德国坦克问题

左泽成 2022213434
丁文皓 2022213431

2024 年 5 月 4 日

1 历史背景

在战争的过程中，西方盟国一直在努力确定德国的生产速度，并以两种主要方式得到这个数据：常规情报收集和统计估计。大部分情况下，统计估计的可信度很大程度上决定于常规情报收集。而有时常规情报收集会和统计估计一起结合使用，如在 D 日前夕估计豹式坦克的生产速度。

盟军指挥机构已经注意到了出现在意大利的五号坦克（豹式坦克），速度快，装备有 75 mm 长管战车炮，是很不寻常的重型坦克，但在法国北部编号就很小，就和虎 I 坦克在突尼斯那样。由于谢尔曼坦克在对阵三号坦克和四号坦克时表现良好，美国陆军对它非常自信，以至于迫不及待地希望与他们相遇。就在 D 日的不久前，有消息称德国手中有大量的五号坦克。

为了确定这种说法的真实性，盟军要尝试估计正在生产坦克的数量。为此，他们要使用被缴获或被摧毁的坦克的序列号。他们主要使用变速箱的号码，因为其为两个完整的序列。底盘和发动机号码也有使用，但它们的编号规则比较复杂。各种其他的零件用来对分析结果进行交叉检查。对轮胎也做了类似的分析，其上观察到了顺序编号（即 1、2、3、……、N）。

对坦克车轮的分析产生了对使用中的车轮模具数量的估计。在与英国车轮制造商讨论过后，他们估计了这么多的模具可以生产多少车轮，进而是每个月可生产的坦克数量。对两辆坦克（每辆 32 个车轮，总计 64 个车轮）车轮的分析的结果是 1944 年 2 月的生产数量估计在 270 左右，大大超出此前预期。

德国战后公布的记录显示，1944 年 2 月一个月的生产量是 276 辆。统计方法结果的精确度是常规情报收集方法所远远不能达到的，而“德国坦克问题”这个词也成为了这种统计分析问题的标志。

估计产量并不是这种序列号分析的唯一应用。它也用于探查更多德国生产的信息，包括工厂数目、工厂的相对重要性、供应链长度（基于生产和使用之间的滞后程度）、生

产工艺的改变、及对诸如橡胶等资源的使用。

2 问题描述

在第二次世界大战期间，德国军队装备有 n 辆坦克。每辆坦克上都刻有一个唯一的序列号，编号属于集合 $\{1,...,n\}$。作为盟军，我们不知道 n 的值，但我们已经缴获了一些德国坦克的样本（当然是无替换地缴获），这些样本的序列号是有序的（$s_1,...,s_k$）。假设在所有坦克中，每辆坦克被缴获的可能性都是相等的，而且 n 是固定的，我们的目标是根据数据（$s_1,...,s_k$）来估计 n。

3 频率论分析

3.1 最小方差无偏估计（MVUE）的定义

对于统计量 $\hat{\theta}$，其最小方差无偏估计（MVUE）的数学定义为：

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

其中，m 是观测到的最大标记（最大样本值），k 是已知的平均标记（平均样本值）。这是一个最小的估计，意味着没有更小的估计值。

3.2 方差

方差的定义如下：

$$\mathrm{var}(\hat{N}) = \frac{1}{k}\frac{(N-k)(N+1)}{k+2}$$

对于小样本 $k \ll N$。

3.3 应用

此公式可应用于基于样本估计整体的问题。例如，如果我们要估计一个集群的大小，可以使用上述公式来计算。

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

我们可以通过分析样本中观测到的最大编号 m 和样本编号的平均值 $\bar{m}$ 来估计整体的大小。

3.4 推导

样本最大值等于 m 的概率为 $\binom{m-1}{k-1}\binom{N}{k}$
样本最大值的期望值为

$$\mu = \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}} = \frac{k(N+1)}{k+1} \tag{1}$$

