Mathematics MSc Dissertation MTHM038, 2020/21

# The German tank problem

A comparison between the Bayesian approach and Frequentist approach.


**Dalal Almudhaf, ID 190888765**

Supervisor: Dr. Wolfram Just


A thesis presented for the degree of

Master of Science in *Financial Mathematics*


School of Mathematical Sciences and *school of Economics and Finance*

Queen Mary University of London

# Declaration of original work

This declaration is made on September 3, 2021.

**Student's Declaration:** I Dalal Almudhaf hereby declare that the work in this thesis is my original work. I have not copied from any other students' work, work of mine submitted elsewhere, or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

## Acknowledgment

I express my sincere gratitude to Dr. Wolfram Just for his guidance and kind help given to me throughout this project, with his suggestions and instructions has served as the major contributor towards the completion of the project. The lecture note he provided, gave me more insight and understanding.

# Abstract

The German tank problem is a problem that came about during World War II, it is about estimating the number of tanks produced by Germany, by having serial numbers from random samples. The project is about determining the size of a population from a random sample.

In this thesis, we will start with a historical background of the problem and the methods used to solve it at that time, throughout the project both Bayesian and Frequentist approaches are explained with comparison between both of them.

# Table of Contents


Chapter 1- The Introduction ........................................................................................ 5

Section 1.1: Introduction ......................................................................................... 5

Section 1.2: The History – “The German Tank Problem” .................................................. 6

Section 1.3: Deriving with the Known Minimum ............................................................... 8

Chapter 2 - The Frequentist Approach “German Tank Problem” ......................................... 10

Section 2.1: Introduction to Frequentist Approach ............................................................ 10

Section 2.1.1: The Analytics ......................................................................................... 10

Chapter 3 – Bayesian Approach “The German Tank Problem” ............................................ 17

Section 3.1: Introduction to the Bayesian Approach .......................................................... 17

Section 3.2: Using the Improper Uniform Prior.................................................................. 18

Section 3.3: Posterior Quantiles and Highest Posterior Density (HPD) intervals .............. 18

Chapter 4 - Frequentist Approach and Bayesian Approach Compared. ................................ 20

Section 4.1: Comparison Between Frequentist Approach and Bayesian Approach ........... 20

Section 4.2: Comparison of Errors ...................................................................... 21

Section 4.3: Assessment with the other priors ................................................................... 22

References ....................................................................................................... 23

Table1: Estimate comparison from statistics and spies to the actual values ............................ 7

Table 2: Findings of the Bayesian Approach ........................................................................ 19

Table 3: Assessment of Bayesian Approach with other Priors ............................................. 22

# Chapter 1- The Introduction

### Section 1.1: Introduction

At the time of World War II, the armies of Germany used tanks to devastating advantage. The allies required estimates which were accurate of their production and deployment of tanks. Two approaches were applied to find the values, first statistics and secondly, the spies (Goerlitz, Battershaw& Millis, 2019). If we assume that the tanks are consecutively labelled beginning with 1: and then we observe the $k$ serial numbers for the unknown number of tanks ($N$) and observed maximum value $m$, hence the best estimate for $N= m(1+1/k)-1$ (Goerlitz, Battershaw& Millis, 2019). This is the German Tank Problem and is a practical application of statistics and mathematics in the real world.

Generally, in statistics, there are several problems which include the estimation of the size of a population. For this reason, statistical methods and instruments are being put in place depending on the size of the sample of a population to help in the approximation or estimation of the size of the population (Bortolussi et al., 2021). The focus of the German Tank Problem, as was applied by the military during World War II, was to give an approximate number of the tanks that were made by the Germans from a sample size by use of the serial numbers of the sample.

Based on the German Tank Problem, there are two main statistical approaches that are being applied and are the use of the frequentist approach and the Bayesian approach. According to frequentist approaches, the data observed is a sample of selected distribution. This distribution of data is known as the possibility: $P(\theta)$, and the goal remains to discover the $\theta$ that maximizes the likelihood (Bortolussi et al., 2021). In logistic regression, for example, data is taken from a Bernoulli distribution, whereas in linear regression, data is sampled from a Gaussian distribution.

In Bayesian approaches, both data and hypotheses are assumed to have probabilities (parameters specifying the distribution of the data). Among the assumptions made by Bayesians are an earlier hypothesis distribution P (θ) and the possibility of data P (θ) (Berg &Hawila, 2020). The prior subjectivity remains a major criticism on the Bayesian reasoning, since various priors can lead to diverse conclusions and posteriors.

