48, 27, 143\}$$

$$\{12,45,23,143,65\}$$

…

Since I only really care about the maximal integer, $\ell = 143$, the other 4 entries can be any other value as long as they are less than 143, so I only can choose $\ell - 1$ or $143 - 1 = 142$ entries for the remaining spots of the sample, i.e. in $m - 1$ or $5 - 1 = 4$ spots. Therefore, I end up with the binomial coefficient $\binom{142}{4}$.

Comment
From this example I see $\binom{\ell-1}{m-1}$ will be in the numerator of the pmf. Therefore, **the general pmf** for when the maximum is ℓ looks like:

$$\mathrm{P}(X_M = \ell) = \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}}$$

Recall that $\mathrm{P}(X_M = \ell) = 0 \; if \; \ell < m$ and if $\ell > N$ for reasons previously stated.

Section 2.1.3 – The Expectation

Now I want to use what I have done to work out a formula for the Expectation value because I will need it to calculate N. The Expectation of the random variable X_M is, essentially, just the average observation for the maximal value of the samples. To compute $E(X_M)$, using the definition of expectation:

$$E(X_M) = \sum_{\ell=m}^{N} \ell \, P(X_M = \ell) = \sum_{\ell=m}^{N} \ell \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}}$$

By definition By the comment above

Remark Note that the sum has ℓ starting at m because ℓ cannot be less than m (the binomial coefficient would be undefined if you had ℓ choose m with $\ell < m$)

There are a few identities of binomial coefficients that I will use in order to eventually solve for N. The following proposition is my first relevant identity.

Proposition 2.1 Let ℓ and m be natural numbers. For binomial coefficients, the following identity holds:

$$\binom{\ell}{m} = \frac{\ell}{m}\binom{\ell-1}{m-1} \qquad \textbf{(1)}$$

Proof: I claim that the left-hand side is equal to the right-hand side. In order to prove this, I will expand side independently from one another using the definition of Binomial Coefficients. Firstly, the right-hand side is, by definition,

$$\binom{\ell}{m} = \frac{\ell!}{m!\,(m-\ell)!}$$

I want to expand the left-hand side enough using what I know in order to end up with something that looks like the above.
Secondly, for the left-hand side using the definition I get,

$$\frac{\ell}{m}\binom{\ell-1}{m-1} = \frac{\ell}{m} \cdot \frac{(\ell-1)!}{(m-1)!\big((\ell-1)-(m-1)\big)!}$$

Combining $\big((\ell-1)-(m-1)\big)!$ We get $(\ell-1-m+1)$ so the 1's cancel

$$= \frac{\ell}{m} \cdot \frac{(\ell-1)!}{(m-1)!\,(\ell-m)!}$$

$$= \frac{\ell \cdot (\ell - 1)!}{m \cdot (m-1)!\,(\ell - m)!}$$

I will use the fact that $\ell(\ell-1)! = \ell!$ In order to simplify in the numerator and denominator, so I end up with the following for the equation of the left-hand side:

$$= \frac{\ell!}{m!\,(\ell - m)!}$$

Therefore, it is clear that both sides are the same:

$$\binom{\ell}{m} = \frac{\ell!}{m!\,(m-\ell)!} = \frac{\ell}{m}\binom{\ell-1}{m-1}$$

■

I will now use this identity in the formula previously given for $E(X_M)$. However, I want the identity to fit in easily. So, my first step is to manipulate $E(X_M)$ so that it looks similar to (1). I will do this by multiplying by a 'special one', $\frac{m}{m}$:

$$E(X_M) = \sum_{\ell=m}^{N} \ell \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}} \cdot \frac{m}{m} = m \cdot \sum_{\ell=m}^{N} \frac{\ell}{m} \cdot \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}}$$

Now $E(X_M)$ looks similar to (1). So, when I apply the identity from (1) to the formula for $E(X_M)$:

$$E(X_M) = \frac{m \cdot \sum_{\ell=m}^{N} \binom{\ell}{m}}{\binom{N}{m}} \qquad \textbf{(*)}$$

I will introduce the next identity in order to manipulate the numerator in (*) further.

