STAT 431 Final Project Report

Yiming Gao[*] Yuran Wang[†] Molin Yang[‡] Rongsheng Zhang[§]

December 12, 2025

Abstract

In this project, we investigate how serious crime rates vary across 21 districts of Los Angeles Police Department (LAPD). We apply a Bayesian hierarchical modeling framework and implement Markov Chain Monte Carlo methods using an open-access dataset from the Data.gov [1] platform. Specifically, we estimate posterior means and confidence intervals for district-level serious crime rates, conduct posterior predictive checks to assess model fit, and perform prior sensitivity analyses to evaluate the effectiveness of our results. Our findings quantify uncertainty in serious-crime rate estimates within Los Angeles and offer a modeling template for analyzing crime patterns in other metropolitan areas around the world.

1 Introduction

1.1 Motivation

Urban crime rarely occurs uniformly across space. Even within a single city, some districts experience a much larger share of serious crimes than others. These severe offenses include homicide, robbery, and aggravated assault. Successfully identifying these spatial differences in serious crime rates is essential for local governments and police departments in that it can generate promising effects. These include more efficient resource allocation, targeted prevention strategies, and evaluating whether particular communities face disproportionate exposure to violence.

1.2 Goal

The Los Angeles Police Department (LAPD) divides the city into 21 geographic areas, each with its own demographic composition and policing needs, creating a natural setting for our analysis. The primary goal of this project is to reflect and compare serious crime rates across the 21 geographic areas in Los Angeles.

[*]Department of Statistics, University of Illinois at Urbana-Champaign (email: yiming32@illinois.edu)
[†]Department of Statistics, University of Illinois at Urbana-Champaign (email: yuran4@illinois.edu)
[‡]Department of Statistics, University of Illinois at Urbana-Champaign (email: moliny2@illinois.edu)
[§]Department of Statistics, University of Illinois at Urbana-Champaign (email: rz36@illinois.edu)
[1] https://catalog.data.gov/dataset/crime-data-from-2020-to-present

However, since different geographic areas are exposed to different total numbers of incidents, comparing raw proportions of serious crimes between districts can produce misleading results. Under the assumption that district-level probability parameters arise from a common population distribution, a hierarchical model achieves a bias–variance tradeoff through partial pooling, allowing each district's estimate to be informed by both its local data and information shared across districts.

Hence, rather than focusing solely on raw sample proportions, we aim to:

- construct a probabilistic model for the number of serious crimes in each area, using the total number of reported crimes as the binomial "exposure";
- estimate area-specific serious-crime rates under a Bayesian hierarchical model that allows areas to share information while retaining their own local patterns; and
- assess the extent of spatial heterogeneity in serious-crime risk, including identifying areas that are credibly higher or lower than the city-wide average after accounting for uncertainty.

2 Data

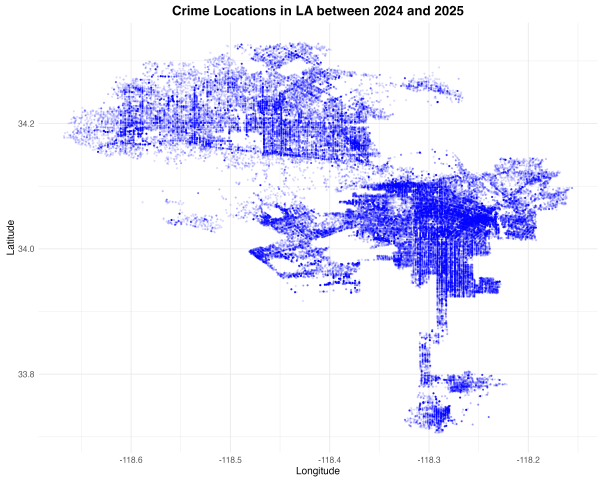
We use the public Crime Data from 2020 to Present dataset maintained by the LAPD and distributed through Data.gov [1]. Each row is a single reported crime incident. For this project, we extract all incidents with a valid date and police area code that occurred between January 1, 2024 and October 31, 2025. After cleaning, we retain crimes from the 21 LAPD Geographic Areas.

From the incident-level data, we keep only variables needed to study crime severity. In particular, `Primary Crime Code 1` is an official LAPD offense code indicating the seriousness of the crime in a particular incident. Following LAPD coding conventions, we classify an incident as a serious crime if its corresponding `Primary Crime Code 1` is less than 300. For each area, we then compute (i) the total number of reported crimes and (ii) the number of serious crimes in the study period; the proportion of serious crimes is the main response in our hierarchical models. We also use the latitude/longitude of each incident only for mapping and exploratory plots.

