STAT7630: Bayesian Statistics
Lecture Slides # 16

Normal Hierarchical Models & Bayesian Version of ANOVA
Chapter 16 (Normal) Hierarchical Models without Predictors

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: November, 2024

Outline

Outline

An Example of Hierarchical Data

- This section focuses on the `spotify` dataset available in the `bayesrules` R package.
- The dataset is a subset of a comprehensive collection of Spotify songs compiled by Kaylin Pavlik in 2019.
- The response variable of interest is the *popularity* score of 350 songs.
- Songs are grouped by artist (bands or solo performers), creating a hierarchical (clustered) data structure.
- Popularity scores for songs by the same artist exhibit potential intra-group correlation, reflecting shared characteristics or fanbase influence.

Complete Pooled Approach

- Initially, we analyze the data under the **complete pooling** assumption, disregarding the hierarchical grouping structure.
- **Notation:**
  - Y_{ij} represents the popularity of the i-th song for the j-th artist.
  - n_j denotes the number of songs attributed to artist j in the dataset.
- For example, the first artist, Mia X, has 4 songs, implying $n_1 = 4$.
- The total sample size is computed as:

$$n = \sum_{j=1}^{44} n_j = n_1 + n_2 + \cdots + n_{44} = 350.$$

Complete Pooled Data Model

- Ignoring the grouping structure, we assume the popularity values follow a normal distribution:

$$Y_{ij} \mid \mu, \sigma^2 \sim N(\mu, \sigma^2).$$

- To assess the assumption of normality, we examine the estimated density of the popularity variable (see next slide).
- Formal Bayesian Normal-Normal model specification:

$$\mu \sim N(50, 52^2), \quad \sigma \sim \text{Exp}(0.048).$$

- Key assumptions:
  - The prior for μ centers around 50, reflecting the plausible range of popularity values (0 to 100).
  - A weakly informative prior is imposed on σ to allow flexibility in variance estimation.

Estimated Density of Popularity

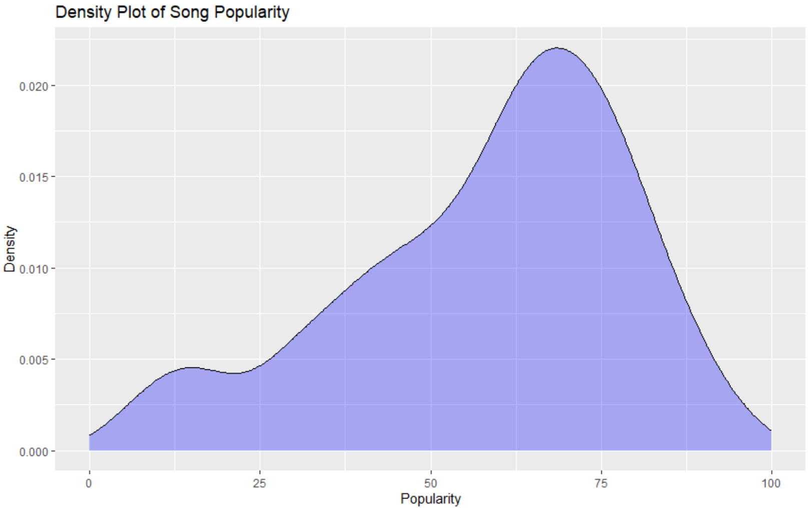

Figure 1: A density plot of the variability in popularity from song to song (with artists pooled).