In this research paper, emphasis will be laid on both the statistical approaches (Frequentist approach and the Bayesian approach) as brought out in the German Tank Problem to solve the practical problems. Furthermore, the paper will illustrate how different basic and advanced concepts of statistics like unbiased estimation, testing of the hypothesis, and maximum likelihood estimation are being applied to find the size of a population from a random sample. Besides, different fundamentals of statistics will be illustrated. The paper has been arranged to start from the introduction and then followed by the history of the German Tank Problem being explained. In the following pages, both the frequentist and Bayesian approaches are discussed. Furthermore, the differences between the two approaches are explained, and lastly, a conclusion is drawn.

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

In the year 1943, the American Embassy's Division of Economic Warfare in the UK began to analyze the serial numbers of the German equipment captured to estimate the production and strength of Germany. The order of the products to be analyzed included: tires, tanks, trucks, guns, flying bombs, and finally, the rockets (Shams, El-Banbi, &Sayyouh, 2017). There was different intelligence derived from the markings of various products. According to Biba (2020), the gathered intelligence assisted in the target systems decision of the air forces, and it is revealed in different instances the German strength in weapons like rockets and tanks.

Subsequent availability of the official statistics of the war on the German war production, it became possible to evaluate the estimate's accuracy that was made.

The second part presented a summary of the estimates of the together with the data that is official besides a comprehensive treatment of particular estimations. The applied markings in analysis of the tanks' were the serial numbers derived both from the captured and destroyed equipment (Tiquet, 2021). The serial numbers provided a small but reliable sample. One assumption was made by the statisticians. Germans logically numbered their tanks in the order of production. This made it possible for those who practice statistics to approximate the total number of produced tanks in a given time duration.

This strategy was key for the battle to be won and assisted the allies in having a better understanding of their opponent. The serial number analysis was not only used in the estimation of production, but it was also used in understanding the general production of Germany, which included; the supply chain length (grounded on the production and use interval), number of factories, and their relative importance, variance in production and application and resource use (Prosdocimi, 2018). After the war, the Ministry of Speer previously tasked with war production of Germany was discovered. The special studies conducted after the war revealed that the estimates of the US and British were very accurate and timely more than the estimates of Germany, as shown in the table below.

| Month | Estimated Statistics | Intelligence Estimate | Records of Germany |
|---|---|---|---|
| Jun-1940 | 169 | 1000 | 122 |
| Jun-1941 | 244 | 1550 | 271 |
| August-1942 | 327 | 1550 | 342 |

*Table1:Estimate comparison from statistics and spies to the actual values. Source: (Mukherjee, 2019)*

The case study of the German Tank dilemma has demonstrated that using statistics to solve such challenges is far more successful than using traditional intelligence-gathering methods. During World War II, the same strategy was used.

**Section 1.3: Deriving with the Known Minimum**

In order to prove this formula: $N = m(1+1/k)-1$

Having a look at k tanks, largest one is marked m, and we know one is the least number. The tanks are sequentially numbered. We look at some extreme instances as a smell test before proving it. For starters, there is no estimation which is smaller as compared to the biggest seen number. Secondly,  In case of several tanks but  only one is seen one (i.e., k = 1), N should be about 2m (Tausch, 2020).

The estimate is acceptable because it basically indicates that if we just have one data point, we can safely assume it was somewhere in the middle. In addition, as we ask more questions, the amount we have to increase the seen highest value drops. If k = 2, for example, m is inflated by about 3/2 factor, implying that our highest observed value is most likely to be about 2/3 of the genuine figure (Castillo et al., 2019). Finally, if k equals N, m therefore must be the similar to N, simplifying this expression ass N = N.

This proof is dividedin two sections. Whereas there is luck because a closed expression form can be can be acquired a closed-form expression, if a fair estimation on the relationship can be established, statistics can be used  in examining the reasonableness of it; To begin the evidence, the probability that the greatest figure equals m is calculated.  Computation is then done on the expected value, and we demonstrate how to convert the expected value to an N. There is need for two combinatorial results.