Proposition 2.2 Let k and m be natural numbers. For binomial coefficients, the following identity holds:

$$\binom{k}{m} = \binom{k+1}{m+1} - \binom{k}{m+1} \qquad \textbf{(2)}$$

Proof: Similar to Proposition 2.1, to work out that both sides equal each other and, therefore, proving the proposition, I will work out each side independently from one another using the definition of Binomial Coefficients.
Let $0 \leq m \leq k$. I know that, by definition,

The right-hand side is:

$$\binom{k}{m} = \frac{k!}{m!\,(n-k)!}$$

To expand the left-hand side:

$$\binom{k+1}{m+1} - \binom{k}{m+1} = \frac{(k+1)!}{(m+1)!\big((k+1)-(m+1)\big)!} - \frac{k!}{(m+1)!\big(k-(m+1)\big)!}$$

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

I will pull out the common factor in the denominator, $(m+1)!$:

$$= \frac{1}{(m+1)!} \cdot \left\{ \frac{(k+1)!}{(k-m)!} - \frac{k!}{(k-m-1)!} \right\}$$

Note that $(k+1)! = (k+1)k!$, so I can now pull out a common factor in the numerator:

$$= \frac{k!}{(m+1)!} \cdot \left\{ \frac{k+1}{(k-m)!} - \frac{1}{(k-m-1)!} \right\}$$

I see that $(k-m)! = (k-m)(k-m-1)!$ In order to pull out another common factor:

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

Taking the portion in brackets, I will simplify further, $\left\{\frac{k+1}{k-m} - 1\right\} = \frac{k+1-(k-m)}{k-m} = \frac{m+1}{k-m}$ and so I now I have a single term:

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

Recognize the common factor between in the denominator:

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

I cancel two terms and also use the fact that $(m+1)! = (m+1)m!$,

$$= \frac{k!}{m!\,(k-m)!}$$

Which is equivalent to:

$$= \binom{k}{m}$$

Therefore, it is clear that both sides are equivalent:

$$\binom{k}{m} = \frac{k!}{m!\,(n-k)!} = \binom{k+1}{m+1} - \binom{k}{m+1}$$

As required.

■

Now I will use (2) to work out the expectation. Recall the formula I determined for the expectation is as follows:

$$E(X_M) = \frac{m \cdot \sum_{\ell=m}^{N} \binom{\ell}{m}}{\binom{N}{m}}$$

Let's try working out a few terms. I will begin by working out the sum in part of the numerator,

$$\sum_{\ell=m}^{N} \binom{\ell}{m} = \binom{m}{m} + \binom{m+1}{m} + \binom{m+2}{m} + \binom{m+3}{m} + \cdots + \binom{N-1}{m} + \binom{N}{m}$$

Note that I can use my identity from (2) for this sum so that it will be easier to work with:

$$\sum_{\ell=m}^{N} \binom{\ell}{m} = \sum_{\ell=m}^{N} \left\{ \binom{\ell+1}{m+1} - \binom{\ell}{m+1} \right\}$$

By replacing each term by the corresponding value, as laid out above (and in (2)) I can see that this is, in fact, a telescoping sum.:

$$= \binom{m}{m} + \binom{m+2}{m+1} - \binom{m+1}{m+1} + \binom{m+3}{m+1} - \binom{m+2}{m+1} + \binom{m+4}{m+1} - \binom{m+3}{m+1} + \cdots + \binom{N}{m+1} - \binom{N-1}{m+1} + \binom{N+1}{m+1} - \binom{N}{m+1}$$

Once I've cancelled the terms I am left with (for the sum in the numerator):

$$\sum_{\ell=m}^{N} \binom{\ell}{m} = \binom{N+1}{m+1}$$

Therefore, once I plug this term back into the formula, replacing the sum with the above value the expectation of X_M is:

$$E(X_M) = \frac{m \cdot \binom{N+1}{m+1}}{\binom{N}{m}}$$

Using the identity (1), I have that $\binom{N+1}{m+1} = \frac{N+1}{m+1} \cdot \binom{N}{m}$ and when I plug that into the formula, I get:

$$E(X_M) = \frac{m \cdot \frac{N+1}{m+1}\binom{N}{m}}{\binom{N}{m}}$$

So, when I cancel the binomial coefficient from the numerator and denominator I am left with:

$$E(X_M) = \frac{m \cdot (N+1)}{m+1}$$

I now have a function that has N in it, so after rearranging we have a formula for N instead:

$$N = \frac{m+1}{m} E(X_M) - 1$$

Section 2.1.4 – Estimating N

I know X_M but I do not know $E(X_M)$ for certain, therefore I will introduce Y, a random variable that is dependent on the random variable X_M.