Meaning of Model Parameters

- In this model, the parameters μ and σ are **global parameters**:
  - They remain constant across all artists in the dataset.
- Interpretation of the parameters:
  - μ: Global mean popularity.
  - σ: Global standard deviation in popularity across songs.
- This model is mathematically equivalent to a normal regression model without predictors:

$$Y_{ij} = \beta_0 + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2).$$

- Estimation can be performed using `stan_glm` with the formula:

$$\texttt{popularity} \sim 1$$

Drawback of Complete Pooling Model

- The posterior mean μ, estimated from this model, provides a single value for the overall mean popularity.
- **Major drawback:** Predictions for new songs from different artists are identical under this model.
- For any artist, the predicted popularity of a new song is the posterior mean:

$$\mathbf{E}(\mu \mid \mathbf{y}) = 58.39.$$

- Using `R`, we can visualize this limitation:
  - Posterior predictive means for each artist (light blue dots) can be plotted against sample means for each artist (dark blue dots) (see the plot in the next slide)
- **Observations:**
  - The posterior predictive means fail to capture inter-artist variability.
  - This demonstrates the model's inability to reflect actual differences in artist popularity — a significant limitation.

Posterior Predictive Intervals of Popularity for Each Artist - Pooled Model

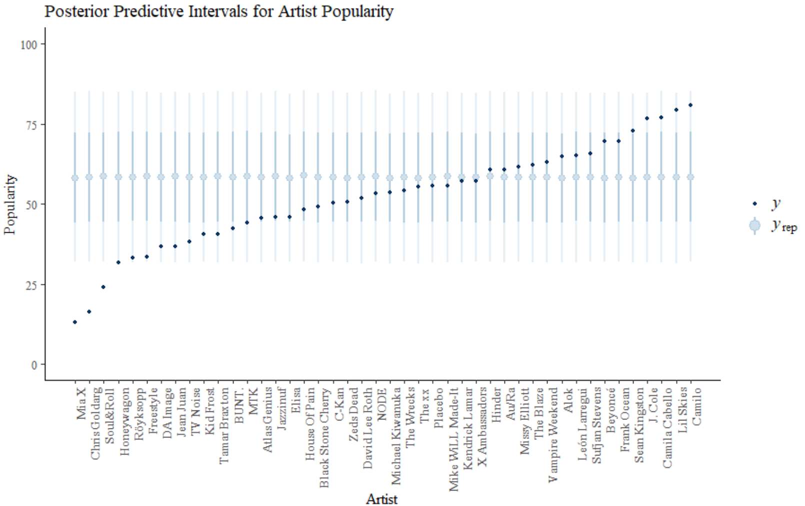

Figure 2: Posterior predictive intervals for artist song popularity, as calculated from a complete pooled model.

No Pooled Model

- The **no pooling** approach allows each artist to have a distinct mean popularity, μ_j:

$$Y_{ij} \mid \mu_j, \sigma \sim N(\mu_j, \sigma^2).$$

- Parameter interpretations:
  - μ_j: Mean song popularity for artist j.
  - σ: Standard deviation in song popularity within each artist.
- Key assumption: σ is same across all artists, meaning the variability in popularity is assumed constant between artists.
- Does this assumption align with reality? (`R` plot in next slide):
  - Evidence suggests σ might differ across artists.
- Despite potential misalignment, we proceed with this model for simplicity, as a shared σ reduces model complexity.

Density Plots of Popularity by Artist

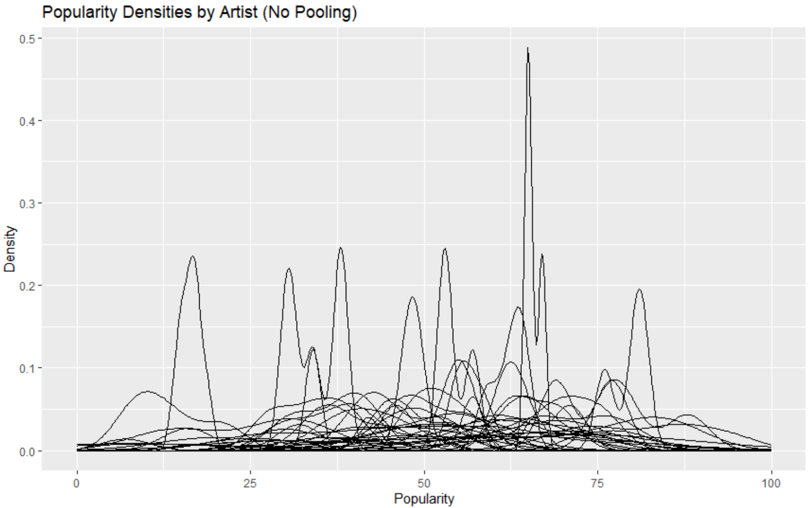

Figure 3: Density plots of the variability in popularity from song to song, by artist.