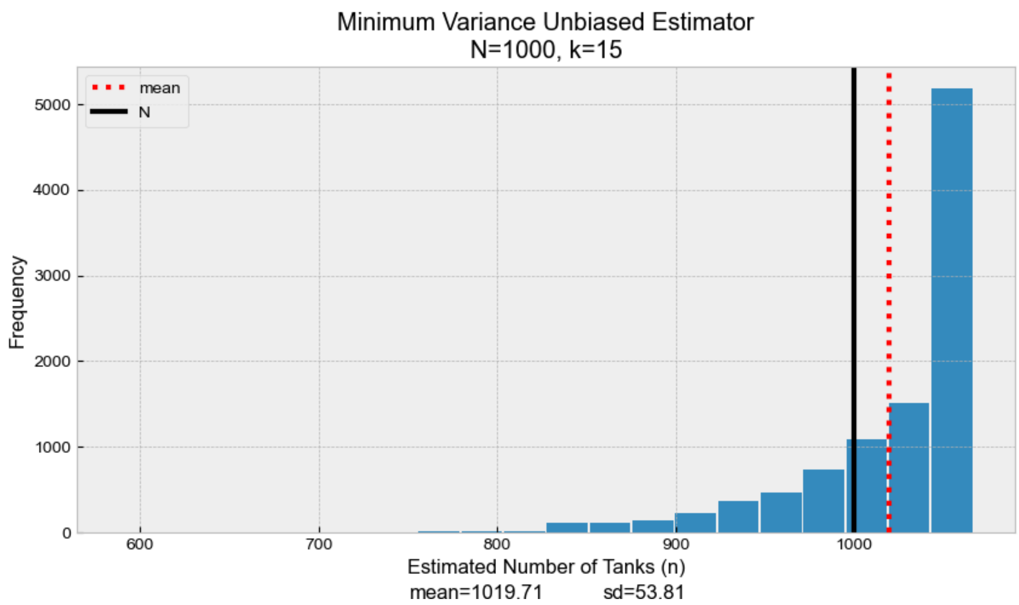
$$\Rightarrow N = \mu\left(1+k^{-1}\right) - 1 \tag{2}$$

因而

$$\mu\left(1+k^{-1}\right) - 1 = E\left[m\left(1+k^{-1}\right) - 1\right] \tag{3}$$

$$\Rightarrow \hat{N} = m\left(1+k^{-1}\right) - 1 \tag{4}$$

为 N 的无偏估计。
我们使用 python 来模拟了这一结果，以直方图的形式呈现:



代码附在了文章末尾。

3.5 总结

这种方法提供了一个对整体大小进行估计的有效手段，尤其是在样本较小的情况下。通过考虑到样本中的最大和平均值，我们能够得到一个对整体的可靠估计。

4 贝叶斯分析

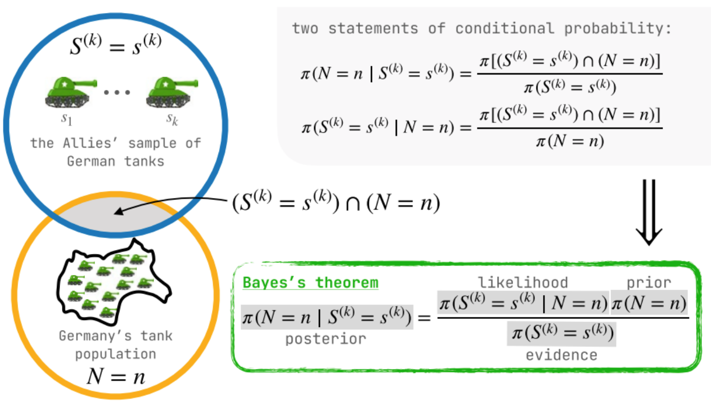
4.1 德国坦克问题的前期研究回顾

频率估计方法中，Kim Border 称德国坦克问题为一个“奇怪的案例”。坦克总数 n 的最大似然估计是观察到的最大序列号 $m(k)$ 中的最大值 $\max_{i\in\{1,\ldots,k\}} s_i$，这是一个有偏估计，因为显然 $m(k) \leq n$。

