A Computational Framework for Leveraging Population-level Data to Explore Humanitarian and Justice-oriented Perspectives

Nathan Alexander, PhD [[1]](#footnote-20)

Hye Ryeon Jang, PhD [[2]](#footnote-22)

# Abstract

While data-driven decision-making has become a gold standard in most industry settings, there is a differential availability of large-scale datasets addressing various humanitarian issues and injustices across the globe. By integrating database development and conceptual reviews, computational practices can be situated within a framework that enables a nuanced analysis of demographic trends, disparities, and their implications on various societal dimensions based on theoretical or conceptual constructs. This study focuses on two cohorts of undergraduate students who engaged in computational practices, analyzing their ability to interpret and model population-level data. Through this experiential learning approach, students developed a deeper understanding of the factors influencing educational attainment, healthcare access, and income inequality. The study’s findings demonstrate that a globally-situated framework and related data not only improve statistical literacy and analytical skills but also foster increased attention to the differential goals of various theoretical constructs when engaging in quantification and computational modeling practices.

*Keywords*: census, data, literacy, population, statistics

## Introduction

## The PRISM Method: Proof, Research, Ideation, Synthesis, and Modeling

This framework integrates scientific inquiry, mathematical proof, modeling, and creative ideation into a unified approach. The PRISM method combines mathematical reasoning with empirical scientific practices and creative problem-solving. It provides a structured yet flexible approach for tackling complex problems across pure and applied domains.

Proof: This initial step of the PRISM method focuses on establishing a solid foundation through rigorous mathematical and logical reasoning. It involves formulating precise definitions and axioms, developing formal mathematical proofs, and applying deductive reasoning to validate logical arguments. Researchers identify necessary and sufficient conditions for their hypotheses and work to establish foundational truths. This stage ensures that subsequent steps are built upon a robust theoretical framework, reducing the likelihood of errors or inconsistencies later in the process.

Research: The research phase is dedicated to gathering and analyzing existing information and data relevant to the problem at hand. This includes conducting comprehensive literature reviews, collecting data from various sources, and analyzing historical trends and patterns. Researchers work to identify gaps in current knowledge and formulate specific research questions and hypotheses. This stage provides crucial context and background, allowing the team to build upon existing work and avoid duplicating efforts. It also helps in refining the problem statement and identifying potential avenues for investigation.

Ideation: During the ideation stage, the focus shifts to creative thinking and the generation of novel approaches. This phase encourages brainstorming potential solutions, exploring unconventional ideas, and applying lateral thinking techniques. Researchers combine concepts from different domains and challenge existing assumptions and paradigms. The goal is to expand the solution space and foster innovative thinking, potentially leading to breakthrough insights. This stage is crucial for addressing complex problems that may require non-traditional approaches.

Synthesis: The synthesis step involves integrating insights from the previous stages to form coherent theories or solutions. This process includes combining ideas from different sources, identifying patterns and connections, and developing unified theories or frameworks. Researchers work to reconcile conflicting information and create a holistic understanding of the problem. This stage is critical for making sense of diverse and sometimes disparate information, leading to a more comprehensive and nuanced approach to the problem.

Modeling: The final stage of the PRISM method focuses on creating tangible representations of the problem or solution. This includes developing mathematical or computational models, creating simulations or prototypes, and designing experiments to test hypotheses. Researchers also work on visualizing data and concepts to better communicate their findings. The modeling stage allows for testing and refinement of ideas in a controlled environment, providing valuable feedback that can inform further iterations of the process. It bridges the gap between theoretical concepts and practical applications, enabling researchers to validate their findings and make predictions about real-world scenarios.

### Algorithm

1. Problem Identification
   * Define the problem or research question
   * Conduct initial background research
2. Rigorous Formulation
   * Develop formal mathematical definitions and axioms
   * State hypotheses and conjectures precisely
3. Ideation and Exploration
   * Brainstorm creative approaches and potential solutions
   * Explore informal examples and special cases
4. Synthesis and Proof
   * Construct formal mathematical proofs where applicable
   * Develop logical arguments and deductive reasoning chains
5. Modeling and Simulation
   * Create mathematical or computational models
   * Implement simulations and generate synthetic data
6. Empirical Testing
   * Design and conduct experiments or observational studies
   * Collect and analyze real-world data
7. Interpretation and Refinement
   * Evaluate results and refine hypotheses/models
   * Identify limitations and areas for further investigation
8. Communication
   * Document process and findings
   * Present results to relevant audiences

### Example: Racial Disparities in Traffic Stops

We apply the PRISM method to investigate racial disparities in traffic stops based on the work of Khajavi (2019).