Formal No Pooling Model

- The **no pooling model** introduces a large number of parameters, specifically $44 + 1 = 45$:

$$Y_{ij} \mid \mu_j, \sigma^2 \sim N(\mu_j, \sigma^2),$$

$$\mu_j \sim N(50, s^2), \quad \sigma \sim \text{Exp}(0.048).$$

- Estimation approach:
    - A regression model with separate coefficients for each artist and no intercept can be specified as:

$$\texttt{popularity} \sim \texttt{artist} - 1.$$

- Prior specification:
    - The priors on μ_j are weakly informative, centered at 50.
    - Weak priors allow the data to dominate, leading to posterior means closely reflecting sample means.
- **Result:** The posterior predictive distribution of popularity for each artist aligns closely with their respective sample means (see `R` plot).

Posterior Predictive Intervals of Popularity for Each Artist - Non-Pooled Model

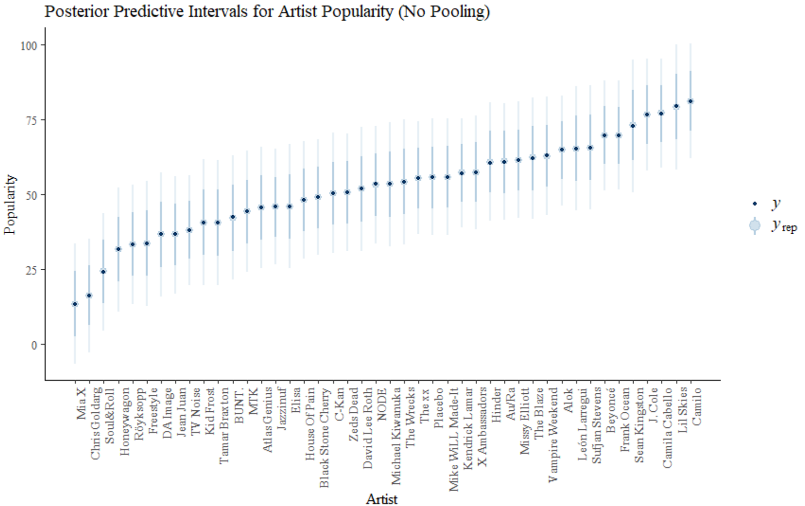

Figure 4: Posterior predictive intervals for artist song popularity, as calculated from a no pooled model.

Drawbacks of the No-Pooling Model

- **Limited Data Sharing:** This model assumes no information is shared across groups (artists), meaning:
    - Data from one artist cannot inform estimates for another artist.
- **Small Sample Size Limitation:**
    - For groups with small sample sizes (e.g., artists with few songs), estimates of mean popularity are imprecise.
- **Lack of Generalizability:**
    - The model cannot predict the mean popularity for an artist outside the sample (e.g., Taylor Swift).
- **Sample-Restricted Inference:**
    - Inferences are limited to the artists included in the dataset, offering no insight into the broader population of artists.

A Better Approach: Hierarchical Model

- A hierarchical model provides a more robust framework for handling this dataset by incorporating three layers:
  1. **Within-group variability:** Describes how song popularity varies within each artist j.
  2. **Between-group variability:** Models how the artist-specific mean song popularity, μ_j, varies across artists.
  3. **Global priors:** Specifies prior distributions for the global parameters μ, σ_y, and σ_μ.
- This approach leverages the hierarchical structure of the data, allowing partial pooling of information across artists while preserving individual characteristics.