Formula used to estimate N:

$$Y = \frac{m+1}{m} X_M - 1$$

So, the expectation of Y should give N:

$$E(Y) = N$$

Remarks on Section 2.1 – Comparing to the naïve estimator
The naïve formula (the 'intuitive formula'):

$$N = x_{max} + x_{max}\frac{1}{sample\ size}$$

The frequentist formula:

$$\boldsymbol{Y = \left(1 + \frac{1}{m}\right) X_M - 1 = X_M + X_M \frac{1}{m} - 1, \quad E(Y) = N}$$

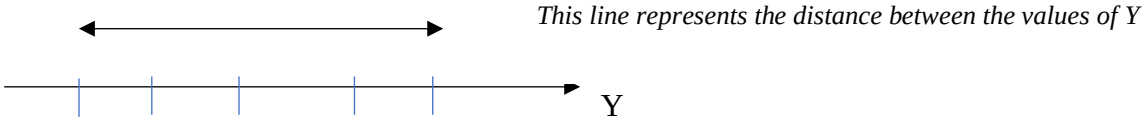
When I compare the two formulas, we see that the mathematical supported formula differs by one which implies that the naïve estimate overestimates, but somehow still gives a good approximation. Which helps to make sense of the mathematical approach because it so closely resembles our intuition. However, after reflecting on the intuition, taking away one from this formula and then taking the expectation of that formula was unexpected.

So, the frequentist formula differs in that it encompasses a mathematical tool, expectation, so essentially, just taking the average of the formula.

Section 2.2.1 – The Variance

Variance is used to measure the fluctuations between the different values of Y across the average value of Y, i.e. $E(Y)$. The variance is quantifying the distance of these fluctuations.

Visualization:



It's ideal for the top line to be small that would mean the variance is small. When the variance is big it means the $E(Y)$ is pretty much useless since Y isn't precise. To have a small variance means the value is likely accurate. Smaller variances are usually a result of having a larger sample size.

From the previous section, I know that $E(Y) = N$ so I want the variance to be small in order to have the most accurate representation of the total number, N. (Which would mean we want a large m.)

Section 2.2.1 – Calculation of Variance

Using the definition of variance, as stated below:

$$Var(Y) = E(Y^2) - (E(Y))^2$$

From the variance I can work out the accuracy of N.

Recall from the previous section, after rearranging, that the formula for the random variable Y is the following:

$$Y = \frac{m+1}{m} X_M - 1$$

So, plugging this into the equation for variance:

$$Var(Y) = Var(\frac{m+1}{m} X_M - 1)$$

So now the variance includes the random value X_M. I will use the following theorem to help me break this down.

Theorem 2.1 Let X be a random variable and a, b be real numbers.

$$Var(aX + b) = a^2 Var(X)$$

Proof: Using the definition of variance,

$$Var(aX + b) = E((aX + b)^2) - (E(aX + b))^2$$

Using rules of expectation, I can work out,

$$= E(a^2X^2 + 2abX + b^2) - (aE(X) + b)^2$$

$$= a^2E(X^2) + 2abE(X) + b - (a^2\big(E(X)\big)^2 + 2abE(X) + b^2)$$

$$= a^2(E(X^2) - \big(E(X)\big)^2)$$

$$= a^2 Var(X)$$

■

So, using both the definition of variance and the theorem, (letting a, b be the obvious terms) I get that I am looking for,

$$\boldsymbol{Var(Y) = (\frac{m+1}{m})^2 Var(X_M)}$$

So, once I know the variance of X_M I can determine the variance of Y from there. I know that, by the definition of variance:

$$Var(X_M) = E\left({X_M}^2\right) - (E(X_M))^2$$

From the previous sections, I already know what $E(X_M)$ is, so I just need to consider how to find $E({X_M}^2)$.

Key Observation: It is "easy" to compute $\mathrm{E}(\mathrm{X_M}(\mathrm{X_M}+1))$
Writing this out completely:

$$E\left(X_M(X_M+1)\right) = E\left({X_M}^2 + X_M\right) = E\left({X_M}^2\right) + E(X_M)$$

So, to summarize,

$$E\left({X_M}^2\right) = E\left(X_M(X_M+1)\right) - E(X_M)$$

Replacing $E\left({X_M}^2\right)$ in the previously stated $Var(X_M)$ equation. We get this **useful formula**:

$$Var(X_M) = E\left(X_M(X_M+1)\right) - E(X_M) - \left(E(X_M)\right)^2 \quad \textbf{(3)}$$

