# CS146 - Computational Methods for Bayesian Statistics

Project 2
Minerva University
Fall 2025
Scheffler, MWh@3PM UTC

November 22, 2025

## Contents

# 1. Summary of findings

> “Which factors most influence attendance — team popularity or day of the week? And can we reliably predict attendance for games where data is missing?”

Below, you can see the starting point of this analysis: data on the attendance per team and day of the week. One need not be a statistician to see that the attendance trends vary a lot depending on the day of the week and on the team that is playing.

## 1.1. Which factors most influence attendance?

Notice how the average attendance (red dotted line) for the teams on the right (Godoy Cruz, Independiente, Argentinos Juniors) is significantly lower than that of the teams on the left, such as Velez Sarsfield and Racing. Similarly, games taking place on Saturday have the highest average attendance compared to those on Thursday, which have the lowest. Most importantly, notice how the *variation* of the average attendance lot larger among the teams than among the days. My analysis supports this finding, which answers the first part of your question: **Team Popularity** is the dominant influence on attendance. While organizing a game on Saturday instead of Thursday will likely result in slight increase in attendance, it still will not compare to the increase in attendance from a low-tier to a top-tier game. The model I created to predict the missing value also makes use of this property, as explained below.

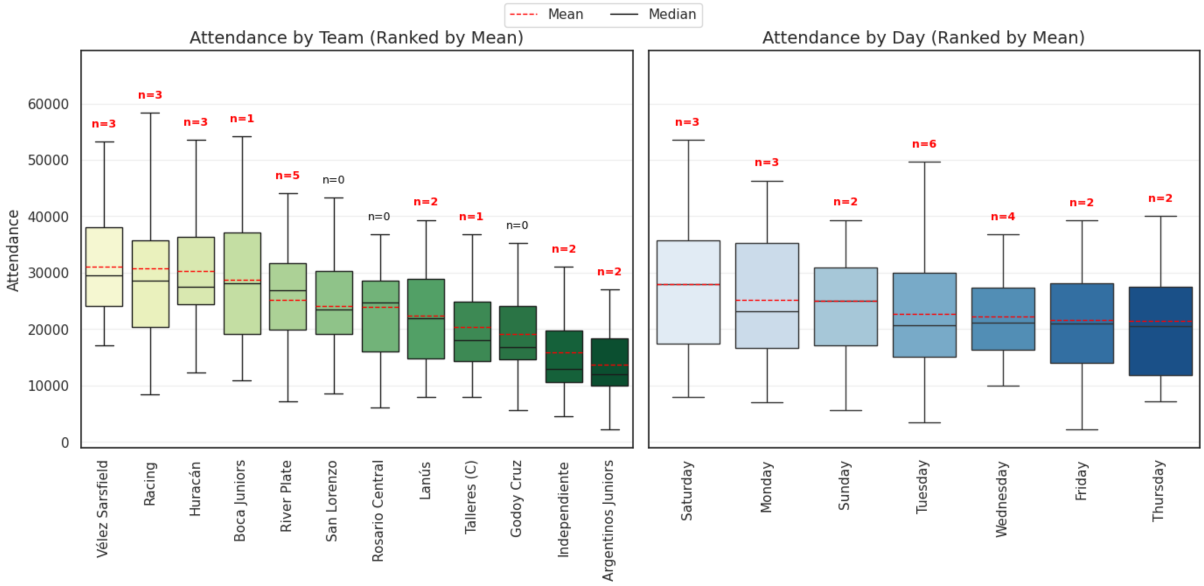

**Figure 1:** Attendance varying by (left) team and (right) weekday. Teams and days were sorted by descending average attendance values. The number of missing datapoints was plotted above each of the boxplots.