Outline

Within-Group Normal Model

- Assume the data values within each group (artist j) follow a normal distribution:

$$Y_{ij} \mid \mu_j, \sigma_y \sim N(\mu_j, \sigma_y^2).$$

- **Key features of the model:**
  - Each artist is allowed to have their own mean song popularity, μ_j, similar to the no-pooling model.
  - σ_y represents the within-group variability, measuring the standard deviation of popularity from song to song for a given artist.
- **Assumption:**
  - The within-group variability σ_y is assumed to be constant across all artists.
  - This assumption may not hold in reality; always verify through diagnostic plots of the data.

Between-Group Layer

- Unlike the no-pooling model, the hierarchical model incorporates a **between-group layer**, recognizing that all sampled artists are drawn from a single population.
- Variability in the artist-specific mean popularities, μ_j, is modeled as: $\mu_j \mid \mu, \sigma_\mu \sim N(\mu, \sigma_\mu^2)$.
- **Parameter interpretations:**
  - μ: The global average of mean song popularity (μ_j) across all artists.
  - σ_μ: The between-group variability, representing the standard deviation of μ_j among artists.
- **Assumption:**
  - Normality is assumed for μ_j.
  - While μ_j is not directly observable, the sample mean song popularity for each artist serves as an estimate.
- **Diagnostic Check:**
  - A density plot of the artist sample means (`R` code) suggests the normality assumption is reasonable (see next slide).

Density Plots of Mean Popularity of Artists

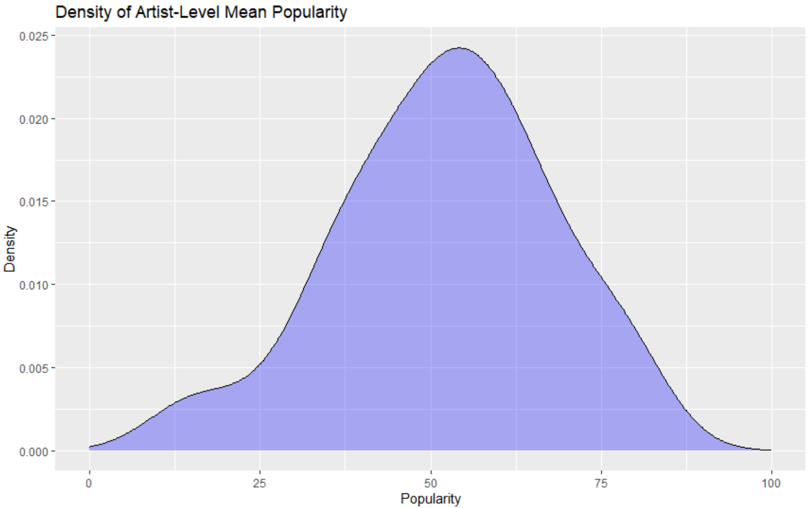

Figure 5: A density plot of the variability in mean song popularity from artist to artist.

Priors on the Global Parameters

- To complete the Bayesian model, priors must be specified for the global parameters μ, σ_y, and σ_μ.
- Following textbook recommendations:
  - **Prior for μ:** $\mu \sim N(50, 52^2)$.
    - The mean of 50 reflects plausible popularity values.
    - The large variance indicates prior uncertainty.
  - **Prior for σ_y:** $\sigma_y \sim \text{Exp}(0.048)$.
    - This choice captures uncertainty about the within-group variability.
    - Other distributions on $(0, \infty)$, such as Gamma or Inverse-Gamma, could also be used.
  - **Prior for σ_μ:** $\sigma_\mu \sim \text{Exp}(1)$.
    Reflects uncertainty in between-group variability.

Analysis of Variance (ANOVA)

- This hierarchical model represents a **Bayesian version** of the classical One-Way Analysis of Variance (ANOVA) model.
- **Objective:** Compare the means of multiple groups by analyzing the relationship between:
  - **Within-group variability** (σ_y^2).
  - **Between-group variability** (σ_μ^2).
- In this example:
  - Groups are defined by the artists.
  - The goal is to estimate the 44 artist-level means $\mu_1, \ldots, \mu_{44}$.
- **Variance decomposition:**

$$\text{Var}(Y_{ij}) = \sigma_y^2 + \sigma_\mu^2,$$

  - σ_y^2: Within-group variance (popularity variability for songs by the same artist).
  - σ_μ^2: Between-group variance (variability in mean popularity across artists).

