UNIVERSITÀ
DI SIENA
1240

Serial number analysis

STATISTICAL DATA ANALYSIS
Ph. D. IN EXPERIMENTAL PHYSICS - XXXVII CYCLE

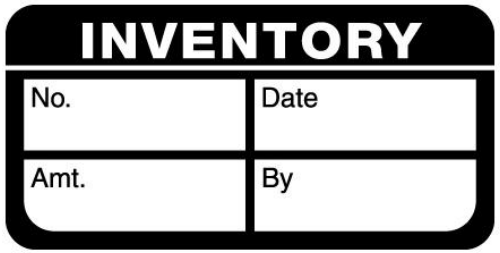
- DIPARTIMENTO SCIENZE FISICHE, DELLA TERRA E DELL'AMBIENTE

FRANCESCO CAMINATI
SIENA 11/05/2023

UNIVERSITÀ DI SIENA
1240

SUMMARY

- **Introduction**
- **Serial Number Analysis**
  - ➤ Gap estimation
  - ➤ Minimum variance unbiased estimation (MVUE)
- **1st example: equipment purchased by a company**
- **2nd example: production estimation during WW2 («German Tank Problem»)**



UNIVERSITÀ DI SIENA
1240

INTRODUCTION

Serial number analysis is the estimation of the total dimension of a population based on the observation of a limited amount of items, as long as these items have sequential serial numbers. We can use the numbers we sampled to predict what is the highest serial number in the pool. This can be used to:

- determine the total number of participants in a marathon
- estimate the production capacity of a factory
- find out how many pieces of equipment a company bought from their inventory number (if no records are easily available)

UNIVERSITÀ DI SIENA 1240

INTRODUCTION – THE PROBLEM

Serial numbers of a particular item are uniformly distributed integers between the starting number a+1 and the final number a+p. Within the whole pool, we observe only n numbers randomly selected without replacement. Our aim is to estimate p from our sample pool.

For simplicity, let's assume that a=0. This means that:

- **Total population:** B = [1,2,3,…n, n+1,…p-1,p];
- **Our sample:** S = [s_1, s_2, s_3,… s_{n-1}, s_n]

UNIVERSITÀ DI SIENA 1240

MEAN/MEDIAN ESTIMATION

The total population must have a mean value, which will coincide with its median as well. This value (M) will be preceded and followed by m-1 other values within B:

$$p=(M-1)+1+(M-1)=2M-1$$

We could take the mean (or median) m of the sample pool S and roughly estimate p as

$$e(p)=2m\text{-}1$$

We know that $p \geq g$, the largest number in our sample S, but it is not guaranteed that $e(p) \geq g$.

UNIVERSITÀ DI SIENA 1240

GAP ESTIMATION

To ensure that $e(p) \geq g$, we could estimate p by adding a certain gap to g.

The gap between g and p could reasonably be equal to the average gap between any other number in our sample pool, including also the gap between 1 and s_1:

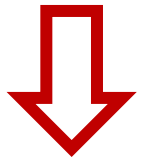
$$p - g = [(s_1 - 1) + (s_2 - s_1 - 1) + \cdots + (s_n - s_{n-1} - 1)]/n$$

$$p - g = \frac{s_n}{n} - 1 = \frac{g}{n} - 1 \rightarrow p = \frac{n+1}{n} g - 1$$

MINIMUM VARIANCE UNBIASED ESTIMATOR

The minimum variance unbiased estimator (MVUE) is the estimator that minimizes the variance. Before we start, let us define:

- Ways to select n integers from the first p integers: $Q = \binom{p}{n}$
- Ways to select n-1 integers from the first g-1: $R = \binom{g-1}{n-1}$



- The probability that g will be the largest integer in a sample of size n is

$$P\{g|n,p\} = \frac{R}{Q} = \binom{g-1}{n-1}\frac{n!\,(p-n)!}{p!} = \frac{(g-1)!\,n(p-n)!}{(g-n)!\,p!}$$

UNIVERSITÀ DI SIENA
1240

MINIMUM VARIANCE UNBIASED ESTIMATOR (2)

Now we will demonstrate that there is only one function e(g) that is an unbiased sufficient estimate of p. Given a sample S of size n, the possible values for our parameter are $p = n, n+1, n+2$ ..., just like the possible values for $g = n, n+1, n+2$ ...