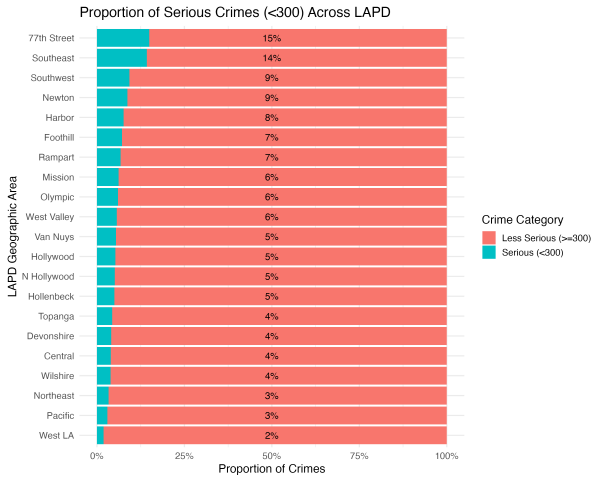
Table 1: Selected Variables from the LAPD Crime Dataset

Variable	**Definition**
Date of Occurrence	Calendar date when the crime occurred (MM/DD/YYYY), converted to a standard date-time format and restricted the study period to 2024–2025.
Police Area Code	Identifies one of the 21 LAPD Geographic Areas. Each area is a distinct administrative region and is the unit of analysis in our study.
Primary Crime Code 1	Official LAPD code for the primary and most serious offense in an incident. We define a serious crime as any incident with `Primary Crime Code 1` < 300.
Location (Longitude, Latitude)	Geographic coordinates of the incident, used to map crime locations and explore spatial patterns across Los Angeles.

Figure 1 provides a brief exploratory summary of the cleaned dataset. The left panel plots all recorded crime locations in Los Angeles over the study period, regions with darker color indicates higher number of crimes. The right panel shows the proportion of serious crimes in each LAPD area, illustrating clear spatial variation in crime severity.



(a) Crime Locations in LA (2024–2025)



(b) Proportion of Serious Crimes by Area

Figure 1: Exploratory Data Analysis of LAPD Crime Data

3 Model

Data and question: Let the LAPD Geographic Areas be indexed by $i = 1, \ldots, K$, where $K = 21$ is the total number of areas observed. We are interested in the number of "serious" crimes (defined as crimes with Crime Code 1 < 300), denoted by y_i, out of the total number of reported crimes N_i for each area during the fixed period (2024-2025). Assuming independent events, the counts can be modeled as:

$$y_i \mid p_i \sim \text{Binomial}(N_i, p_i),$$

where p_i is the underlying serious crime rate for area i.

Hierarchical prior: To borrow strength across areas and enable shrinkage, we place a shared conjugate prior on the proportions:

$$p_i \mid \alpha, \beta \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta), \quad \text{for } i = 1, \ldots, K.$$

To allow the data to inform the overall mean and variance of these rates, we assign independent hyperpriors to the shape parameters α and β:

$$\alpha \sim \text{Gamma}(a_1, b_1), \quad \beta \sim \text{Gamma}(a_2, b_2).$$

In our analysis, we set weakly informative fixed hyperparameters $a_1 = a_2 = 2$ and $b_1 = b_2 = 0.5$ which provide mild regularization while allowing the data to dominate posterior inference.

Posterior and computation: By Bayes' rule, the joint posterior distribution is proportional to the product of the Binomial likelihood, the Beta priors, and the Gamma hyperpriors:

$$\pi(\mathbf{p}, \alpha, \beta \mid \mathbf{y}) \propto \left[\prod_{i=1}^{K} \text{Binomial}(y_i \mid N_i, p_i) \times \text{Beta}(p_i \mid \alpha, \beta) \right] \times \text{Gamma}(\alpha \mid a_1, b_1) \times \text{Gamma}(\beta \mid a_2, b_2).$$

Since the full posterior is analytically intractable, we use Markov Chain Monte Carlo (MCMC) methods implemented in JAGS to approximate the posterior distributions of the parameters $\mathbf{p} = (p_1, \ldots, p_K)$, α, and β.

4 Computational Strategy

We implemented the Bayesian hierarchical Beta–Binomial model in JAGS, whose Gibbs sampling framework efficiently handles models with conjugate full conditional distributions. The hierarchical structure enables partial pooling across the 21 LAPD areas, yielding stabilized esti-

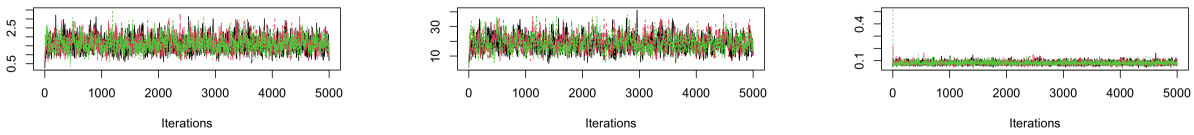
mates for districts with limited data. To assess convergence and avoid dependence on initial values, we ran three parallel chains with deliberately over-dispersed starting points for (α, β): $(0.5, 0.5)$, $(2.0, 2.0)$, and $(5.0, 1.0)$.

For each chain, the first 1,500 iterations were discarded as burn-in. We then obtained 100,000 additional iterations per chain, resulting in 300,000 posterior draws in total. No thinning was applied, as autocorrelation remained moderate and storage constraints were minimal. The large effective sample size supports precise estimation of posterior means, standard deviations, and credible intervals for all monitored parameters.

5 Diagnostic and Robustness Check

5.1 Traceplot

Figure 2 provides visual evidence of satisfactory MCMC performance. Good mixing is indicated by the rapid fluctuation of the chains around a stable level and the substantial overlap across chains, with no visible trends or drift. These visual patterns are characteristic of samples drawn from the stationary posterior distribution and therefore suggest that the MCMC algorithm has successfully converged.