## 1.2. Can we reliably predict the attendance for missing data?

These are two separate questions. Firstly, yes, we can predict missing attendance data by using our knowledge of how the attendance records are distributed over the days and teams. While this provides us with nice average estimates for each of the missing data points, we should be aware that if we want to be very certain about the location of that point, we do not look at a point-estimate but rather an interval of plausible values. This is the second question about ”reliability” that is a little harder to answer. For instance, if we want to be 95% sure about the attendance of the River Plate Game on Thursday, we are looking at a very large range of 7,920-54,130. However, if we only look at the predicted average attendance, we obtain an easier-to-read but less precise value of 24,616 attendees.

When we plot the newly predicted points onto the known plot from above, we can see again that my model assigns larger importance to the team that is playing than the day the game is happening on. For instance, if you look at point 16 (Argentinos Juniors game on a Saturday), we see that it is rather significantly lower than the average attendance of a game on Saturday. However, this is because *Argentinos Juniors* are playing, which is the team that attracts the lowest numbers of fans to their games (see the left plot). At the same time, on the team plot, point 16 represents a 'popular' game given that *Argentinos Juniors* are playing where the attendance is larger than the average. This is because it is happening on a *Saturday*, which attracts more fans than usual. Find the list of calculated values at the end of this paper. Finally, a recommendation is also to specifically improve the ticketing system at those games where we currently see the most missing data: River Plate ($n = 5$), Velez Sarsfield ($n = 3$), Racing ($n = 3$), and Huracan ($n = 3$). While I was able to make some predictions, we will always have more certainty about the outcome if the ticket data is collected properly. For a more detailed analysis of the methods used to predict these values, read the following sections.

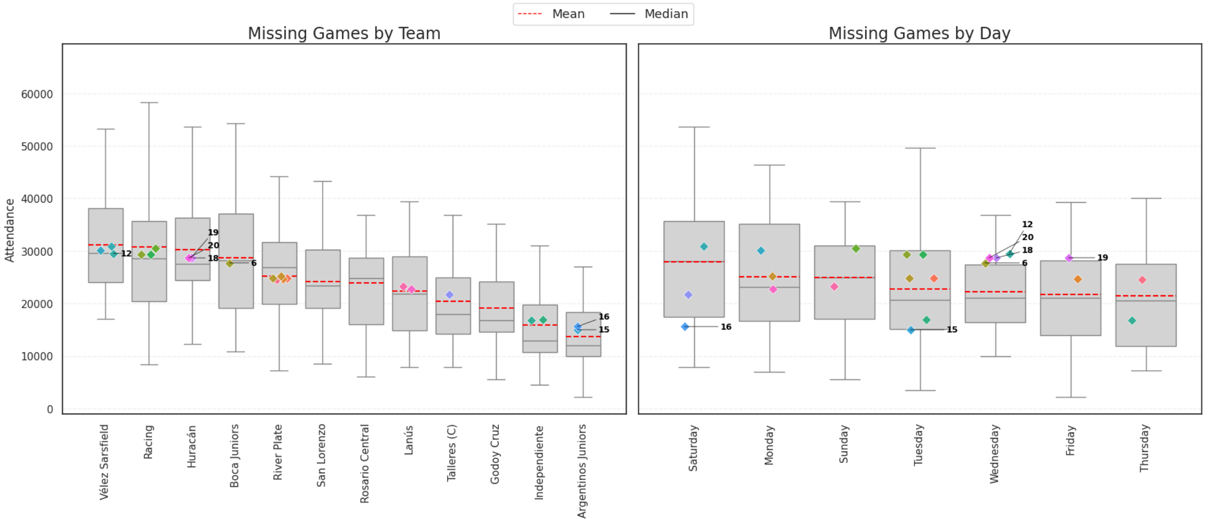

**Figure 2:** Predicted attendance of the missing games plotted on the day of the event and the competing team. Points that fall outside of the middle 50% of the day-boxplot were numbered.

# 2. Model 1

First, we create a 'complete pooling' model whereby we assume that all attendance measurements come from a single distribution. This way, we only estimate the parameters of a single distribution and assume that this single estimate is applied to all the teams and days of the week. This is likely an under-fit of the data as we assume zero variation among teams and days of the week.