1. Problem Identification
   * Research question: Are there racial disparities in traffic stop rates?
   * Background: Review existing literature on racial profiling in policing
2. Rigorous Formulation
   * Define terms: traffic stop, racial disparity, benchmark population
   * Hypothesis: Traffic stop rates for minority drivers are disproportionately higher than for white drivers, controlling for relevant factors
3. Ideation and Exploration
   * Brainstorm potential contributing factors: implicit bias, over-policing in certain neighborhoods, socioeconomic factors
   * Explore simplified scenarios and thought experiments
4. Synthesis and Proof
   * Develop statistical framework for measuring disparities
   * Prove mathematical properties of chosen disparity metrics
5. Modeling and Simulation
   * Create agent-based model of traffic stops
   * Simulate various scenarios with synthetic data
6. Empirical Testing
   * Analyze real traffic stop data from police departments
   * Conduct statistical tests to assess disparities
7. Interpretation and Refinement
   * Evaluate results and refine initial hypotheses
   * Identify limitations and propose further studies
8. Communication
   * Prepare research paper and data visualizations
   * Present findings to policymakers and community stakeholders

## Mathematical models

To examine the foundations of the two examples provided (traffic stops and fatal police shootings), we should employ several mathematical models and statistical techniques. Here’s an outline of the key models and approaches:

## 1. Poisson and Negative Binomial Regression

These models are appropriate for analyzing count data, such as the number of traffic stops or fatal shootings.

* **Poisson Regression**: Used when the variance equals the mean.
* **Negative Binomial Regression**: Preferred when there’s overdispersion (variance > mean).

The model takes the form:

Where is the count outcome, and are predictor variables.

## 2. Logistic Regression

This model is useful for binary outcomes, such as whether a stop resulted in a search or not.

Where is the probability of the event occurring.

## 3. Multilevel Models

These models account for hierarchical data structures, such as stops nested within officers or precincts.

Where indexes individuals and indexes groups.

## 4. Propensity Score Matching

This technique helps control for confounding variables when estimating causal effects.

Where is the propensity score, is the treatment indicator, and is a vector of covariates.

## 5. Difference-in-Differences

This approach can be used to analyze the impact of policy changes on racial disparities over time.

Where is the treatment group indicator, is the post-treatment time period indicator, and is their interaction.

## 6. Veil of Darkness Test

This test, mentioned in the examples, uses logistic regression to compare the racial composition of stops during daylight versus darkness.

Where is the probability of a minority driver being stopped, and is an indicator for stops made after dark.

## 7. Synthetic Control Method

This method could be used to construct a synthetic control group for analyzing the impact of policy changes on racial disparities in policing.

Where is the outcome for the treated unit, are observed covariates, are unobserved factors, and are factor loadings.

These models provide a robust framework for analyzing racial disparities in policing, allowing researchers to control for various confounding factors and test different hypotheses about the nature and causes of these disparities[1][2][3][4].

Citations: [1] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11181091/> [2] <https://assets.ipums.org/_files/mpc/wp2013-05.pdf> [3] <https://academic.oup.com/pnasnexus/article/1/4/pgac144/6652221> [4] <https://www.ppic.org/publication/racial-disparities-in-traffic-stops/> [5] <https://learn.arcgis.com/en/projects/examine-racial-disparities-in-police-stops/> [6] <https://www.sciencedirect.com/science/article/abs/pii/S0927537120301160> [7] <https://phys.org/news/2024-07-connecticut-racial-disparities-traffic-states.html> [8] <https://en.wikipedia.org/wiki/Scientific_method>

### Race and Traffic Stop Data

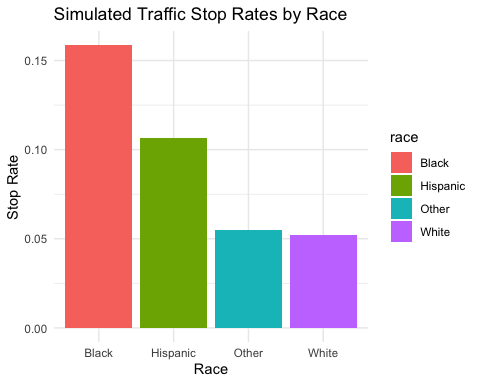
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
  
set.seed(123)  
  
# Generate synthetic population  
population <- data.frame(  
 race = sample(c("White", "Black", "Hispanic", "Other"), 10000, replace = TRUE,   
 prob = c(0.6, 0.2, 0.15, 0.05)),  
 age = rnorm(10000, mean = 40, sd = 15),  
 speed = rnorm(10000, mean = 65, sd = 10)  
)  
  
# Simulate biased stop probabilities  
population <- population %>%  
 mutate(  
 stop\_prob = case\_when(  
 race == "White" ~ 0.05,  
 race == "Black" ~ 0.15,  
 race == "Hispanic" ~ 0.10,  
 race == "Other" ~ 0.07  
 ),  
 stopped = rbinom(n(), 1, stop\_prob)  
 )  
  