If $p = n$, then $g = n$ with P=1, and for the statistic e(g) (which is based only on g) t be an unbiased estimator of p, $e(g) = n$ when $g = n$. So

$$E\{e(g)|p = n\} = e(n) = n$$

If $p = n+1$, the only possible results are $g = n$ and $g = n+1$, and

$$E\{e(g)|p = n+1\} = e(n)P\{g = n|p = n+1\} +$$

$$+e(n+1)P\{g = n+1|p = n+1\} = n+1$$

MINIMUM VARIANCE UNBIASED ESTIMATOR (3)

In general, if e(g) is unbiased, for $h = n, n+1, n+2$ ...

$$e(h) = \left[h - \sum_{i=n}^{n-1} e(i) P\{g = i | p = h\} \right] \Big/ P\{g = h | p = h\}$$

By recursion, we can determine e(h) uniquely, so e(g) is the only unbiased estimate of p that is based on g only (see ref. 1):

$$E\{g|n,p\} = \sum_{g=n}^{p} g \frac{(g-1)!\, n(p-n)!}{(g-n)!\, p!} = \sum_{g=n}^{p} \frac{g!\, n(p-n)!}{(g-n)!\, p!} =$$

$$= \frac{(p+1)!\, n(p-n)!}{(p-n)!\, (n+1)p!} = \frac{(p+1)n}{(n+1)}$$

UNIVERSITÀ DI SIENA
1240

MINIMUM VARIANCE UNBIASED ESTIMATOR (4)

So, the expected value of

$$e = \frac{g(n+1)}{n} - 1 = p$$

Since this is the only unbiased estimate e(g) based on g, we see that $e = e(g)$. As you can see, it coincides with the «gap estimate» we previously did.

By further calculations, we can obtain its variance:

$$\sigma^2(e) = \frac{(p+1)(p-n)}{(n+2)n}$$

UNIVERSITÀ DI SIENA
1240

MINIMUM VARIANCE UNBIASED ESTIMATOR (5)

If the initial number of our population is a+1, instead of 1, the problem can be reduced to the analysis we just did by subtracting a to all the numbers.

If a is unknown too, then the difference d between g and the smallest number in our sample can be estimated as

$$E\{d\} = \frac{(p+1)(n-1)}{(n+1)}$$

The minimum variance unbiased estimator for p is then

$$f = \frac{d(n+1)}{(n-1)} - 1$$

UNIVERSITÀ DI SIENA 1240

EXAMPLE 1: EQUIPMENT PURCHASED

Let’s say that a company or institution bought a certain amount of equipment many years ago. If no records are available, the amount of total equipment bought can be calculated from the serial numbers of a sample of such equipment.

In ref. 2, as an example, Goodman calculates the amount of office furniture (desks, bookcases,…) that the Division of Social Sciences of the University of Chicago bought over twenty years before.

UNIVERSITÀ DI SIENA
1240

EXAMPLE 1: EQUIPMENT PURCHASED (2)

From 31 pieces of such equipment, the following serial numbers were recorded:

S = [83,135,274,380,668,895,955,964,1113,1174,1210,1344,1387,1414,1610, 1668,1689,1756,1865,1874,1880,1936,2005,2006,2065,2157,2220,2224,2396, 2543, 2787]

The unbiased estimate of the total number is

$$f = \frac{\mathrm{d}(\mathrm{n}+1)}{(\mathrm{n}-1)} - 1 = (2787 - 83)\frac{32}{30} = 2884{,}3$$

After days of accurate research, a secretary from the administrative office found an official record that set p=2885.