The first pascal's identity is:

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

The simplest way to demonstrate this is to consider the two sides as distinct methods of numbering how many methods we can select r persons in a population of n + 1 people, where correctly n of this population are in a single set and there is accurately one individual in the other. It's easier to see now. if we rewrite it like this:

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

Because of this, it is acceptable $\frac{1}{0} = \frac{1}{1} = 1$. Note that the left side is selecting r people from a combined group of n + 1 persons, but the right side is selecting r people with the first summand indicating that we should not select a person from a group of only one person, and the second summand indicating that we must select that person.

The second identity is based on binomial coefficient sums:

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

For the inductive step, we assume

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

Then

$$\sum_{m=k}^{N+1} \frac{m}{k} = \sum_{m=k}^{N} \frac{m}{k} + \frac{N+1}{K}$$

$$= \binom{N+1}{k+1} + \binom{N+1}{k}$$

$$= \binom{N+2}{k+1}$$

Where the final equality is following from pascal's identity, and this concludes the proof.

**The probability of m being the maximum sample**. Assume that the random variable is M for the highest observations while value observed is m. It's worth noting that witnessing a number that is either smaller or Larger than N has a 0% chance of happening. We assert that in the case of r k ≤ m ≤ N that:

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

# Chapter 2 - The Frequentist Approach "German Tank Problem"

### Section 2.1 Introduction to Frequentist Approach

Many things are assigned a serial number, which usually begins with one and increases by 1 with each subsequent unit. During World War II, for example, German tanks possessed a lot of serialized parts. Allied statisticians might estimate the overall number of manufactured tanks by Germans over some given period of time by gathering values of these figures (Fehlmann&Kranich, 2017). According to Bortolussi et al. (2021), the notion on serial numbers faced in the field existed to be samples from a discrete uniform population distribution with a beginning value of one and ending at an anonymous value of N, in which N represents the real number of built tanks. The difficulty here is estimating value of N using facts from captured/destroyed tanks in battle.

If you pull k balls with replacement from an urn with balls numbered from one to N, the greatest observed value in the sample becomes credible N estimation.

### Section 2.1.1: The Analytics

This is demonstrated through a simple simulation:

Sample Urn <- function (k, N = 10000, r = 100) {

```
    urn <- 1: N
    sample Max <- replicate (r, max (sample (urn, k, replace = TRUE)))
    sample Max
}
k <- seq (1, 10000, length .out = 100)
sample Max <- apply (k, sample Urn)
sample Max <- data. Frame (max Value = do. Call (c, sample Max), k = rep (k, each = 100))
ggplot (sample Max, aes (x = k, y = max Value)) + geom_ point () + scale_ y_ continuous
("Sample Maximum")
```

Despite their best efforts, the Allies were unable to capture enough tanks to rely on asymptotic theory. The bias of this estimate is high for small sample sizes. There is sense in this because maximum sample cannot be more than, but must be the similar to, maximum value of support and any chance of a draw in the sample is proportional to the size of the draw (and sample space).

Furthermore, tanks were not being sampled with replacement. The serial numbers of the tank were discovered after it the capture or disability in combat. As a result, after examination of the tank, there cannot be anymore observation. Hence slightly complicating pmf for M, however it offers a very accurate N estimate for a sample fixed than minus it. Nonetheless, it is possible to produce an N, N estimator which is not biased for small size sample as well as reflecting the fact that there can be creation of sampling without replacements. By beginning, consider the following:

E (M; k) $=\sum_{m=0}^{N} mPr(m)$

Or the maximum sample expected, M, provided a K draw size is majorly the total number of m products and pr (m) as m $\in$ 1, 2, We can separate the summing into two sections if we notice that we have seen k tanks and that N<k is not conceivable because we are sampling without replacement and have seen k distinct tanks.

E (M;k) = $\sum_{m=0}^{k-1} mPr(m) + \sum_{m=0}^{N} mPr(m)$

Because Pr (N-=m) = 0 for $m$<k, this reduces to

E (M; k) = $\sum_{m=0}^{N} mPr(m)$

The number of methods to select k1 tanks in a sample set of m1 tanks divided by the sum total of methods to select k tanks in a total population set of N tanks is represented by the pmf of M since the sampling is without replacement. This is my least favorite section of math statistics after questions involving cards, yet it looks as though the equation is given by

Pr (M=m) = $\frac{\substack{m-1\\k-1}}{\substack{n\\k}}$

If this is plugged for Pr(m) in the above summation, we come up with something like this

E (M; k) = $\sum_{m=k}^{N} m \, \frac{\substack{m-1\\k-1}}{\substack{N\\k}}$

Even without expanding the binomial coefficients into a factorial form, this is unattractive. As a result, we focus on expansion.