Proportion of Variance Explained

- The proportion of total variance in Y_{ij} explained by within-group and between-group differences is given by:
  - **Within-group variance:** $\frac{\sigma_y^2}{\sigma_\mu^2 + \sigma_y^2}$: Proportion of $\text{Var}(Y_{ij})$ explained by differences within each group (artist).
  - **Between-group variance:** $\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_y^2}$: Proportion of $\text{Var}(Y_{ij})$ explained by differences between groups (artists).
- The term $\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_y^2}$ also measures the **within-group correlation**, such as the correlation between the popularity of songs by the same artist.
- **Model implication:** The model forces this correlation to be positive, which is reasonable for most real-world scenarios.

Fitting the Bayesian Model

- **Posterior analysis** is performed using the `stan_glmer` function.
- **Formula syntax:**
  - Unlike `stan_glm`, the grouping variable (`artist`) is specified using:

$$\texttt{popularity} \sim (1 \mid \texttt{artist}).$$

  - This accounts for the hierarchical structure of the data.
- **Model fit assessment:**
  - The `pp_check` function compares the posterior predictive density with the observed data density.
  - This diagnostic tool helps evaluate the adequacy of the model fit (refer to `R` code).
- **Fit quality:** The Normal hierarchical model provides a reasonable fit to the data, though not perfect (see next slide).

Density Plots of Simulated and Observed Popularity

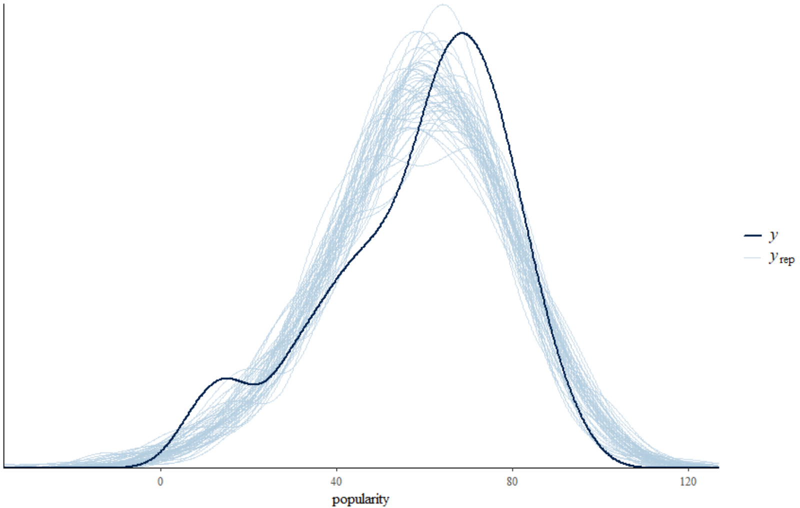

Figure 6: 100 posterior simulated datasets of song popularity (light blue) along with the actual observed popularity data (dark blue).

Outline

Posterior Inference about Model Parameters

- **Posterior inference** for global parameters, such as point estimates and credible intervals, can be computed easily in `R`.
- **Example results:**
  - Posterior point estimate for μ: $\widehat{\mu} = 52.5$.
  - 80% credible interval for μ: $(49.3, 55.7)$.
  - Posterior estimates for standard deviations:
    $\widehat{\sigma}_\mu = 15.1, \quad \widehat{\sigma}_y = 14.0.$
- **Estimated within-group correlation:**

$$\frac{\widehat{\sigma}_\mu^2}{\widehat{\sigma}_\mu^2 + \widehat{\sigma}_y^2} = \frac{15.1^2}{15.1^2 + 14.0^2} = 0.54.$$

- **Interpretation:** This indicates a moderate positive linear association in popularity values for songs from the same artist.