EXAMPLE 2: GERMAN TANK PROBLEM

One of the most famous examples of serial number analysis occured during World War 2, the so-called «German tank problem». The Allied intelligence used serial number analysis to evaluate the monthly production of several types of German weapons (tanks, airplanes, trucks, half-tracks, rockets,…) by analyzing the serial numbers on captured assets.
Even though the results of the statistical analysis were very different from the estimates of the intelligence, time proved that the former was right, as reported in ref. 3.

EXAMPLE 2: GERMAN TANK PROBLEM INTRODUCTION

Above: Logo of the Office of Strategic Services, the American intelligence agency during WW2

During the early stages of the conflict, Allied economic intelligence proved to be inadequate. Conventional intelligence estimates were based on:

- pre-war data extrapolated from American and British experience;
- standard tables of equipment requirements derived from German order of battle estimates.

Models based on this data were unreliable, since they allowed for many degrees of freeedom which could be restricted only using unrealistic and/or very rigid assumptions.
Data on specifics plants was no better: interrogations and secret sources conveyed a large amount of conflicting data.

UNIVERSITÀ DI SIENA 1240

EXAMPLE 2: GERMAN TANK PROBLEM STATISTICAL APPROACH

By 1943, the Allied had collected a quite large sample of German equipment, and almost every single piece was labelled with:

- name and location of the manufacturer;
- date of manufacture;
- production serial number;
- miscellaneous markings (trade marks, mold numbers, cast numbers…)

These markings had two purposes:

- quality control
- spare parts control

A collaboration between the British Ministry of Economic Warfare and the Economic Warfare Division of the American Embassy lead to a technique that allowed to extrapolate strategic information from the analysis of these serial markings.

UNIVERSITÀ DI SIENA
1240

EXAMPLE 2: GERMAN TANK PROBLEM RESULTS

Type of Tire	Estimated Average Monthly Production, Jan.–Mar. 1943	Actual Average Monthly Production, 1943[1]	Percentage Error
Truck and Passenger car	147,000	159,700	8% −
Aero	28,500	26,400	8% +
Total	175,500	186,100	6% −

Type of Truck	Estimated Production for 1942	Speer Ministry Statistics	Percentage Error
Light truck	16,500	14,436	15% +
Medium truck	62,300	53,439	17% +
Heavy truck	18,500	11,952	35% +
Total	97,300	79,827	22% +

Above: Comparison between production estimates obtained by serial number analysis and data from the German War and Armaments Ministry for tyres and trucks production (ref. 3).

Below: comparison between serial number estimate, intelligence estimate (American and British combined), and actual monthly tank production for the 1940-42 period (ref. 3).

Date	Estimated Monthly Production		Monthly Production Speer Ministry
	Serial Number Estimate	Munitions Record 10 Aug. 42	
June, 1940	169	1000	122
June, 1941	244	1550	271
August, 1942	327	1550	342

UNIVERSITÀ DI SIENA 1240

CONCLUSION

Serial number analysis is a very powerful tool for population estimation based on a limited sample pool. From this data, one can easily estimate the total size of the population, its variance, and even determine confidence intervals.

In particular, it was a very efficient method for economic intelligence evaluations during World War 2, as demonstrated by the data obtained from the German War Ministry after the war.

UNIVERSITÀ DI SIENA 1240

Thank you very much for your attention

UNIVERSITÀ DI SIENA 1240

REFERENCES

1. Goodman, L., “Serial Number Analysis”, *Journal of the American Statistical Associacion,* **49** (265), Dec. 1952, pp. 622-633 (doi:10.2307/2280780)
2. Goodman, L., “Some practical techniques in serial number analysis”, *Journal of the American Statistical Associacion,* **47** (260), Mar. 1954, pp. 97-112 (doi:10.2307/2281038)
3. Ruggles, R., Brodie, H., "An Empirical Approach to Economic Intelligence in World War II”, *Journal of the American Statistical Associacion,* **42** (237), 1947, pp. 72-91 (doi:10.1080/01621459.1947.10501915)