E (M; k) = $\sum_{m=k}^{N} m \, \frac{\frac{(m-1)}{(k-1)!((m-1)-(k-1))!}}{\substack{N\\k}}$

And if we get them into the expression

E (M; k) = $\sum_{m=k}^{N} m \, \frac{\frac{m!}{(k-1)!((m-1)-(k-1))!}}{\substack{N\\k}}$

If the expansion could be converted back to a binomial coefficient, that would be much nicer. It's close to $\substack{m\\k}$, however a $\frac{1}{k}$ The multiplier is required. We don't modify the value by multiplying by k both the top of the expansion and bottom of the expansion, and this result in:

$$E\ (M;\ k) = \sum_{m=k}^{N} \frac{\frac{km!}{k(k-1)!((m-1)-(k-1))!}}{\substack{N \\ k}}$$

Which can be written as:

$$E\ (M;\ k) = \sum_{m=k}^{N} \frac{k\left[\frac{m}{k}\right]}{\left[\frac{N}{k}\right]}$$

This doesn't appear to be much of an improvement at a first impression. However, we now only get the totals of one binomial coefficient, with no multiplications taking on within totals. Remember that we previously had the pmf of ***M*** as

$$Pr\ (M=m) = \left(\frac{\substack{m-1 \\ k-1}}{\substack{n \\ k}}\right)$$

Going by what pmf refers to, the totals of Pr (M=m) with m support is 1. Since Pr (m) values are 0 for **m<k,** it results in

$$1 = \sum_{m=k}^{N} m \left(\frac{\substack{m-1 \\ k-1}}{\substack{N \\ k}}\right)$$

There can be pulling out of the $\frac{1}{\substack{N \\ k}}$ to give

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

For this expression to be true.

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

Returning to the expectation, $\sum_{m=k}^{n} \binom{m}{k}$ Produces a similar result, however with the "1" part missing. Since we are aware that there is no effect on the expression's even by making changes to the indices by either adding or subtracting a constant value as long as all the indices are changed, the expectation can again be written as

$$\mathrm{E}(\mathrm{M};\mathrm{k}) = \frac{1}{\binom{N}{k}} \sum_{m=k+1}^{N} \binom{m-1}{k-1}$$

Let's get an expression written in the right formula with it depending on the application of probability mass function.

Considering correlation $\binom{N}{k}$= $\sum_{m=k}^{N} \binom{m-1}{k-1}$ and replacing, we reach to

$$\mathrm{E(M;k)} = \frac{1}{\binom{N}{k}} \binom{N+1}{k+1}$$

No more unpresentable summations. Expanding the binomial coefficients, derives the equation to:

E (M; k) = k $\frac{\frac{(N+1)!}{(k+1)!(N-k)}}{\frac{N!}{k!(N-k)!}}$

E (M; k) = $\frac{k(N+1)!k!}{(k+1)!N!}$

Upon realizing that $\frac{K!}{(K+1)}$ is cancelling up to $\frac{1}{k+1}$ and that $\frac{(n+1)!}{n!}$ minimizes up to n+1, this is then written as E (M; k) = $\frac{k(N+1)}{(k+1)}$

Solving the N equation,

N = $\frac{Mk+M}{k} - 1$

Hence,

E(N) = E $\left(\frac{Mk+M}{k} - 1\right) = \frac{Mk+M}{k} - 1$

Note that it can still be toned down to E (N) = M $\left(1 + \frac{1}{K}\right) - 1$

For simplicity.