(a) α (b) β (c) μ_p

Figure 2: Trace plots for the monitored parameters.

Note: Each panel shows posterior samples from three parallel MCMC chains after burn-in.

In addition to the visual diagnostics discussed above, we also include the Gelman–Rubin statistic (PSRF) as a supplementary convergence check. The PSRF compares the variability within each MCMC chain to the variability between chains; when all chains have converged to the same posterior distribution, these quantities should be nearly identical, producing values very close to 1. Our results in 2 show PSRF values essentially equal to 1, providing additional confirmation of convergence.

5.2 Posterior Predictive Check

We performed posterior predictive check (PPC) to validate that our proposed model provides an adequate fit.

Figure 3 displays the posterior predictive distribution of the total serious crime count across all areas, yielding a Bayesian p-value of 0.52. This value indicates that approximately 52% of the replicated datasets produced totals greater than or equal to the observed count. A p-value close to 0.5 suggests the model adequately reproduces the key summary statistic of the data, providing evidence that our hierarchical Beta-Binomial model captures the essential features of the crime distribution across LAPD areas.

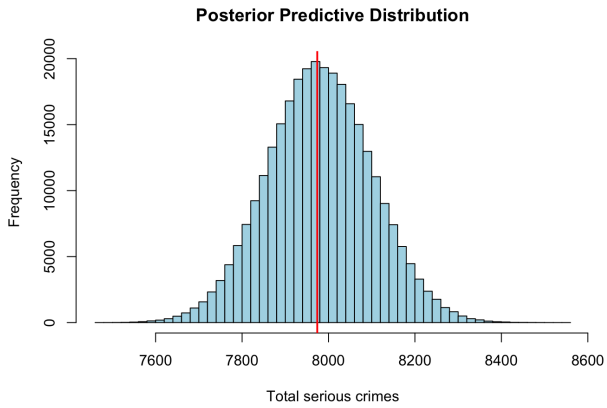


Figure 3: Posterior Predictive Distribution of Total Serious Crimes.

Note: The distribution of total serious crimes predicted by the model (histogram) compared to the observed value (red line). The observed data falls within the central region of the predictive distribution, indicating adequate model fit.

5.3 Sensitivity Analysis

To validate the stability of our posterior inferences against prior specification, we conducted a sensitivity analysis by comparing the baseline model ($\alpha, \beta \sim \text{Gamma}(2, 0.5)$) with a more diffuse alternative ($\alpha, \beta \sim \text{Gamma}(1, 0.25)$). Figure 4 displays the correspondence between the posterior mean estimates of the area-level serious crime rates (p_i) under these two regimes. The estimates align almost perfectly along the 45-degree identity line, indicating that the results are virtually invariant to the choice of hyperpriors. This confirms that the inference is primarily data-driven, with the likelihood of the observed data dominating the prior information.

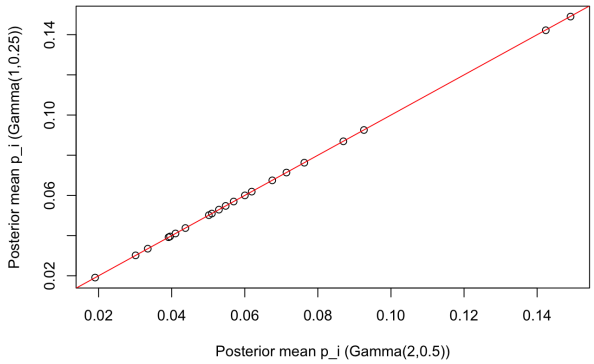


Figure 4: Sensitivity of posterior means to hyperprior choice.

Note: Each point compares the posterior mean estimate of p_i under the baseline hyperpriors (horizontal axis) versus diffuse hyperpriors (vertical axis). The red line denotes the identity line. Both models were fit using identical MCMC settings (300,000 post–burn-in draws across three chains).

6 Results and Discussions

The hierarchical model provides estimated serious crime rates (p_i) for all 21 LAPD geographic areas.

Table 2 presents the posterior summaries for the hyperparameters and the global serious crime rate. The convergence diagnostics provide strong evidence of model stability: all monitored parameters exhibit negligible Monte Carlo Standard Errors and Gelman-Rubin statistics of approximately 1.01, confirming that the MCMC sampler has successfully converged to the stationary distribution. Statistically, the posterior estimates for the hyperparameters yield a global mean crime rate of approximately 0.08, which serves as the population-level baseline toward which individual area estimates are shrunk.

Table 2: Posterior summaries for hyperparameters and the global serious crime rate.

Parameter	Mean	SD	2.5%	97.5%	MCSE	$\hat{R}$
α (shape 1)	1.61	0.40	0.93	2.50	0.0021	1
β (shape 2)	18.79	5.38	9.71	30.68	0.028	1
μ_p (serious crime rate)	0.08	0.01	0.06	0.11	0.000026	1

Note: Each row reports posterior summaries based on 300,000 MCMC draws from three parallel chains. Mean and SD denote the posterior mean and standard deviation. Columns 4 and 5 give the bounds of the central 95% credible interval. MCSE is the time-series Monte Carlo standard error accounting for autocorrelation. $\hat{R}$ is the Gelman–Rubin convergence diagnostic.