# Analyze stop rates  
stop\_rates <- population %>%  
 group\_by(race) %>%  
 summarize(  
 stop\_rate = mean(stopped),  
 n = n()  
 )  
  
# Visualize results  
ggplot(stop\_rates, aes(x = race, y = stop\_rate, fill = race)) +  
 geom\_col() +  
 labs(title = "Simulated Traffic Stop Rates by Race",  
 y = "Stop Rate",  
 x = "Race") +  
 theme\_minimal()



This code generates a synthetic population with racial demographics, simulates traffic stops with built-in racial disparities, and produces a visualization of the results. Researchers could use this as a starting point to explore various scenarios and test analytical methods before applying them to real-world data[1][2].

Citations: [1] <http://www.contrib.andrew.cmu.edu/~avigad/Papers/method.pdf> [2] <https://dept.math.lsa.umich.edu/~jchw/PrimerOnProof.pdf> [3] <https://math.osu.edu/undergrad/non-majors/resources/proof-method> [4] <https://www.reddit.com/r/math/comments/25f70a/does_the_study_of_pure_mathematics_use_the/> [5] <https://en.wikipedia.org/wiki/Scientific_method> [6] <https://www.sciencebuddies.org/science-fair-projects/science-fair/steps-of-the-scientific-method> [7] <https://math.libretexts.org/Courses/SUNY_Schenectady_County_Community_College/Discrete_Structures/03%3A_Constructing_and_Writing_Proofs_in_Mathematics/3.06%3A_Review_of_Proof_Methods>

Certainly. Let’s apply the PRISM method to analyze fatal police shootings using the Washington Post’s Fatal Force database. This example will focus on racial disparities in fatal police shootings.

## Fatal Police Shootings Analysis

1. Problem Identification
   * Research question: Are there racial disparities in fatal police shootings?
   * Background: Review the Washington Post’s Fatal Force database methodology[1][3]
2. Rigorous Formulation
   * Define terms: fatal police shooting, racial disparity, population benchmark
   * Hypothesis: Black individuals are disproportionately represented in fatal police shootings compared to their population share
3. Ideation and Exploration
   * Consider factors: geographic distribution, armed vs. unarmed victims, mental health crises
   * Explore trends over time and potential policy impacts
4. Synthesis and Proof
   * Develop statistical framework for measuring disparities
   * Calculate per capita rates and risk ratios
5. Modeling and Analysis
   * Use R to analyze the Washington Post’s Fatal Force database
   * Perform statistical tests and create visualizations
6. Empirical Testing
   * Compare findings to other studies and government data
   * Assess limitations of the database and potential biases
7. Interpretation and Refinement
   * Evaluate results and refine initial hypotheses
   * Identify areas for further investigation, such as local-level analyses
8. Communication
   * Prepare data visualizations and summary statistics
   * Draft report for policymakers and community stakeholders

### R Code for Analysis

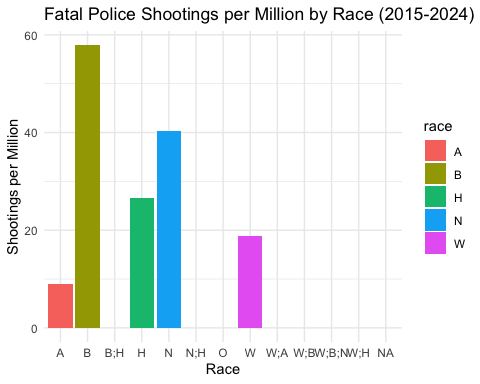
Here’s R code to analyze the Washington Post’s Fatal Force database:

library(dplyr)  
library(ggplot2)  
library(readr)  
  
# Load the data  
url <- "https://raw.githubusercontent.com/washingtonpost/data-police-shootings/master/v2/fatal-police-shootings-data.csv"  
shootings <- read\_csv(url)

## Rows: 10033 Columns: 19  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (12): threat\_type, flee\_status, armed\_with, city, county, state, locati...  
## dbl (4): id, latitude, longitude, age  
## lgl (2): was\_mental\_illness\_related, body\_camera  
## date (1): date  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Calculate shootings by race  
shootings\_by\_race <- shootings %>%  
 group\_by(race) %>%  
 summarize(count = n()) %>%  
 mutate(percentage = count / sum(count) \* 100)  
  
# Calculate per capita rates (using approximate 2020 US population data)  
population <- data.frame(  
 race = c("W", "B", "H", "A", "N"),  
 population = c(236992000, 41100000, 62100000, 19860000, 3400000)  
)  
  
shootings\_per\_capita <- shootings\_by\_race %>%  
 left\_join(population, by = "race") %>%  
 mutate(per\_million = (count / population) \* 1000000)  
  