In order to establish the estimator performance against the application of mere maximum sample, in collecting a justly minor sample, the practical number of N to be used is 300 under the historical context.

```
germanTankSim <- function(k, N) {
    tanks <- 1:N
    m <- max(sample(tanks, k))
    nhat <- m * (1 + 1/k) - 1
    nhat
}
sampleMaxEstimator <- function(k, N) {
    tanks <- 1:N
    m <- max(sample(tanks, k))
    m
}

N <- 300
k <- round(seq(0.01, 1, 0.01) * N)
k[k == 0] <- 1  # doesn't make sense with k = 0
tanks <- sapply(rep(k, each = 1000), germanTankSim, N)
tanks <- data.frame(adjustedMax = tanks, k = rep(k, each = 1000))
tanks$sampleMax <- sapply(rep(k, each = 1000), sampleMaxEstimator, N)
```

When the observation of the corrected estimate of the estimator is done as a determinant of tanks that are seen, a straight decline in estimate of the variance and a relative less biased N estimator are visualized.

```
ggplot (tanks, aes (x = k)) + geom_jitter (aes (y = adjusted), alpha = 0.1, position =
position jitter (width = 1,

    height = 1)) + geom_hline (aes (yintercept = 300), lty = 2) + scale_x_continuous
("Number of Observed Tanks") +

    scale_y_continuous ("Unbiased Estimate of Number of Total Tanks")
```

Moving towards the sample maximum, it can clearly show that the biasness leans towards the low value of total number of tanks. When the probability is provided for a drawn of small size, for example, sample population of 10%, having the highest population value, it becomes reasonable. The bias reduces as the number tanks observed increases and tanks converges to the exact figure when $k \sim N$.

```
ggplot (tanks, aes (x = k)) + geom_jitter (aes (y = sampleMax), alpha = 0.1, position
= position jitter (width = 1,

    height = 1)) + geom_hline (aes (yintercept = 300), lty = 2) + scale_x_continuous
("Number of Observed Tanks") +

    scale_y_continuous ("Sample Max")
```

Computation of the mean square error where k value for both the estimators, we can see that the estimator that is adjusted shows a more accurate N estimation as compared to the raw sample mean.

```
mse <- function (value, N) {

    mean ((value - N) ^2)

}

esd <- aggregate (cbind (tanks$adjustedMax, tanks$sampleMax), by = list(tanks$k),

    mse, N = 300)

names(esd) <- c ("k", "adjustMax", "sampleMax")

esd <- data. Frame (k = rep (esd$k, 2), mse = c (esd$adjustMax, esd$sampleMax),

    estimator = rep (c ("Adjusted", "Sample Max"), each = 100))

ggplot (esd, aes (x = k, y = mse, color = estimator)) + geom_line()
```

As it began with the tank during World War II, this method can now be applied practically for the estimation of the large values depending on the sequential serial numbers. For example, it can be used to compute the end of humanity.

# Chapter 3 – Bayesian Approach “The German Tank Problem”

## Section 3.1: Introduction to the Bayesian Approach

Bayesian line of thought to the German Tank Problem is to take into consideration that the probability of the number of tanks for the enemy is equivalent to number of tanks observed and equal to, and the largest serial numbers are equal to.

There is an assumption for an appropriate N distribution; then, the distribution of posterior is to be computed through theorem of Bayes. For example;

$$P\,(N/X_n) = \frac{P\left(\frac{Xn}{N}\right)P(N)}{P(Xn)} = \frac{P\left(\frac{Xn}{N}\right)P(N)}{P\left(\frac{Xn}{N}\right)P(N)} \qquad \text{for } X_n \leq N < \infty \text{ and 0, otherwise}$$

various choices can be taken as an initial N distribution.

An improper uniform before on every positive integer, for example;

$P(N) \infty 1$ for $N = 0, 1,,,,,,, \infty$ and 0, otherwise

A proper uniform distribution with an upper limit k for N, for example;

$P(N) = \frac{1}{K+1}$ for $0 \leq N \leq k$ and 0, else.

A geometric, Poisson or Negative Binomial distribution.

**Section 3.2: Applying the Improper Uniform Prior:**

Under the improper uniform prior, the posterior distribution is then given by,

P (N/ $X_n$) = $\frac{\frac{n-1}{n}\binom{Xn}{n}}{\binom{N}{n}}$ If N = $X_n$, $X_n$ + 1,,,,,,,, ∞ and 0, else which is a shifted factorial distribution.

The posterior distribution is enormously skewed to the positive. The mode of posterior is at $X_n$.

The posterior mean is E (N/ $X_n$) = $\frac{(n-1)(Xn-1)}{n-2}$ for n > 2

And the posterior variance V (N/ $X_n$) = $\frac{(n-1)(Xn-1)(Xn-n+1)}{(n-2)2(n-3)}$ for n > 3

Quantile measures are more appropriate than moment measurements for summarizing posterior information since the distribution of posterior is strongly skewed positively. In this scenario, posterior median becomes the appropriate estimate of N, and the posterior quartile deviation is the measure of estimate accuracy.