Complementing the Bayesian analysis, Table 3 summarizes the raw empirical proportions for the 21 LAPD geographic areas. These represent the frequentist maximum likelihood estimates calculated under the assumption that each area is independent. The data reveal substantial spatial heterogeneity in crime distribution. The 77th Street exhibit the highest frequentist rates, but ar-

eas such as West LA and Pacific show the lowest proportions, clustering between 0.02 and 0.03. The majority of districts fall within an intermediate band of 0.04 to 0.08. This observable variation—combined with the differing sample sizes across districts—provides the empirical motivation for adopting a hierarchical Bayesian approach to stabilize estimates in areas with higher statistical noise.

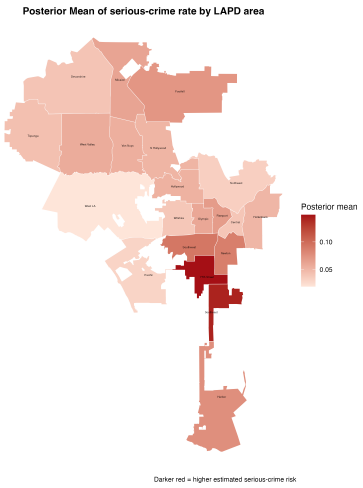
Table 3: Frequentist estimates of serious crime rates by LAPD area.

Area	y_i	N_i	$\hat{p}_i$	Area	y_i	N_i	$\hat{p}_i$	Area	y_i	N_i	$\hat{p}_i$
77th Street	1012	6777	0.15	Mission	332	5369	0.06	Southeast	800	5611	0.14
Central	403	10214	0.04	N Hollywood	369	7249	0.05	Southwest	770	8314	0.09
Devonshire	241	5892	0.04	Newton	469	5390	0.09	Topanga	242	5548	0.04
Foothill	282	3951	0.07	Northeast	175	5260	0.03	Van Nuys	315	5761	0.06
Harbor	405	5309	0.08	Olympic	346	5771	0.06	West LA	102	5416	0.02
Hollenbeck	216	4317	0.05	Pacific	245	8163	0.03	West Valley	301	5292	0.06
Hollywood	325	6152	0.05	Rampart	378	5602	0.07	Wilshire	246	6306	0.04

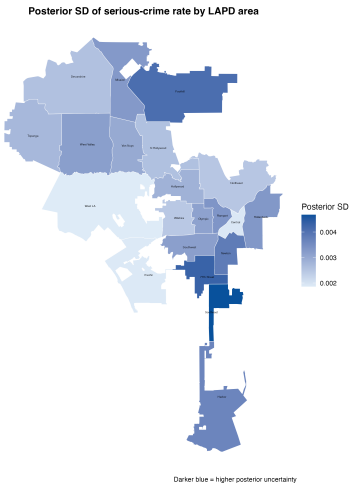
Note: Each row corresponds to one LAPD geographic area. y_i is the number of serious crimes in 2024–2025, N_i is the total number of reported crimes, and $\hat{p}_i$ is the frequentist estimate of the serious crime rate obtained as the sample proportion y_i/N_i.

To investigate the spatial pattern of serious crime rates, we obtained the official LAPD division boundaries from the department's public mapping resources and overlaid our posterior estimates onto the corresponding geographic regions.

Figure 5 presents the resulting spatial summaries. Panel (a) shows the posterior mean estimates of the serious-crime rates across the 21 areas, revealing clear geographic variation in crime intensity. Panel (b) displays the posterior standard deviations, which identify regions with greater estimation uncertainty and therefore less stable local information.



(a) Posterior mean of p_i



(b) Posterior standard deviation of p_i

Figure 5: Spatial posterior estimates of serious crime rates across LAPD areas.

Note: The left panel displays the posterior mean serious crime rate for each area, where darker shades indicate higher estimated risk. The right panel shows the posterior standard deviation, reflecting uncertainty; darker regions correspond to greater uncertainty due to smaller sample sizes or greater variability.

To further quantify these differences and visualize estimation uncertainty, Figure 6 presents the posterior distributions for each area, ranked by their posterior mean. The red dots represent the raw sample proportions (y_i/N_i), while the boxplots depict the 95% credible intervals. Notably, the precision of these estimates varies across the spectrum: areas with lower crime rates (bottom-left) exhibit markedly narrower credible intervals compared to those with higher rates (top-right). This pattern reflects both the theoretical mean-variance relationship of the Binomial distribution—where variance increases with the proportion p—and the variation in sample sizes, where districts with larger total counts yield more precise posterior inference.