# Visualize results  
ggplot(shootings\_per\_capita, aes(x = race, y = per\_million, fill = race)) +  
 geom\_col() +  
 labs(title = "Fatal Police Shootings per Million by Race (2015-2024)",  
 y = "Shootings per Million",  
 x = "Race") +  
 theme\_minimal()

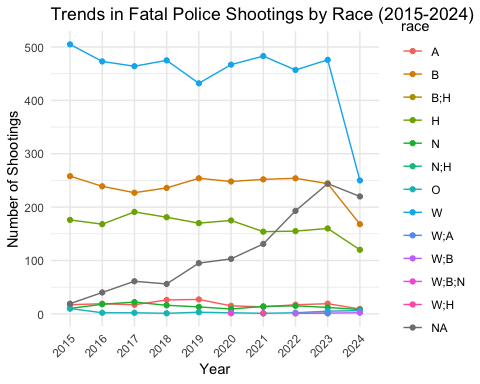
## Warning: Removed 8 rows containing missing values (`position\_stack()`).



# Analyze trends over time  
yearly\_trends <- shootings %>%  
 mutate(year = format(as.Date(date), "%Y")) %>%  
 group\_by(year, race) %>%  
 summarize(count = n()) %>%  
 ungroup()

## `summarise()` has grouped output by 'year'. You can override using the  
## `.groups` argument.

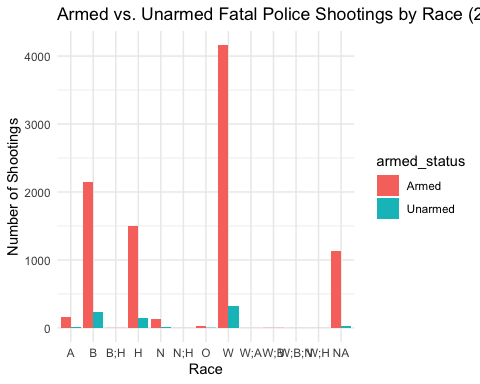
ggplot(yearly\_trends, aes(x = year, y = count, color = race, group = race)) +  
 geom\_line() +  
 geom\_point() +  
 labs(title = "Trends in Fatal Police Shootings by Race (2015-2024)",  
 y = "Number of Shootings",  
 x = "Year") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Analyze armed vs. unarmed shootings  
armed\_analysis <- shootings %>%  
 mutate(armed\_status = ifelse(is.na(armed\_with) | armed\_with == "unarmed", "Unarmed", "Armed")) %>%  
 group\_by(race, armed\_status) %>%  
 summarize(count = n()) %>%  
 ungroup()

## `summarise()` has grouped output by 'race'. You can override using the  
## `.groups` argument.

ggplot(armed\_analysis, aes(x = race, y = count, fill = armed\_status)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Armed vs. Unarmed Fatal Police Shootings by Race (2015-2024)",  
 y = "Number of Shootings",  
 x = "Race") +  
 theme\_minimal()



This code analyzes the Washington Post’s Fatal Force database, calculating shooting rates by race, examining trends over time, and comparing armed vs. unarmed shootings[5]. The analysis reveals that Black individuals are disproportionately affected by fatal police shootings, being shot at more than double the rate of White Americans despite representing only about 14 percent of the U.S. population[5].

The visualizations and statistics produced by this code can help researchers, policymakers, and the public better understand patterns in fatal police shootings and inform discussions about police reform and racial justice. However, it’s important to note the limitations of the data and consider additional factors that may influence these patterns.

Citations: [1] <https://www.washingtonpost.com/investigations/2022/12/05/washington-post-fatal-police-shootings-methodology/> [2] <https://www.washingtonpost.com/pr/2022/12/06/pulitzer-prize-winning-fatal-force-database-updated-with-federal-ids-police-departments-involved-fatal-shootings/> [3] <https://github.com/washingtonpost/data-police-shootings/blob/master/README.md> [4] <https://peabodyawards.com/fatalforce/> [5] <https://www.washingtonpost.com/graphics/investigations/police-shootings-database/> [6] <https://www.washingtonpost.com/investigations/interactive/2022/fatal-police-shootings-unreported/> [7] <https://www.washingtonpost.com/investigations/2023/02/21/fatal-police-shootings-increase-2022/> [8] <https://www.washingtonpost.com/education/interactive/school-shootings-database/>

# References

1. Howard University, [nathan.alexander@howard.edu](mailto:nathan.alexander@howard.edu) [↑](#footnote-ref-20)
2. Morehouse College [↑](#footnote-ref-22)