Leo Goodman 导出了坦克总数的最小方差无偏估计器。为了直观理解 $\hat{n}$，注意到 n 必须大于或等于 $m(k)$，如果在排序后的序列号 $s_1, \ldots, s_k$ 中观察到大（或小）间隔（包括在最小序列号之前的间隔），则 n 很可能远大于 $m(k)$。公式中 n 的估计量基于间隙大小来量化；$\frac{m(k)}{k} - 1$ 是间隙的平均大小。Goodman 还导出了 n 的频率学两侧 $1-\alpha$ 置信区间 $m(k) \leq n \leq x$，其中 x 是满足下面条件的最大整数：

$$\frac{(m(k)-1)^k}{(x)^k} \geq \alpha$$

符号 $(n)_k$ 表示阶乘。



4.2 贝叶斯方法概述

贝叶斯方法对德国坦克问题的处理赋予每个可能的坦克数量一个概率，从而量化不确定性，并提供一个机会将关于坦克数量的先验信息或信念纳入解决方案。我们将通过一个例子来说明。最后，我们回顾了在其他上下文中与德国坦克问题类似的问题。

上限 Ω 必须是有限的，因为该函数

$$f(n) = \lim_{\Omega\to\infty} \begin{cases} \frac{1}{\Omega-k} & \text{if } k \leq n < \Omega \\ 0 & \text{otherwise} \end{cases}$$

的结果为 $f(n)=0$，而这不是一个概率质量函数。

因而

$$(n \mid m,k)=\begin{cases}\frac{(m|n,k)}{\sum_{n=m}^{\Omega-1}(m|n,k)} & \text{if } m\le n<\Omega\\ 0 & \text{otherwise}\end{cases}$$

如果 $\sum_{n=m}^{\infty}(m \mid n,k)<\infty$，那么不受欢迎的变量 Ω 就能从表达式中消失。

$$(n \mid m,k)=\begin{cases}0 & \text{if } n<m\\ \frac{(m|n,k)}{\sum_{n=m}^{\infty}(m|n,k)} & \text{if } n\ge m\end{cases}$$

当 $k\ge 1$ 时，敌方坦克数量分布的众数为 m。

当 $k\ge 2$ 时，敌方坦克数量“等于”n 的可信度为

$$(N=n \mid M=m\ge k, K=k\ge 2)=\begin{cases}0 & \text{if } n<m\\ \frac{k-1}{k}\frac{\binom{m-1}{k-1}}{\binom{n}{k}} & \text{if } n\ge m\end{cases}$$

而敌方坦克数量 N “大于”n 的可信度为

$$(N>n \mid M=m\ge k, K=k\ge 2)=\begin{cases}1 & \text{if } n<m\\ \frac{\binom{m-1}{k-1}}{\binom{n}{k-1}} & \text{if } n\ge m\end{cases}$$

当 $k\ge 3$ 时，N 的均值有限：

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

当 $k\ge 4$ 时，N 的标准差有限：

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

4.3 一辆坦克的模型

从 n 辆坦克的总体中随机观察一辆坦克，当 $m\le n$ 时，其序列号为 m 的概率为 $\frac{1}{n}$，而当 $m>n$ 时概率为零。用艾弗森括号表示法可以写成：

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

这是 m 的条件概率质量分布函数。

当 m 为定值时，这是一个 n 的似然函数。

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

极大似然估计的坦克总数为 $N_0 = m$。

总概率为无穷大，因为尾部为一个调和级数数列。

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

但

$$\sum_n \mathcal{L}(n)[n < \Omega] = \sum_{n=m}^{\Omega-1} \frac{1}{n} \tag{5}$$

$$= H_{\Omega-1} - H_{m-1} \tag{6}$$

其中 H_n 为调和数。

可信度质量分布函数依赖于先前的限制 Ω:

$$(N = n|M = m, K = 1) \tag{7}$$

$$=(n|m) = \frac{[m \leq n]}{n} \frac{[n < \Omega]}{H_{\Omega-1} - H_{m-1}} \tag{8}$$