## 2.1. Setting up Model 1

As we are modeling the number of attendees to football games, we are dealing with count data. However, as we do not know the upper bound, like in a fixed number trials, a Binomial model is unsuitable. So we might consider using a Poisson model. As there is no reason to assume that the mean and the variance are the same here, a Gamma-Poisson (Negative Binomial) model might be the best choice here as it has a parameter $\alpha$ that models the dispersion of the data.

$$\begin{aligned} y_i &\sim \text{Gamma-Poisson}(\mu, \alpha) && \text{(Likelihood)} \\ \mu &\sim \text{Half-Normal}(\sigma = 20000) && \text{(Prior for population mean)} \\ \alpha &\sim \text{Exponential}(\lambda = 10) && \text{(Prior for population variance)} \end{aligned} \tag{1}$$

As $\mu$ models the mean number of attendees at a football game, this value must be positive. So the support of this prior should be restricted to positive values only. If the Half-Normal distribution is given a large value for its standard deviation to represent the order of magnitude of the attendance observations ranging in between approximately 10,000-60,000 (plausible stadium capacities), it will likely be a plausible and not too informative prior. This is acceptable as the large sample size of $n = 240$ will likely overwhelm the prior.

Secondly, the prior for the variance parameter $\alpha$ must also be restricted to positive values. Here, we can use a simple exponential prior with $\lambda = 1$. The variance of a negative binomial distribution is $\sigma^2 = \mu + \frac{\mu^2}{\alpha}$. Similarly, the mean of the exponential prior is $\frac{1}{\lambda} = 1$. This way, we allow the model to adapt to different levels of dispersion. If $\alpha$ becomes very large, $\sigma^2 \approx \mu$ as $\frac{\mu^2}{\alpha} \to 0$, which would make it approach a Poisson distribution. However, the model can adapt and reduce the value of $\alpha$ to model the larger dispersion.

To check whether the chosen model produces plausible data, we can take samples from the likelihood function and show the distribution of possible data values using only the priors (Figure 3). While many of the sample distributions over the attendance have peaks centered at low attendance values, they are all plausible in that they do not have negative values or sudden peaks at high values of the attendance that would fall out of the plausible range: for instance, if there was a peak at around 3,000,000, we would consider that sample implausible as no stadium can hold a capacity that large - the "Rungrado 1st of May Stadium" with 114.000 seats currently holds the highest-capacity record [source]. Having established that the priors are plausible, we can use the observed data to obtain the posterior.

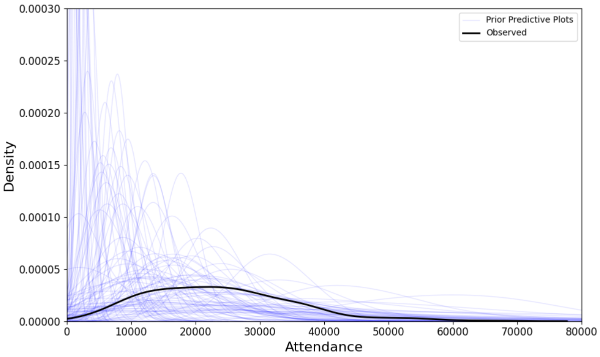

**Figure 3:** Prior Predictive Plot with 100 samples for the complete pooling Model 1. Note that the y-range was restricted to [0,0.0003] in order to show the observed data more clearly.

