Group 5: Neha Karna, Prafulla Shrestha, Aidan Stewart, Josh Lefdal, Shivam Bhardwaj

A Monte Carlo simulation to establish an upper confidence bound for random match probability in biometric identification systems.

Biometric identification systems (BIS) offer a more secure alternative to traditional methods like cards, keys, or PINs, which are vulnerable to theft or loss (de Luis-García et al., 2003). BIS uses physiological traits—such as fingerprints, facial recognition, or iris scans—that are difficult to replicate (Wayman, 2000). The system digitally captures and compares these traits to stored templates for authentication or identification, with the probability of error increasing as the number of comparisons grows.

A crucial aspect of a successful biometric identification system (BIS) is its **biometric capacity**, which is often assessed by the **random match probability (RMP)**. RMP measures the likelihood that BIS will fail to distinguish between two different profiles, mistakenly identifying them as the same person. The **lower the RMP**, the more efficient the system is. For example, in forensic DNA analysis, RMP refers to the probability that a randomly selected person from a population other than the suspect could have the same DNA profile. A smaller RMP indicates a higher likelihood that the DNA sample belongs to the suspect, rather than being a coincidental match (National Research Council, 1996). In BIS, we aim to demonstrate that RMP is below a certain threshold, focusing on the **1-α upper confidence bound** to ensure system accuracy.

Confidence interval means a range of values we expect our estimate (RMP) to fall between if we repeatedly sample our data. The upper and lower values in the confidence interval are the **upper and lower confidence bounds**, respectively. **Confidence level** is the percentage of times we expect to reproduce an estimate between the upper and lower confidence bound (Bevans, 2023). If we maintain confidence level of 95 % while designing our confidence bounds, it means that we are confident 95 out of 100 times that our estimate will fall between the upper and lower confidence bound.

Coverage probability refers to the proportion of times that a confidence interval contains the true parameter value. In other words, if we construct many upper confidence bounds, the coverage probability is the fraction of those intervals that actually contain the true value of the RMP that we are estimating.

Monte Carlo simulation is a mathematical model that uses repeated random sampling to estimate outcomes of mathematical functions. Also known as multiple probability simulation, the Monte Carlo method can be considered a series of coin flips, which uses randomness to give crucial insight into what may happen tomorrow (weather forecast or economic recession). Although no single universal Monte Carlo method exists, many simulations are based on a series of probability density functions (PDFs) (Harrison, 2010).

Multinomial distribution is the generalization of binomial distribution when we extend the number of categories beyond two, in our case almost 50 categories. However, we risk oversimplifying the real-world case scenarios, which might involve more dynamic distribution. While using a multinomial distribution for Monte Carlo simulation is efficient, the complexities increase as we increase the sample size and simulation size using the multinomial distribution, which results in a longer computational time. The major strength of using multinoulli may be flexibility in defining categories or discrete data such as fingerprints or facial recognition.

Methodology

To set up the Monte Carlo model, we define inputs as randomly selected profiles (input: samp.size) and estimate RMP and upper bounds.

Let X_i and $X_j$be the i^{th} and j^{th} randomly selected biometric profile from a population of 3000 profile. If X_i *and* X_j match then, $m_{ij} = 1$, else $m_{ij} = 0$. The RMP, for $i \neq j$ can be denoted by $E(m_{ij}) = \theta$. We are assuming multinomial distribution of our population where there are more than two categories.

Then the estimated random match probability, from a sample of n randomly selected profiles, denoted by $\hat{\theta}_n$, given by,

$$\hat{\theta}_n = \binom{n}{2}^{-1} \sum_{1\leq i} \sum_{<j\leq n} m_{ij}$$

We have three statistical approaches for the upper confidence bounds:

UB1: i.i.d. Binomial Assumption

$$\boldsymbol{UB_1} = \hat{\theta}_n + 1.645\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{\binom{n}{2}}}$$

UB2: Independent Comparisons Binomial Assumption

$$UB_2 = \hat{\theta}_n + 1.645\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{\frac{n}{2}}}$$

UB3: A U-Statistic Approach

$$UB_3 = \hat{\theta}_n + 1.645\frac{2\hat{\rho}}{\sqrt{n}}$$

Where,

$$\hat{\rho}^2 = \frac{1}{n-1}\sum\nolimits_{i=1}^{n}\left(\hat{\theta}_i - \hat{\theta}_n\right)^2$$

Here,
$\hat{\theta}_i$is the proportion of profiles in the sample that is a match with i^{th} profile.

So, the estimated RMP from n sample is given by, $\hat{\theta}_n = n^{-1}\sum_{i=1}^{n}\hat{\theta}_i$.

Next, we repeatedly sample from the PDFs (i.e., number of simulations or sim.size in our R code) to generate large numbers of outputs, take the mean, and use the results, including **coverage probability,** to recommend one of the three upper confidence bounds (UB1, UB2, or UB3).