N 的均值为

$$\sum_n n \cdot (n|m) = \sum_{n=m}^{\Omega-1} \frac{1}{H_{\Omega-1} - H_{m-1}} \tag{9}$$

$$= \frac{\Omega - m}{H_{\Omega-1} - H_{m-1}} \tag{10}$$

$$\approx \frac{\Omega - m}{\log\left(\frac{\Omega-1}{m-1}\right)} \tag{11}$$

5 频率学派与贝叶斯学派

贝叶斯学派与频率学派是在统计推断问题上目前最为流行，应用最广的两大学派。频率学派认为概率是物质世界的一种客观属性, 并不因认识主体的不同而发生变化; 贝叶斯学派把概率看作对物质世界的一种主观认识, 是认识主体对物质世界信息量掌握多少的一种度量。贝叶斯统计方法是一种以贝叶斯公式为核心，以先验信息和后验信息为综合依据，以“辩证”推断为主要特征的统计方法。与经典的统计归纳推理方法相比，它采用了一种全新的思维范式，将不确定参数看作随机变量，并以贝叶斯理论为基础，将获取数据前人们的主观信念作为先验信息与样本信息进行综合，再根据贝叶斯定理推导出参数的后验概率分布，最后以该后验分布为基础，利用模拟方法进行参数的统计推断。

5.1 概率的解释

5.1.1 频率解释

在频率解释中，概率是根据事件在重复试验中出现的频率来近似计算的。如果事件 A 在 N 次试验中发生了 M 次，那么其概率可以近似表示为 $P(A) \approx \frac{M}{N}$。大数定律支持了这一近似理论，即当 N 趋于无穷大时，事件发生的频率几乎始终趋近于其概率。换句话说，当试验次数 N 足够大时，事件发生的频率会稳定在其概率附近，并且随着 N 的增大，这种稳定性会变得更加显著。这就是概率的频率解释。

然而，频率解释只适用于可以在相同条件下重复进行的试验。对于一些不可重复的事件，比如总统竞选或体育比赛结果，频率解释就无能为力了。因此，频率解释在处理一次性事件时存在局限性。

5.1.2 贝叶斯解释（主观解释）

贝叶斯学派认为概率是合理信度的度量，反映了个体的知识状态和主观信仰，因此称为主观概率。根据贝叶斯观点，概率是主观的，它取决于个体的信仰和经验。贝叶斯方法允许对一次性事件进行概率估计，因为它将概率解释为个体的信仰程度。在贝叶斯观点下，先验概率是在实验或观察之前或之外根据经验积累得出的。相较于频率解释，贝叶斯解释的优势在于允许对一次性事件进行概率估计，并且对决策的影响较小。

在贝叶斯分析中，概率值的微小差异并不会对决策产生显著影响，因为实际中往往会考虑多个事件的概率，并且这些概率之间存在一定的关系。此外，在数据分析中，当数据量足够大时，个体的主观概率对结果的影响通常较小。

5.2 两大学派在统计推断上的区别

两大学派在统计推断上的区别源于它们对概率理解的不同，这直接影响了参数的假设、样本的作用以及统计推断方法。频率学派将参数视为固定的常数，而贝叶斯学派则将参数视为随机变量。参数代表了某一系统特征的未知部分，例如在盒子中红球和白球的数量问题中，红球的比例就是一个未知参数。虽然实际的红球比例是确定的常数，但由于样本数据的随机性和测量误差，我们无法准确获知参数的真实值，因此贝叶斯学派将未知参数视为随机变量。

在频率学派中，对未知参数的推断是基于样本及其分布中的信息，既考虑实际抽到的样本又考虑理论上可能抽到但在本次抽样中未抽到的样本。而贝叶斯学派则通过综合先验分布的信息（即抽样前对参数的认识）和样本信息来对未知参数进行推断，只利用本次抽到的样本，不考虑未抽到的样本。尽管贝叶斯方法需要对先验分布做出假设，但它能够更充分地利用已有的信息，从而做出更准确的推断。

在区间估计方面，频率学派和贝叶斯学派也存在差异。例如，对于一个参数的估计结果是 $[0, 100]$，置信水平为 90%，频率学派会将这个区间理解为包含参数的概率为 90%，而贝叶斯学派则认为正确的解释是，如果多次抽样，每次都按照这种方法来估计区间，约有 90% 的区间会包含参数。这种差异主要源于频率学派考虑所有可能的样本，而贝叶斯学派仅考虑抽到的样本，使得贝叶斯学派对结果的解释更易于理解和接受。