**Section 3.3: Posterior Quantiles and Highest Posterior Density (HPD) intervals**

Now we'll look at how to calculate posterior quantiles. Let $N_q$ be quantile of the posterior be, i.e. the smallest integer satisfying; $N^i$

$$\sum_{Ni}^{Nq} = X_n \, P(X_n/N^i) = \sum_{Ni}^{Nq} = Xn \, \frac{\frac{n-1}{n}\binom{Xn}{n}}{\binom{N}{n}} \geq q$$

Or equivalently

$$\sum_{Ni}^{Nq} = X_n \, P\,(X_n / N^i) = \sum_{Ni}^{\infty} = \text{Nq} + 1 \;\; \frac{\frac{n-1}{n}\binom{Xn}{n}}{\binom{N}{n}} = \frac{(\text{Xn}-1)!(\text{Nq}-\text{n}+1)!}{(\text{Xn}-\text{n})!\text{Nq}} \leq 1 - q$$

This allows us to make calculation on the median and any posterior quantile minus having to sum the posterior distribution explicitly. The real-world polynomial to be resolved $N_q$ is

$$N_q\,(N_q - 1) \ldots (N_q - n + 2) - \frac{(\text{Xn}-1)!}{(1-q)(\text{Xn}-1)!} = 0$$

where $N_q = [N_q]$, the smallest integer large than $N_q$.

Since the posterior distribution mode is constantly at $X_n$ and the posterior distribution is monotone decreasing as N is increasing, calculation of the highest posterior density (HPD) intervals reduces to the calculation of quantiles of the posterior. For example, the HPD-interval of level q equals [ $X_n$, $X_n$, + 1, Nq ]

**Results found;**

The number N is set to 1000. The necessary simulated data is acquired and demonstrate computation using R. The simulation and computation R code were adapted from the work. For sample sizes of n=5, 10, 20, 50, 75 and 100, we make the necessary calculations. The following table summarizes the findings:

| n | Frequentist Estimate | Estimate of Standard error | Bayesian Estimate (Improper Uniform Prior) | Posterior Quartile Deviation | Shortest 95% Confidence Interval [Xn, L] | 95% Highest Posterior Density Interval (Xn, L) |
|---|---|---|---|---|---|---|
| 5 | 1123.40 | 187.00 | 1113.00 | 159.00 | 1702.00 | 1978.00 |
| 10 | 1097.90 | 99.40 | 1078.00 | 66.80 | 1345.00 | 1391.00 |
| 20 | 1029.05 | 48.55 | 1017.00 | 29.50 | 1136.00 | 1146.00 |
| 50 | 975.14 | 18.63 | 970.00 | 10.50 | 1013.00 | 1015.00 |
| 75 | 1008.28 | 12.77 | 1004.00 | 7.00 | 1034.00 | 1035.00 |
| 100 | 1009.00 | 9.49 | 1006.00 | 5.00 | 1027.00 | 1028.00 |

*Table 2: Findings of the Bayesian Approach. Source: (Mukherjee, 2019)*

# Chapter 4 - Frequentist Approach and Bayesian Approach Compared.

### Section 4.1: Comparison Between Frequentist Approach and Bayesian Approach

On average, the estimates on frequentist as well as Bayesian (with incorrect prior uniform) are close to the true value of the parameter. In both approaches, variability in the estimates diminishes as the sample size grows, as expected.

The Bayesian approach continually yields lesser estimates as compared to frequentist approach. Furthermore, the frequentist technique standard error as well as the posterior quartile deviation in the Bayesian method reflects some variability in the estimations in these two approaches (Chen et al., 2018). As a result, the Bayesian approach's point estimates are better, at least in these cases.

When interval estimates produced by the two methodologies are compared, it's important to remember that the two approaches' interpretations of the data differ. A confidence interval of 95% shortest length makes sure that a supplied interval includes real parameter value on average in the 95 % repeats of a fundamental random experiment, according to theory of frequentist method (Guo et al., 2017).

The 95 percent HPD interval in the Bayesian approach provides the variety in the parameters of the posterior distribution that covers 95 percent of an entire curved area, with all interval points having an advanced density or mass compared to some points not within the interval, hence presenting variety of values of reasonable parameter (Berg, 2021). In terms of the intervals length, the N estimate of the interval for these two different approaches can be compared. The situation is therefore seen that the frequentist technique yields shorter intervals.