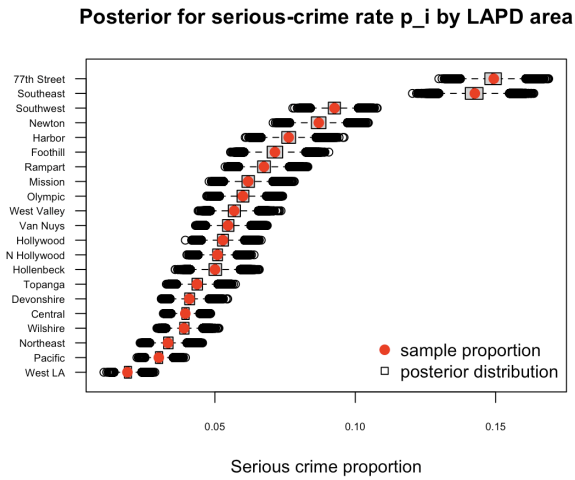


Figure 6: Posterior distributions of the area-specific serious-crime rates.

Note: Each boxplot summarizes the posterior distribution of p_i for a given LAPD area, sorted by the posterior mean. The red dots denote the empirical proportions y_i/N_i. The horizontal extent of each boxplot represents the width of the central 95% credible interval for p_i.

These posterior findings corroborate the preliminary patterns observed in our exploratory analysis in Figure 1. The distinct spatial gradient—characterized by elevated risk in southern divisions versus lower risk in western districts—persists even after hierarchical regularization. This consistency confirms that the geographic disparities identified in the raw data are driven by a robust underlying signal rather than stochastic fluctuations in local crime counts.

7 Conclusion

To sum up, this study applies a Bayesian hierarchical Beta–Binomial model to estimate serious-crime rates across the 21 LAPD geographic areas. The model shows strong MCMC convergence and produces stabilized posterior estimates that appropriately shrink extreme sample proportions, especially in regions with smaller crime counts.

The results reveal clear geographic disparities: southern districts exhibit the highest underly-

ing crime risk, while western areas remain comparatively low. By quantifying uncertainty and borrowing strength across areas, the hierarchical approach provides more reliable estimates than raw proportions and offers a statistically principled foundation for resource allocation and policy decisions.

8 Group Members Contributions

The contributions of each group member to this project are listed below:

- **Yiming Gao:** Performed MCMC convergence diagnostics (including trace plots and Gelman-Rubin statistics) and wrote the Model section.
- **Yuran Wang:** Conducted the sensitivity analysis, created the posterior spatial maps, and drafted the Results section.
- **Molin Yang:** Responsible for determining the priors and hyperpriors for the final model, performed PPC and its interpretations, and drafted the Introduction and Data sections.
- **Rongsheng Zhang:** Handled data cleaning and visualization, developed the Bayesian hierarchical model in JAGS, and compiled the Appendix and code.

All members participated in the interpretation of the results and reviewed the final manuscript.

9 AI Usage Statement

We use generative AI tools (ChatGPT and Gemini) to assist with debugging R code and refining the academic English and formatting of this report. No AI tools were used to conduct the statistical inference or interpret the findings. All final content, code, and results were verified by the group members.

References

City of Los Angeles (2025). *Crime Data from 2020 to Present*. https://catalog.data.gov/dataset/crime-data-from-2020-to-present. Accessed November 2025.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reich, Brian J. and Sujit K. Ghosh (2019). *Bayesian Statistical Methods*. Boca Raton, FL: CRC Press.

Appendix

Figures

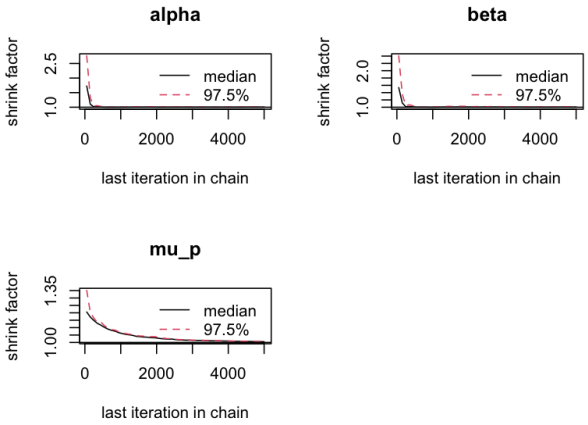
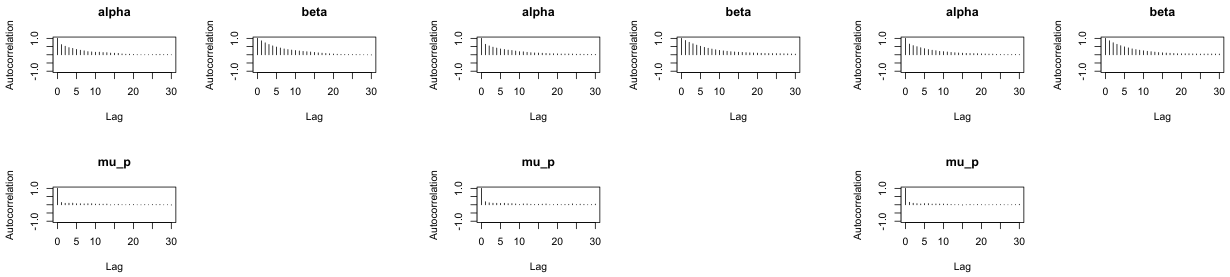


Figure 7: Gelman-Rubin Convergence Diagnostic.