在假设检验方面，两大学派也有不同。贝叶斯学派同等对待每个假设，其目的是找出给定问题的最佳答案，而频率学派则考虑各个假设的利弊后果。因此，在实施步骤上，频率学派需要确定原假设与备择假设，然后找出检验统计量的分布；而贝叶斯学派则简单方便得多，只需要根据后验概率分布来对参数属于各个假设的概率进行判断。

5.3 两大学派的过去与未来

贝叶斯学派的创始人贝叶斯早在 18 世纪就已提出贝叶斯公式, 但之后约 200 年, 始终是频率学派占据着统计学界的主导地位。当时最具影响力的统计学家大都属于频率学派, 如 Neyman、Fisher、Pearson 等。直到 20 世纪五六十年代, 贝叶斯学派开始逐步繁荣, 这要归因于该学派理论的不断完善特别是无信息先验分布的发展、统计决策理论对贝叶斯理论的使用以及贝叶斯理论在医学等领域中取得的显著成果, 时至今日, 贝叶斯学派已发展到足以和频率学派分庭抗礼的程度, 国际著名统计期刊刊登贝叶斯方法文章的比重日益增大。在当今统计学界, 众多统计学家也不再明确区分自己是属于哪个学派。当学者研究出一种频率学派意义下的新方法新模型后, 通常他本人或者其他学者也会研究出贝叶斯方法下的形式。虽然两大学派在具体方法理论的某些方面有殊途同归的结果, 但他们在哲学层面上的争议始终存在。统计作为一种分析数据的方法工具, 未来哪个学派更能占优, 这可能要取决于他们在实际应用中的表现, 尤其是在处理大数据问题中的表现。

6 代码

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rcParams

rcParams["font.family"] = "sans-serif"
rcParams["font.sans-serif"] = ["Arial"]

np.random.seed(1_294_837)

N = 1000
```

```python
k = 15
Ns = np.linspace(1, N + 1)
iterations = 10000

def sample(Ns, k):
    return np.random.choice(Ns, k, replace=False)

if __name__ == "__main__":
    samples = [sample(Ns, k) for _ in range(iterations)]
    estimator_value = [max(s) + max(s) / k - 1 for s in samples]
    mean = np.mean(estimator_value)

    plt.figure(figsize=(10, 5))
    plt.style.use("bmh")
    plt.hist(estimator_value, bins=20, rwidth=0.95)
    plt.axvline(mean, color="r", linewidth=3, linestyle=":", label="
       mean")
    plt.axvline(N, color="k", linewidth=3, linestyle="-", label="N")
    plt.legend()
    min_ylim, max_ylim = plt.ylim()
    plt.title(f"Minimum Variance Unbiased Estimator\\nN={N}, k={k}")
    plt.gcf().text(0.40, 0.0, "mean={:.2f}".format(np.mean(
       estimator_value)), fontsize=12)
    plt.gcf().text(0.56, 0.0, "sd={:.2f}".format(np.std(
       estimator_value)), fontsize=12)
    plt.xlabel("Estimated Number of Tanks (n)")
    plt.ylabel("Frequency")
    plt.savefig("estim.png", dpi=300)
    plt.show()
```

参考文献

[1] 范超. 概率是物质属性还是主观认识——频率学派与贝叶斯学派的区别 [J]. 中国统计,2016,(08):40-41.

[2] 裴治捷. 浅析统计学中贝叶斯估计方法和经典频率学派估计方法的不同 [J]. 科技视界,2014,(28):217. DOI:10.19694/j.cnki.issn2095-2457.2014.28.167.

[3] 赵玉震, 王成, 孙增国, 等. 贝叶斯学派与频率学派在统计推断上的差异 [J]. 电子设计工程,2013,21(13):21-24. DOI:10.14022/j.cnki.dzsjgc.2013.13.001.

[4] Simon, C.M. A Bayesian Treatment of the German Tank Problem. Math Intelligencer (2023). https://doi.org/10.1007/s00283-023-10274-6