Here I just need to work out what $E(X_M(X_M+1))$ is. By definition (as previously stated in Section 2.1.3):

$$E(X_M) = \sum_{\ell=m}^{N} \ell\, P(X_M = \ell)$$

Therefore,

$$E\left(X_M(X_M+1)\right) = \sum_{\ell=m}^{N} \left(\ell(\ell+1)\right) P(X_M = \ell) = \sum_{\ell=m}^{N} \left(\ell(\ell+1)\right) \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}}$$

Recalling Proposition 2.1:

$$E\left(X_M(X_M+1)\right) = \frac{1}{\binom{N}{m}} \sum_{\ell=m}^{N} \left(\ell(\ell+1)\right) \binom{\ell-1}{m-1}$$

Now taking a quick break in to think about what's in the sum, $\ell(\ell+1)\binom{\ell-1}{m-1}$ in more detail. It would be nice to find an identity for:

$$\frac{\ell(\ell+1)}{m(m+1)}\binom{\ell-1}{m-1}$$

$$=\frac{\ell(\ell+1)}{m(m+1)}\cdot\frac{(\ell-1)!}{(m-1)!\big((\ell-1)-(m-1)\big)!}$$

$$=\frac{\ell(\ell+1)}{m(m+1)}\cdot\frac{(\ell-1)!}{(m-1)!\,(\ell-m)!}$$

$$=\frac{\ell(\ell+1)(\ell-1)!}{m(m+1)(m-1)!\,(\ell-m)!}$$

$$=\frac{(\ell+1)!}{(m+1)!\,(\ell-m)!}$$

$$=\frac{(\ell+1)!}{(m+1)!\,((\ell+1)(m+1))!}$$

$$=\binom{\ell+1}{m+1}$$

Which gives me the **nice identity**: (4)

$$\frac{\ell(\ell+1)}{m(m+1)}\binom{\ell-1}{m-1}=\binom{\ell+1}{m+1}$$

When I multiply by the denominator I get:

$$\ell(\ell+1)\binom{\ell-1}{m-1}=\binom{\ell+1}{m+1}m(m+1)$$

Now going back to my calculation of $E(X_M(X_M+1))$, I can now replace $\ell(\ell+1)\binom{\ell-1}{m-1}$ with my identity:

$$E\big(X_M(X_M+1)\big)=\frac{1}{\binom{N}{m}}\sum_{\ell=m}^{N}\binom{\ell+1}{m+1}m(m+1)$$

$$=\frac{m(m+1)}{\binom{N}{m}}\sum_{\ell=m}^{N}\binom{\ell+1}{m+1}$$

Taking another break here, I will use proposition 2.2,

$$\sum_{\ell=m}^{N} \binom{\ell}{m} = \binom{N+1}{m+1}$$

Now going to look at the sum I am interested in again,

$$\sum_{\ell=m}^{N} \binom{\ell+1}{m+1}$$

I will let $k = \ell + 1$, so I get the following,

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

Now again resuming my calculation of $E(X_M(X_M+1))$,

$$E\big(X_M(X_M+1)\big) = \frac{m(m+1)}{\binom{N}{m}} \cdot \binom{N+2}{m+2}$$

When I simplify this value:

$$= \frac{m(m+1)}{\frac{N!}{m!\,(N-m)!}} \cdot \frac{(N+2)!}{(m+2)!\big((N+2)-(m+2)\big)!}$$

$$= \frac{\big(m(m+1)\big)(m!\,(N-m)!)}{N!} \cdot \frac{(N+2)!}{(m+2)!\,(N-m)!}$$

$$= \frac{m}{N!} \cdot \frac{(N+2)!}{m+2}$$

$$= m \cdot \frac{(N+2)(N+1)}{m+2}$$

$$= \frac{m}{m+2}(N+2)(N+1)$$

Now using my useful variance formula, I can work out the following:

$$Var(X_M) = \underbrace{E\big(X_M(X_M+1)\big)} - E(X_M) - \big(E(X_M)\big)^2$$

$$= \frac{m}{m+2}(N+2)(N+1) - \left(\frac{m(N+1)}{m+1}\right) - \left(\frac{m(N+1)}{m+1}\right)^2$$

$$= \frac{m}{m+2}(N+2)(N+1) - \frac{m(N+1)}{m+1}\left(1 + \frac{m(N+1)}{m+1}\right)$$