- We have used the **means of pairwise distance** to estimate RMP. Pairwise distance is how close or different two values are. As we have discussed, RMP is the probability that a randomly selected biometric profile other than the suspected one will be a match, it makes sense to calculate the mean of pairwise distance.
- We have chosen a range of **sample sizes from 30-500**. In a real scenario, we may have to deal with a sample size as low as 30 (a small number of possible suspects at a crime scene) or as high as 500 or even 1000 (when there is no clue and a criminal can be at large). No matter what the sample size is, we want our BIS system to be efficient in terms of the biometric capacity as measured by RMP.
- We have used **simulation sizes of 1500 and 3000**. Higher simulation size means, we are repeatedly sampling random data to get large output. The idea with the large output is that we wanted to be as representative as possible of all the combinations that may be encountered in the real scenario. In short, the bigger the simulation size, the more accurate our Monte Carlo is going to be.
- We have assumed **equal and unequal RMPs** among our categories **(3 categories: 40, 50, 60)**. This may be important when we want to define the number of categories and in situations where probabilities of biometric profiles falling under one category may not be the same (the real scenario may be more complex than a simple coin flip or a rolling of a six-sided fair die). In **equal RMP simulation**, each category has an equal chance of being selected while in **unequal RMP simulation**, some categories may be chosen more often than others.

Algorithm

- The simulation was iterated over the combination of categories (40,50 and 60), sample size (30, 60, 120, 240 and 500), and simulation size (1500 and 3000).
- Samples were drawn from a multinomial distribution assuming both equal probabilities and unequal probabilities separately.
- Each random sample falls into at least one of the categories.
- For each pair of samples, the distance is based on the difference between their categorical values.
- If two samples are in the same category, their distance will be zero because they both will have 1 in the same column of the category.
- The logical operator == creates a logical vector where every random match is identified as success (1) and failure (0).
- The estimated RMP is the proportion of successes out of the total pair of observations over the simulation. We are interested in having a lower RMP that defines the biometric capacity of BIS.
    - For UB3 estimated RMP is calculated as the mean of sum of proportion the proportion of match for each sample profile.
- The upper bound is calculated using UB1, UB2 and UB3.

- Coverage probability is the proportion of times that the true value of RMP lies below the calculated UB. Hence, a higher coverage probability would represent an efficient upper confidence bound.

Results

Equal Random match probabilities

UB1: The simulation gave coverage probability ranging from 0.8853 to 0.9747 (simulation size:1500) and 0.8747 to 0.9633 (simulation size: 3000) across different sample sizes at multiple categories. The minimum coverage probability for both simulations was observed when the sample size was 30 for 50 categories. For all other combinations, coverage probability was higher than 0.90 which is good as we are aiming for a 95% confidence interval.

UB2: The simulation was consistent in terms of coverage probability of 100 % across all sample sizes at multiple categories for both simulation sizes.

UB3: The simulation gave coverage probability from 0.9427 to .9873 across different sample sizes at multiple categories (simulation size:1500) and 0.9483 to 0.9853 (simulation size: 3000). The minimum coverage probability was observed when the sample size was 30 for 40 categories (simulation size:1500) and 30 for 50 categories (simulation size:3000).

Un-equal Random match probabilities

UB1: The simulation gave maximum coverage probability of 0.914 at sample size 60 for 50 categories (simulation size: 1500) and 0.920 at sample size 60 for 40 categories (simulation size: 3000). The coverage probability dropped as sample size increased across all categories for both simulation sizes with minimum coverage probability observed at maximum sample size of 500 (0.7567 for simulation size:1500 and 0.7490 for simulation size:3000).

UB2: The result for UB2 un-equal RMP was consistent with what we observed for UB2 equal RMP, coverage probability of 100%.

UB3: The coverage probability ranged from 0.9045 to 0.9620 (simulation size:1500) and 0.9263 to 0.9650 (simulation size:3000). For all combinations of sample size and categories, coverage probability was higher than 0.90 which is good as we are aiming for 95 % confidence interval.

Conclusions

- UB1 gives the lowest upper bound and has the closest coverage probability to 0.95 in reference to equal probability; hence, it is our recommendation, for equal random match probability to utilize UB1 for the approximate upper confidence bound while using equal probability.
- The higher upper bound calculated by UB2 must be one of the reasons for near the 100% coverage probability. However, since we are interested in lower RMP, or alternatively, lower upper bound, we can conclude that UB2 is inefficient.
- UB3 was similar to UB1 in coverage probability regarding equal probability, however, UB1 was closer to 0.95. Additionally, while utilizing unequal probability, UB3 was closest to 0.95 and would be our recommendation for the approximate upper confidence bound.

Group 5: Neha Karna, Prafulla Shrestha, Aidan Stewart, Josh Lefdal, Shivam Bhardwaj

Recommendation

- **Equal RMP: UB1**
- **Un-equal RMP: UB3**

Group 5: Neha Karna, Prafulla Shrestha, Aidan Stewart, Josh Lefdal, Shivam Bhardwaj

References

Baumer, Ben & Kaplan, Daniel & Horton, Nicholas. (2021). Modern Data Science with R. 10.1201/9780429200717.

de Luis-García, R., Alberola-López, C., Aghzout, O., & Ruiz-Alzola, J. (2003). Biometric identification systems. Signal Processing, 83(12), 2539-2557. https://doi.org/https://doi.org/10.1016/j.sigpro.2003.08.001

Wayman, J. (2000). National Biometric Test Center Collected Works.

National Research Council, Division on Earth and Life Studies, Commission on Life Sciences, & Committee on DNA Forensic Science: An Update. (1996). The Evaluation of Forensic DNA Evidence. National Academies Press.

Bevans, R. (2023). Understanding Confidence Intervals | Easy Examples & Formulas. Scribbr. Retrieved October 21, 2024, from https://www.scribbr.com/statistics/confidence-interval/

Harrison, R. L. (2010). Introduction To Monte Carlo Simulation. AIP Conf Proc, 1204, 17-21. https://doi.org/10.1063/1.3295638