Note: The shrink factor approaches 1 as the number of iterations increases, indicating that the chains have converged to the stationary distribution.



(a) Chain 1 (b) Chain 2 (c) Chain 3

Figure 8: Autocorrelation Plots for MCMC Chains.

Note: The rapid decay of autocorrelation across all three chains confirms that the samples are effectively independent and mixing well.

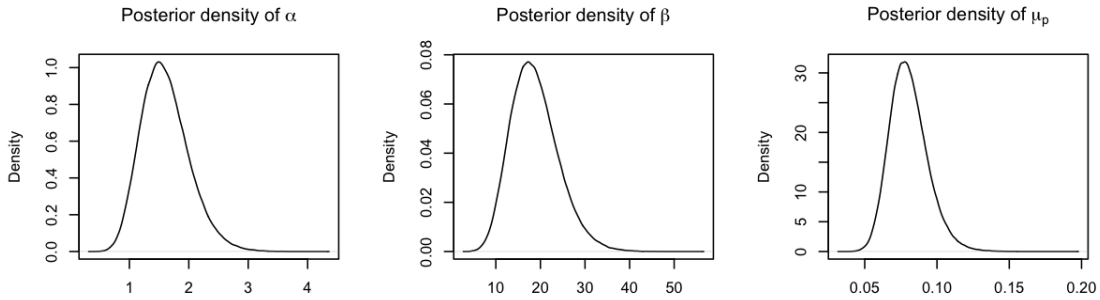


Figure 9: Posterior Density Estimates.

Note: Density plots for the hyperparameters α and β, and μ_p. The smooth, unimodal distributions further support model stability.

Code

```
##
    ------------------------------------------------------------------------

library(dplyr)
library(lubridate)
library(janitor)

crimes = read_csv("~/Downloads/Crime_Data_from_2020_to_Present.csv",
    show_col_types = FALSE) %>%
  clean_names() %>%
  mutate(date_occ = mdy_hms(date_occ)) %>%
  filter(year(date_occ) %in% c(2024, 2025))

head(crimes)
print(dim(crimes))




##
    ------------------------------------------------------------------------

library(dplyr)
library(ggplot2)
library(scales)
```

```r

area_serious_counts <- crimes %>%
  mutate(
    serious = if_else(crm_cd_1 < 300,
                      "Serious (<300)",
                      "Less Serious (>=300)")
  ) %>%
  group_by(area_name, serious) %>%
  summarise(count = n(), .groups = "drop")

area_serious_rate <- area_serious_counts %>%
  group_by(area_name) %>%
  mutate(rate = count / sum(count))

ordered_areas <- area_serious_rate %>%
  filter(serious == "Serious (<300)") %>%
  arrange(rate) %>%
  pull(area_name)

area_serious_rate$area_name <- factor(area_serious_rate$area_name,
   levels = ordered_areas)

ggplot(area_serious_rate, aes(x = area_name, y = rate, fill = serious))
    +
  geom_col(position = "fill") +
  geom_text(
    data = subset(area_serious_rate, serious == "Serious (<300)"),
    aes(label = percent(rate, accuracy = 1)),
    position = position_fill(vjust = 0.5),
    color = "black",
    size = 4
  ) +
  coord_flip() +
  labs(
    title = "Proportion of Serious Crimes (<300) Across LAPD",
    x = "LAPD Geographic Area",
    y = "Proportion of Crimes",
    fill = "Crime Category"
```

```
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_minimal(base_size = 14)



##
    ------------------------------------------------------------------------

library(ggplot2)
library(dplyr)


ggplot(crimes, aes(x = lon, y = lat)) +
  geom_point(alpha = 0.1, color = "blue", size = 0.5) +
  labs(title = "Crime Locations in LA between 2024 and 2025", x = "
     Longitude", y = "Latitude") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10)
  )


##
    ------------------------------------------------------------------------

library(dplyr)

# By area_name
serious_by_area <- crimes %>%
  mutate(serious = crm_cd_1 < 300) %>%   # TRUE = serious crime
  group_by(area_name) %>%
  summarise(
    y = sum(serious, na.rm = TRUE),      # serious crimes count
    N = n(),                             # total crimes in area
    prop_serious = y / N,
```

```r
    .groups = "drop"
  ) %>%
  arrange(area_name)

serious_by_area
n_areas <- nrow(serious_by_area)

y <- serious_by_area$y
N <- serious_by_area$N


##
    ------------------------------------------------------------------------


# hyperprior
a1 <- 2; b1 <- 0.5   # alpha ~ Gamma(2, 0.5)
a2 <- 2; b2 <- 0.5   # beta  ~ Gamma(2, 0.5)

jags_data_serious <- list(
  y  = y,
  N  = N,
  K  = n_areas,
  a1 = a1, b1 = b1,
  a2 = a2, b2 = b2
)


##
    ------------------------------------------------------------------------


serious_model_string <- "
model {
  # likelihood: Binomial for serious crimes in each LAPD area
  for (i in 1:K) {
    y[i] ~ dbin(p[i], N[i])      # Y_i | p_i ~ Binomial(N_i, p_i)
    p[i] ~ dbeta(alpha, beta)    # area-level serious crime rate
  }