$$= \frac{m}{m+2}(N+2)(N+1) - \frac{m(N+1)\big(m+1+m(N+1)\big)}{(m+1)(m+1)}$$

$$= m(N+1)\left(\frac{N+2}{m+2} - \frac{m+1+m(N+1)}{(m+1)(m+1)}\right)$$

$$= m(N+1)\left(\frac{N+2}{m+2} - \frac{m(N+2)+1}{(m+1)^2}\right)$$

$$= m(N+1)\left(\frac{(m+1)^2(N+2) - (m(N+2)+1)(m+2)}{(m+2)(m+1)^2}\right)$$

$$= m(N+1)\left(\frac{(m+1)^2(N+2) - m(N+2)(m+2) - (m+2)}{(m+2)(m+1)^2}\right)$$

$$= m(N+1)\left(\frac{(N+2)((m+1)^2 - m(m+2)) - (m+2)}{(m+2)(m+1)^2}\right)$$

$$= m(N+1)\left(\frac{(N+2)(1) - (m+2)}{(m+2)(m+1)^2}\right)$$

$$= m(N+1)\left(\frac{(N+2)(1) - (m+2)}{(m+2)(m+1)^2}\right)$$

$$= m(N+1)\left(\frac{N-m}{(m+2)(m+1)^2}\right)$$

So now I have what I need to calculate the variance of Y. Using the formula, I mentioned at the beginning of the section:

$$Var(Y) = (\frac{m+1}{m})^2 \underbrace{Var(X_M)}$$
$$= (\frac{m+1}{m})^2 \cdot (\frac{N-m}{(m+2)(m+1)^2}) m(N+1)$$

And after cancelling some terms and further simplifying, I am left with:

$$= \frac{1}{m}\left(\frac{N-m}{(m+2)}\right)(N+1) = \frac{N+1}{m}(\frac{N-m}{m+2})$$

So, the variance of Y is:

$$\boldsymbol{Var(Y) = \frac{N+1}{m}(\frac{N-m}{m+2})}$$

Which makes sense because as the sample size, m, gets larger the better (smaller) the variance.

Chapter 3 – The Analysis Through Simulation

Section 3.1: Simulations in R

The simulations were performed in R. The following is my input values, followed by an explanation of the layout of the code and the output for the code.

Input:
floor(runif(50,1,251))
I used this command to generate my sample of size 50 from a total of 250 integers starting at 1. The floor function ensures that my values are integers.

max(floor(runif(50,1,251)))
I used this code to find X_M I will use the max function in front of the previous.

So, to generate values for X_M, Y, and standard deviation (the square root of the variance) my code looks like:

```
xm<-replicate(10,max(floor(runif(50,1,251))))
xm
        ((50+1)/50)*xm-1
```

sqrt(((250+1)/50)*((250-50)/(50+2)))

Output:

```
[1] 248 250 249 250 250 246 244 250 245 249

 [1] 251.96 254.00 252.98 254.00 254.00 249.92 247.88 254.00 248.90 252.98

[1] 4.394052
```

Summary of Key points

- "The German Tank Problem" was a population estimator problem used by the Western Allies during WWII that aimed to use statistical methods to find a precise estimate of monthly tank production by the Germans based on a sample of serial numbers from captured or destroyed tanks.

- The Frequentist approach to this problem was to assume there is a fixed number of tanks produced in a particular month, N, and construct a probability mass function (pmf) using a random variable to denote the maximal value of the sample set.

- The general pmf for when the maximum is some value $\ell, m \leq \ell \leq N$, looks like:

$$\mathrm{P}(X_M = \ell) = \frac{\binom{\ell-1}{m-1}}{\binom{N}{m}}$$

- The pmf is used and manipulated using binomial coefficient identities to produce the expectation of the maximal value:

$$E(X_M) = \frac{m \cdot (N+1)}{m+1}$$

- After some carefully rearranging, an equation with a new random variable, Y, dependent on X_M is used to solve for an estimation of the total number of tanks:

$$Y = \frac{m+1}{m} X_M - 1, \qquad E(Y) = N$$

- Variance is used to measure the fluctuations between the different values of Y across the average value of Y, i.e. $E(Y)$. The variance quantifies the distance. The variance

gives a picture of how accurate the estimate is. The formula for variance for this problem is given by:

$$Var(Y) = \frac{N+1}{m}\left(\frac{N-m}{m+2}\right)$$

- "The German Tank Problem" has been hugely influential in regard to the power of Mathematics, application of statistics, wartime tactics and the end of WWII.

References

[1] Michael Hochster, PhD, Quora, https://www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians, 2016

[2] Queen Mary University of London, Introduction to Statistics Notes, 2020

[3] Wikipedia, German tank problem, https://en.wikipedia.org/wiki/German_tank_problem, 2019

[4] Wikipedia, Frequentist Inference, https://en.wikipedia.org/wiki/Frequentist_inference, 2018