Posterior Inference about Group-Specific Parameters

- Posterior inference for group-specific parameters, μ_j (e.g., artist-level mean popularity), includes point and interval estimates.
- **Example results:**
  - For Beyoncé:
    Point estimate: $\widehat{\mu}_{\text{Beyoncé}} = 69.1$,
    80% credible interval: $(65.6, 72.7)$.
  - For Vampire Weekend:
    Point estimate: $\widehat{\mu}_{\text{Vampire Weekend}} = 61.6$,
    80% credible interval: $(54.8, 68.5)$.
- **Observation:**
  - Credible interval widths vary across artists (see next slide).
  - Artists with smaller sample sizes have wider credible intervals, reflecting greater uncertainty (e.g., Frank Ocean vs. Lil Skies).

Posterior Credible Intervals for Popularity

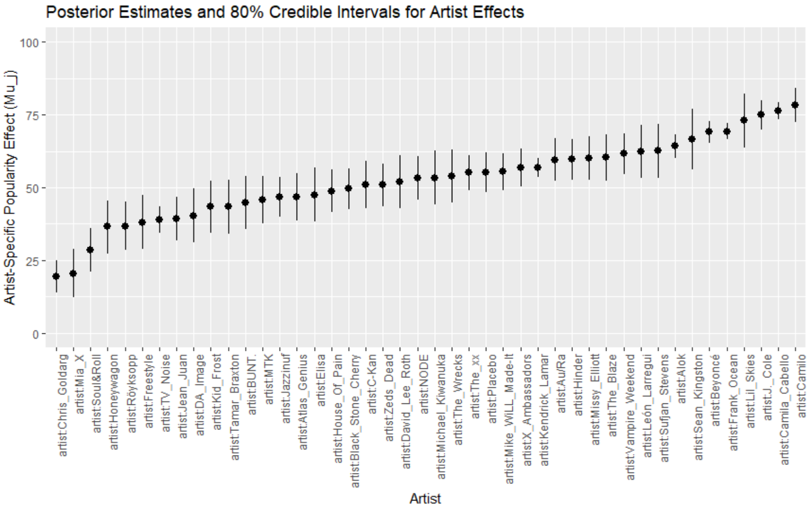

Figure 7: 80% posterior credible intervals for each artist's mean song popularity.

Posterior Prediction for an Artist in the Sample

- To predict the popularity of a new song by an artist in the sample (e.g., Vampire Weekend):
  - An 80% prediction interval for the popularity of a new song is:

$$(42.5, 80.8).$$

- **Key observation:**
  - The prediction interval is significantly wider than the 80% credible interval for Vampire Weekend's mean popularity, μ_j.
- **Why?**
  - The credible interval reflects uncertainty in the mean popularity μ_j, averaged across all songs.
  - The prediction interval accounts for the additional variability in individual song popularity within the group, making it naturally wider.
- **Conclusion:** It is logical that we can estimate an artist's mean popularity with more precision than the popularity of a single song.

Posterior Prediction for an Artist Not in the Sample

- Predicting the popularity of a new song by an artist not in the sample (e.g., Taylor Swift) is possible with the hierarchical model:
  - Recall: The no-pooling model could not accommodate this scenario.
  - The hierarchical model leverages information about the broader population to make predictions.
- **Steps in the prediction process:**
  1. Simulate values for μ_j (Taylor Swift's mean popularity) from: $\mu_j \sim N(\mu, \sigma_\mu^2)$, while allowing μ and σ_μ to vary according to their posterior distributions.
  2. Simulate song popularity values, Y, from: $Y \sim N(\mu_j, \sigma_y^2)$, while varying σ_y according to its posterior distribution.
- **Result:**
  - An 80% prediction interval for Taylor Swift's new song popularity: $(25.9, 78.9)$.

Is This Prediction Accurate?