```

```
  # hyperpriors for Beta shape parameters (partially informative)
  alpha ~ dgamma(a1, b1)
  beta  ~ dgamma(a2, b2)

  # derived quantity: overall mean serious-crime rate across areas
  mu_p <- alpha / (alpha + beta)
}
"

# write as bug
writeLines(serious_model_string, con = "serious_model.bug")



##
    ------------------------------------------------------------------------


library(rjags)
library(coda)

# three chain
inits_list_serious <- list(
  list(alpha = 0.5, beta = 0.5),
  list(alpha = 2.0, beta = 2.0),
  list(alpha = 5.0, beta = 1.0)
)

params_pre <- c("alpha", "beta", "mu_p")

mod_serious <- jags.model(
  file     = "serious_model.bug",
  data     = jags_data_serious,
  inits    = inits_list_serious,
  n.chains = 3,
  n.adapt  = 0
)

pre_samp_serious <- coda.samples(
```

```
  model = mod_serious,
  variable.names = params_pre,
  n.iter = 5000
)

# traceplot
traceplot(pre_samp_serious[, "alpha"])
traceplot(pre_samp_serious[, "beta"])
traceplot(pre_samp_serious[, "mu_p"])

# Gelman-Rubin PSRF
gelman.diag(pre_samp_serious, autoburnin = FALSE)

gelman.plot(pre_samp_serious, autoburnin = FALSE)

# Autocorrelation plots for one chain
autocorr.plot(window(pre_samp_serious, 1500)[[1]])
autocorr.plot(window(pre_samp_serious, 1500)[[2]])
autocorr.plot(window(pre_samp_serious, 1500)[[3]])



##
    ------------------------------------------------------------------------


# rebuild the model + burn-in
mod_serious <- jags.model(
  file     = "serious_model.bug",
  data     = jags_data_serious,
  inits    = inits_list_serious,
  n.chains = 3,
  n.adapt  = 0
)

# drop first 1500 as burn-in
update(mod_serious, 1500)

params_post <- c("alpha", "beta", "mu_p", "p")
```

```r

post_samp_serious <- coda.samples(
  model = mod_serious,
  variable.names = params_post,
  n.iter = 100000
)

summary(post_samp_serious)



## ------------------------------------------------------------------------

post_mat <- as.matrix(post_samp_serious)

alpha_draws <- post_mat[, "alpha"]
beta_draws  <- post_mat[, "beta"]
mu_p_draws  <- post_mat[, "mu_p"]

# Some summary
c(
  mean_alpha = mean(alpha_draws),
  sd_alpha   = sd(alpha_draws)
)

c(
  mean_beta = mean(beta_draws),
  sd_beta   = sd(beta_draws)
)

quantile(mu_p_draws, c(0.025, 0.5, 0.975))  # overall serious crime rate 95% CI

# plot posterior density
par(mfrow = c(1, 3))
plot(density(alpha_draws),
     main = expression(paste("Posterior density of ", alpha)),
```

```r
     xlab = "")
plot(density(beta_draws),
     main = expression(paste("Posterior density of ", beta)),
     xlab = "")
plot(density(mu_p_draws),
     main = expression(paste("Posterior density of ", mu[p])),
     xlab = "")


##
    ------------------------------------------------------------------------


# find all p[i]
p_cols <- grep("^p\\[", colnames(post_mat))
p_draws <- post_mat[, p_cols]

p_summary <- apply(
  p_draws,
  2,
  function(z) c(
    mean = mean(z),
    sd   = sd(z),
    q2.5 = quantile(z, 0.025),
    q50  = quantile(z, 0.50),
    q97.5 = quantile(z, 0.975)
  )
)

p_summary <- as.data.frame(t(p_summary))
p_summary$area_name <- serious_by_area$area_name

# Sort by posterior mean
p_summary <- p_summary %>%
  arrange(desc(mean))

p_summary
```

```r
##
    ------------------------------------------------------------------------

# Box plot with posterior mean and CI
ord <- order(colMeans(p_draws), decreasing = FALSE)

par(mar = c(5, 10, 4, 2) + 0.1)

boxplot(
  p_draws[, ord],
  horizontal = TRUE,
  names = serious_by_area$area_name[ord],
  las = 1,
  cex.axis = 0.60,
  xlab = "Serious crime proportion",
  main = "Posterior for serious-crime rate p_i by LAPD area"
)

points(
  serious_by_area$prop_serious[ord],
  1:ncol(p_draws),
  col = "red", pch = 19, cex = 1.2
)

legend("bottomright", legend = c("sample proportion", "posterior
    distribution"),
       pch = c(19, 22), pt.cex = c(1.2, 1), col = c("red", "black"),
          bty = "n")



##
    ------------------------------------------------------------------------

library(sf)
library(dplyr)
library(ggplot2)