## 2.2. Model 1 Posterior

To confirm whether we obtained quality samples from the posterior distribution, we check if the samplers converged. All the rounded $\hat{R}$-values are 1.0, which is below the conventional threshold of $< 1.01$. This suggests that there is hardly any variance in between the chains, which again implies that all chains explored the entire posterior space well. Additionally, the effective sample size (ESS) is larger than $3,800$ for all the parameters, which is a strong value showing little autocorrelation and a useful acceptance rate as well as a healthy exploration of the posterior space. The sampler does not get stuck in single locations but instead properly explores the posterior space. We obtain a posterior mean value of approximately 23699 (Table 1), which in this model represents the global average of the attendance. At the same time, the posterior mean for $\alpha \approx 4$, which is a small value that represents high levels of dispersion. This means that while the average is at 23699, the model also considers large fluctuations plausible. This is not surprising considering the distribution of the data, which varies from values as low as a couple thousands to values up to 60,000. Lastly, we want to check the model fit after incorporating the data. It is notable that some of the posterior samples produce larger density values for the peak at around 20,000 than the original model. However, overall, the posterior-predictive plot in Figure 5 shows that the samples from the posterior (plotted in blue) closely align with the observed data (black), which tells us that this model can do a good job at modeling the data as long as we do not split it up along its categories: we now want to know how well this model predicts the attendance for each day and team, as this is our primary goal. It is no surprise that the posterior predictive distributions look the same for each team, as this is how we set up the model (Figure 6). Variation in the attendance from one team to the next is assumed to be 0, so the posterior predictive outputs the same datasets for each team. While this mostly models the data

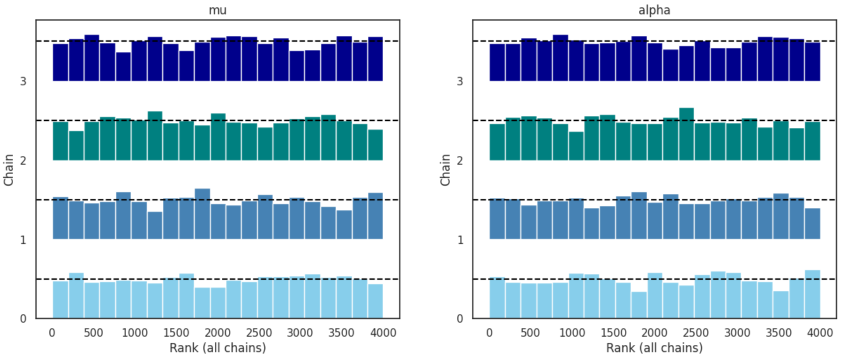

**Figure 4:** Rank Plot for the posterior sampling process for Model 1. Ranks are approximately uniformly distributed for each of the chain, suggesting that all chains explored the posterior space well.

| | mean | sd | hdi_3% | hdi_97% | mcse_mean | mcse_sd | ess_bulk | ess_tail | r_hat |
|---|---|---|---|---|---|---|---|---|---|
| mu | **23698.896** | 795.629 | 22267.556 | 25270.736 | 12.846 | 12.782 | 3874.0 | 2718.0 | 1.0 |
| alpha | **3.938** | 0.361 | 3.326 | 4.682 | 0.006 | 0.005 | 3851.0 | 2894.0 | 1.0 |

**Table 1:** Posterior Value Table for Model 1. The ESS and $\hat{R}$ values suggest sampler convergence.

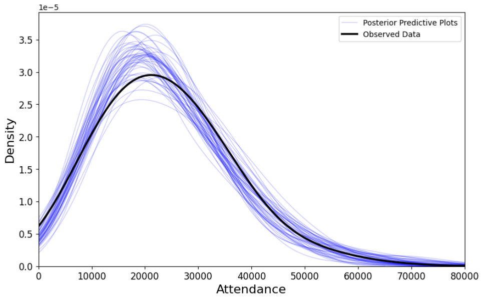

**Figure 5:** Posterior Predictive Plot with 100 samples for Model 1. Samples are plotted in blue and show close alignment with the observed data.

well, the different attendance distribution in some teams, like Argentinos Juniors, or days, like Saturday, are not captured very well.