**Section 4.2: Comparison of Errors;**

The following graph compares the errors (error = value estimated - true value) for samples of different sizes in the two different approaches:

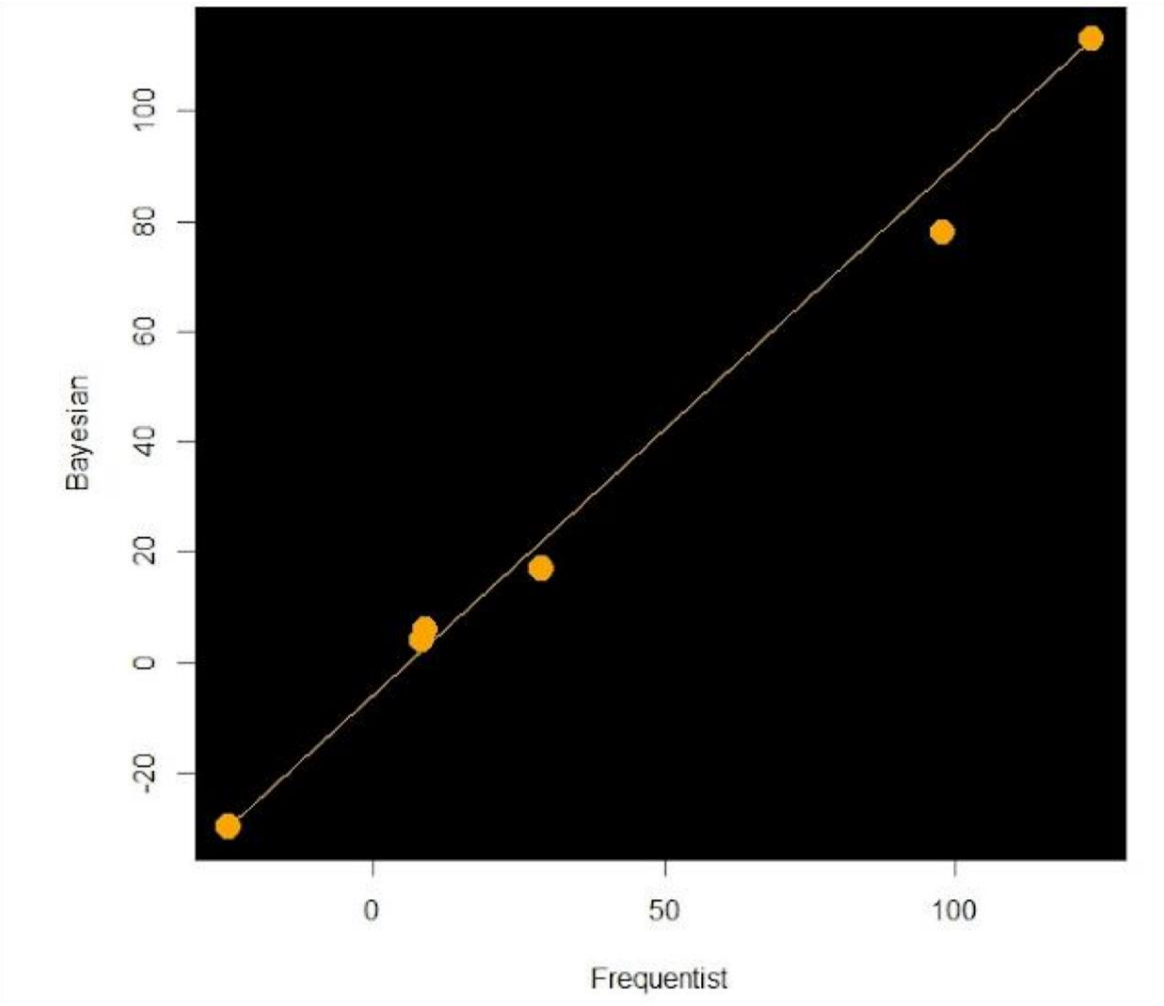


*Figure 1: Comparison between the two approaches. Source: (Mukherjee, 2019)*

It is evidenced that the errors are closely the same for the Frequentist Approach and Bayesian Approach because the line y=x is very close to the lying points.