```

```r
lapd_sf <- st_read("~/Downloads/LAPD_Divisions.geojson", quiet = TRUE)

area_lookup <- crimes %>%
  distinct(area, area_name) %>%
  arrange(area)

area_lookup

name_match <- tibble::tibble(
  area_name = c(
    "Central",
    "Rampart",
    "Southwest",
    "Hollenbeck",
    "Harbor",
    "Hollywood",
    "Wilshire",
    "West LA",
    "Van Nuys",
    "West Valley",
    "Northeast",
    "77th Street",
    "Newton",
    "Pacific",
    "N Hollywood",
    "Foothill",
    "Devonshire",
    "Topanga",
    "Olympic",
    "Mission",
    "Southeast"
  ),
  APREC = c(
    "CENTRAL",
    "RAMPART",
    "SOUTHWEST",
    "HOLLENBECK",
    "HARBOR",
```

```
    "HOLLYWOOD",
    "WILSHIRE",
    "WEST LOS ANGELES",
    "VAN NUYS",
    "WEST VALLEY",
    "NORTHEAST",
    "77TH STREET",
    "NEWTON",
    "PACIFIC",
    "NORTH HOLLYWOOD",
    "FOOTHILL",
    "DEVONSHIRE",
    "TOPANGA",
    "OLYMPIC",
    "MISSION",
    "SOUTHEAST"
  )
)

p_map <- p_summary %>%
  select(area_name, mean) %>%
  left_join(name_match, by = "area_name")

lapd_map <- lapd_sf %>%
  left_join(p_map, by = "APREC")

ggplot(lapd_map) +
  geom_sf(aes(fill = mean), color = "white", size = 0.2) +
  geom_sf_text(aes(label = area_name), size = 1.2, color = "black") +
  scale_fill_gradient(
    low  = "#fee5d9",
    high = "#a50f15",
    name = "Posterior mean"
  ) +
  labs(
    title   = "Posterior Mean of serious-crime rate by LAPD area",
    caption = "Darker red = higher estimated serious-crime risk"
  ) +
```

```r
  theme_void(base_size = 10) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "right"
  )



##
    ------------------------------------------------------------------------

library(dplyr)
library(ggplot2)
library(sf)

p_sd_map <- p_summary %>%
  select(area_name, sd) %>%
  left_join(name_match, by = "area_name")

lapd_sd_map <- lapd_sf %>%
  left_join(p_sd_map, by = "APREC")

ggplot(lapd_sd_map) +
  geom_sf(aes(fill = sd), color = "white", size = 0.2) +
  geom_sf_text(aes(label = area_name), size = 1.2, color = "black") +
  scale_fill_gradient(
    low  = "#deebf7",
    high = "#08519c",
    name = "Posterior SD"
  ) +
  labs(
    title   = "Posterior SD of serious-crime rate by LAPD area",
    caption = "Darker blue = higher posterior uncertainty"
  ) +
  theme_void(base_size = 10) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "right"
```

```
  )


##
    ------------------------------------------------------------------------


library(rjags)
library(coda)
library(dplyr)

run_hyper_model <- function(a1, b1, a2, b2, label) {
  data_list <- list(
    y  = y,
    N  = N,
    K  = n_areas,
    a1 = a1, b1 = b1,
    a2 = a2, b2 = b2
  )

  mod <- jags.model(
    file     = "serious_model.bug",
    data     = data_list,
    n.chains = 3,
    n.adapt  = 0
  )

  update(mod, 1500)

  samp <- coda.samples(
    model = mod,
    variable.names = c("alpha", "beta", "p"),
    n.iter = 40000
  )

  mat <- as.matrix(samp)
  p_cols <- grep("^p\\[", colnames(mat))

  list(
```

```r
    label      = label,
    alpha_mean = mean(mat[, "alpha"]),
    beta_mean  = mean(mat[, "beta"]),
    p_means    = colMeans(mat[, p_cols])
  )
}

# baseline
fit_base <- run_hyper_model(2, 0.5, 2, 0.5, "Gamma(2,0.5)")

fit_diffuse <- run_hyper_model(1, 0.25, 1, 0.25, "Gamma(1,0.25)")

fit_conc <- run_hyper_model(4, 1, 4, 1, "Gamma(4,1)")

# put all into a dataframe
sens_df <- data.frame(
  area_name = serious_by_area$area_name,
  base      = fit_base$p_means,
  diffuse   = fit_diffuse$p_means,
  conc      = fit_conc$p_means
)

head(sens_df)


cor(sens_df$base, sens_df$diffuse)
cor(sens_df$base, sens_df$conc)


plot(
  sens_df$base, sens_df$diffuse,
  xlab = "Posterior mean p_i (Gamma(2,0.5))",
  ylab = "Posterior mean p_i (Gamma(1,0.25))"
)
abline(0, 1, col = "red")

# Posterior Predictive Distribution
y_rep <- matrix(NA, nrow = nrow(post_mat), ncol = n_areas)
```

```r
for (i in 1:nrow(post_mat)) {
  for (j in 1:n_areas) {
    y_rep[i, j] <- rbinom(1, N[j], post_mat[i, paste0("p[", j, "]")])
  }
}

T_obs <- sum(y)
T_rep <- apply(y_rep, 1, sum)

# Bayesian p-value
mean(T_rep >= T_obs)

hist(T_rep, breaks = 50, col = "lightblue",
     main = "Posterior Predictive Distribution",
     xlab = "Total serious crimes")
abline(v = T_obs, col = "red", lwd = 2)
```