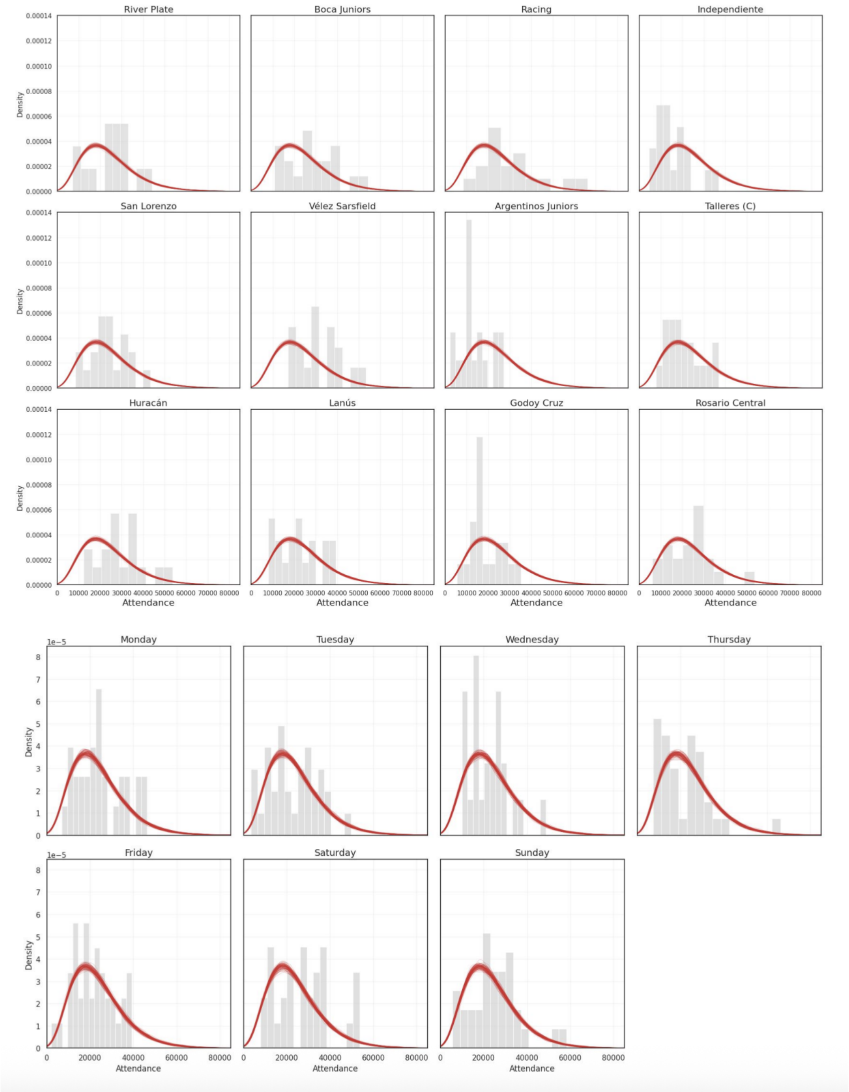

**Figure 6:** Posterior Predictive Plot of the complete pooling model for each weekday and team with 100 plots each. The posterior predictive plots are identical, which means that the model fits differently well depending on the day and the team.

# 3. Model 2

So far, we have ignored the possibility that attendance might vary between teams and days of the week. This is not entirely sound, as it is plausible to think that attendance depends on the playing team: for instance, attendance would surely vary between a game of Bayern München, Germany's 'hyped' football champion with 75,024 seats in their stadium, or SC Paderborn, the Bundesliga's newcomer with 15,000 seats and a smaller fan-base. The same counts for different weekdays, as games on Mondays are likely less attended (where people are likely busier with work) than on Saturdays (where fans might be more inclined to go and watch a game). We use a hierarchical model to include the idea that there is variance between each of these two groups of pools (seven days of the week and twelve football teams). Now, each observation of the attendance includes information about the day of the week on which it was observed, as well as the team that played on that day.