- **Real-world applicability:**
  - Do we truly believe the prediction interval for Taylor Swift's new song popularity? **Probably not.**
  - If the "new artist" were someone with no prior fame, the interval might be reasonable.
  - However, Taylor Swift is one of the most globally recognized and successful artists, so her song's popularity would likely fall in the higher range.
- **Improving the model:**
  - To better capture Taylor's exceptional status, a more realistic model could include artist-level covariates, such as:
    - Number of past Grammy nominations.
    - Historical radio airplay or streaming metrics.
  - Including such predictors could refine the model's predictions for artists with unique characteristics.
- **Next steps:** Chapter 17 explores hierarchical models augmented with predictor variables, offering a more nuanced approach to modeling.

Outline

Shrinkage

- We visualize predictions for new song popularities for all 44 artists (see next slide):
  - Light blue: Point and interval predictions from the hierarchical model; Dark blue: Sample mean popularity for each artist.
- **Observation:**
  - The plot demonstrates the phenomenon of **shrinkage**.
  - Hierarchical model predictions shrink (or pull) the artist-specific sample means toward the global sample mean.
- **Model comparison:**
  - **Complete-pooling model:** Predicts song popularity using the global mean.
  - **No-pooling model:** Predicts song popularity using the artist's own mean.
  - **Hierarchical model:** Balances these extremes, combining global and group-specific information.
- Shrinkage reflects the hierarchical model's ability to pool information across artists while respecting individual group differences.

Posterior Credible Intervals for Popularity + Observed Mean Popularity

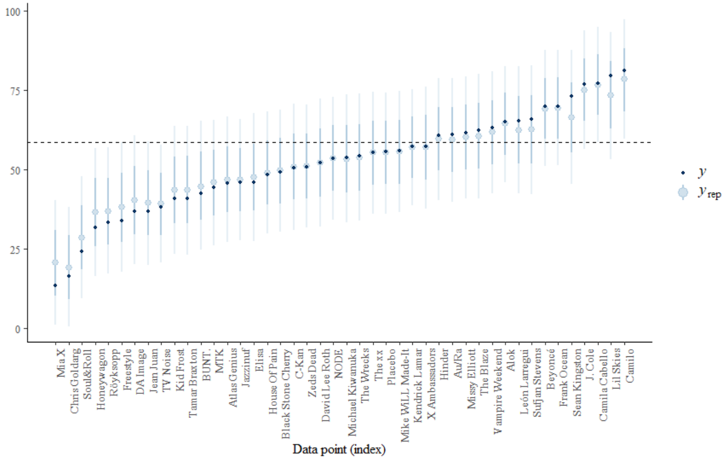

Figure 8: Posterior predictive intervals for artist song popularity, as calculated from a hierarchical model. The horizontal dashed line represents the average popularity across all songs.

How Much Shrinkage?

- **Key observation:**
  - Artists with the **smallest sample sizes** experience the most shrinkage toward the global mean.
  - These artists also have the **widest credible intervals** for their μ_j estimates, reflecting greater uncertainty.
- **Rationale:**
  - With less data for an artist, the model borrows information from other artists in the population to improve predictions.
  - For artists with large sample sizes (e.g., Frank Ocean), the model relies more on their own data, reducing shrinkage.

How Much Shrinkage?

- **Free throw analogy:**
  - Consider two basketball players:
    - Player A: Made 98 out of 100 free throws.
    - Player B: Made 3 out of 3 free throws.
  - Which player would you predict has a higher probability of making their next free throw?
  - Intuitively, Player A's estimate is more reliable due to the larger sample size, demonstrating the concept of shrinkage in practice.

Grouping Variable or Predictor?

- **Why treat "artist" as a grouping variable instead of a categorical predictor?**
  - If all levels of the variable in the sample are the only levels of interest, it should be included as a **predictor**.
  - Example: In a Poisson model for academic awards, the variable "track" (`academic`, `vocational`, `general`) represented all possible levels and was treated as a predictor.
- **Spotify example:**
  - The artists in the dataset are a **random sample** from a larger population of artists.
  - Treating "artist" as a grouping variable allows the model to generalize to the entire population of artists, including those not in the sample.
- **Key distinction:**
  - This aligns with the classical distinction between:
    - **Fixed effects:** Used when all levels of the variable are of interest; **Random effects:** Used when levels represent a random sample from a larger population.