**Section 4.3: Assessment with the other priors.**

We use the right uniform prior (choosing 2000 as the top limit, which is an acceptable choice) and the negative binomial prior to performing identical calculations. We use the right uniform prior (choosing 2000 as the top limit, which is an acceptable choice) and the negative binomial prior to performing identical calculations. Below table summarizes the results achieved utilizing various priors in comparison to the estimates obtained using the improper uniform prior:

| **n** | **Bayesian Estimate (Proper Uniform Prior)** | **Posterior Standard deviation (Proper Uniform Prior)** | **Bayesian Estimate (Negative Binomial Prior)** | **Posterior Standard deviation (Negative Binomial Prior)** |
|---:|---:|---:|---:|---:|
| 5 | 1176.77 | 230.48 | 1788.15 | 274.61 |
| 10 | 1120.62 | 131.60 | 1578.84 | 254.89 |
| 20 | 1034.44 | 57.01 | 1228.68 | 177.04 |
| 50 | 975.92 | 19.82 | 991.38 | 35.16 |
| 75 | 1008.63 | 13.30 | 1013.90 | 18.59 |
| 100 | 1009.19 | 9.78 | 1011.80 | 12.41 |

*Table 3: Assessment of Bayesian Approach with other Priors. Source: (Mukherjee, 2019)*

It can be observed that the above priors are performing worse compared to the improper uniform prior for the case where points are estimated.

## References

Berg, A. (2021). Bayesian Modeling Competitions for the Classroom. *RevistaColombiana de Estadística*, *44*(2), 243-252.

BERG, A., & HAWILA, N. (2020). INTRODUCING BAYESIAN INFERENCE WITH THE TAXICAB PROBLEM.

Biba, S. (2020). Ganging up on Trump? Sino-German Relations and the Problem with Soft Balancing against the USA. *Journal of Chinese Political Science*, *25*(4).

Bortolussi, L., Cairoli, F., Paoletti, N., Smolka, S. A., &Stoller, S. D. (2021). Neural predictive monitoring and a comparison of frequentist and Bayesian approaches. *International Journal on Software Tools for Technology Transfer*, 1-26.

Bortolussi, L., Cairoli, F., Paoletti, N., Smolka, S. A., &Stoller, S. D. (2021). Neural predictive monitoring and a comparison of frequentist and Bayesian approaches. *International Journal on Software Tools for Technology Transfer*, 1-26.

Castillo, O., Valdez, F., Soria, J., Amador-Angulo, L., Ochoa, P., &Peraza, C. (2019). A comparative study in fuzzy controller optimization using bee colony, differential evolution, and harmony search algorithms. *Algorithms*, *12*(1), 9.

Chen, J., Huang, B., Ding, F., &Gu, Y. (2018). Variational Bayesian approach for ARX systems with missing observations and varying time-delays.*Automatica*, *94*, 194-204.

Fehlmann, T. M., &Kranich, E. (2017, October). A new approach for continuously monitoring project deadlines in software development. In *Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement* (pp. 161-169).

Goerlitz, W., Battershaw, B., & Millis, W. (2019). *History of the German General Staff*. Routledge.

Guo, F., Kodamana, H., Zhao, Y., Huang, B., & Ding, Y. (2017). Robust identification of nonlinear errors-in-variables systems with parameter uncertainties using variational Bayesian approach. *IEEE Transactions on Industrial Informatics*, *13*(6), 3047-3057.

Mukherjee, S. (2019, March 21). *The German Tank PROBLEM: Frequentistvs Bayesian approach*. The German Tank Problem: Frequentistvs Bayesian Approach. http://statscuriousminds.blogspot.com/2019/03/the-german-tank-problem-frequentistvs.html.

Prosdocimi, I. (2018). German tanks and historical records: the estimation of the time coverage of ungauged extreme events. *Stochastic environmental research and risk assessment*, *32*(3), 607-622.

Shams, M., El-Banbi, A., &Sayyouh, H. (2017, April). A comparative study of proxy modelling techniques in assisted history matching. In *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*. OnePetro.

Tausch, A. (2020). The political geography of Shoah knowledge and awareness, estimated from the analysis of global library catalogues and Wikipedia user statistics. *Jewish Political Studies Review*, *31*(1/2), 7-123.

Tiquet, P. (2021). *German Tank Destroyers*. Casemate.