## 3.1. Setting up Model 2

For this model, we use the same Gamma-Poisson likelihood function to model the count data of attendees at a game. However, this time, we use adaptive priors for the random effects of different days ($\beta$) and different teams ($\gamma$), so the model can determine the variation in attendance within each of their groups. To model the expected attendance $\mu$, a link function was used to map the values for the random effects of the given team and day to a positive parameter space. We also introduce an additional parameter $b$ to represent the expected attendance for a 'baseline' log-game-attendance (when $\beta = \gamma = 0$). An exponential distribution with $\lambda = 10$ was chosen to better represent the order of magnitude of this estimate, with $e^{10} = 22026$, which would be a plausible baseline value for the attendance given that stadiums usually have a capacity of 10,000-60,000. This creates a mapping to the positive space as we can exponentiate the sum as follows:

$$\log(\mu_k) = b + \beta_{\text{TEAM}[k]} + \gamma_{\text{DAY}[k]} \rightarrow \mu_k = \exp(b + \beta_{\text{TEAM}[k]} + \gamma_{\text{DAY}[k]}) \tag{2}$$

This exponential expression will guarantee that all $\mu_k > 0$.

Next, the parameters $\beta$ (team random effects) and $\gamma$ (weekday random effects) were modeled as the deviation from the baseline $b$ by modeling them as Normal distributions with mean 0 and standard deviation $\sigma_t$ and $\sigma_d$, respectively. This way, we allow the random effects to be different for each day of the week, and for each of the 12 teams. As the standard deviations of the priors have to be positive, we model the hyperpriors using an exponential distribution with rate parameter $\lambda = 1$. This is a weakly informative prior that allows partial pooling to happen and 'shrink' estimates for teams with low amounts of data towards the mean.

Lastly, we also use a parameter $\alpha$ to model the overall dispersion of the model, which represents the variation of the attendance caused by unmeasured factors (e.g. whether the day is a public holiday, whether it snows that day, which opponent the team is playing...).

$$\begin{aligned}
y_k &\sim \text{Gamma-Poisson}(\mu_k, \alpha) && \text{(Likelihood for game } k\text{)} \\
\log(\mu_k) &= b + \beta_{\text{TEAM}[k]} + \gamma_{\text{DAY}[k]} && \text{(Log-link Linear Model)} \\
b &\sim \text{Normal}(10, 1) && \text{(Prior for Global Intercept)} \\
\beta_j &\sim \text{Normal}(0, \sigma_t) && \text{(Prior for Team Random Effects for } j = 1..12\text{)} \\
\gamma_m &\sim \text{Normal}(0, \sigma_d) && \text{(Prior for Weekday Random Effects for } m = 1..7\text{)} \\
\sigma_t &\sim \text{Exponential}(1) && \text{(Hyperprior for Team Standard Deviation)} \\
\sigma_d &\sim \text{Exponential}(1) && \text{(Hyperprior for Weekday Standard Deviation)} \\
\alpha &\sim \text{Exponential}(0.01) && \text{(Prior for Dispersion Parameter)}
\end{aligned} \tag{3}$$

Just like for Model 1, we first perform a plausibility check of the model setup and plot 100 samples from the prior predictive distribution (Figure 7). These look very similar to the previous model whereby the distributions fall within a plausible range of attendance values. There are no negative or implausible extreme values, so we can now use the data to get posterior samples.

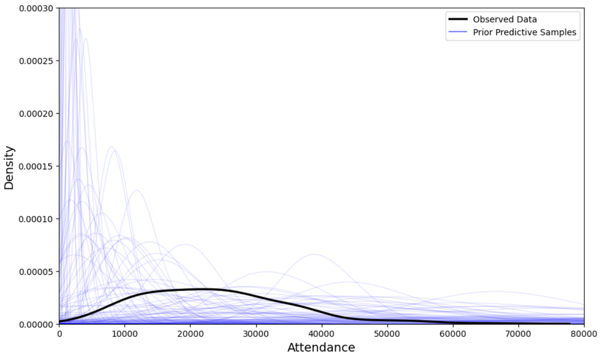

**Figure 7:** Prior Predictive Plot with 100 samples for the hierarchical Model 2. Again, the y-range was restricted to [0,0.0003] in order to show the observed data more clearly.

## 3.2. Sampling Problems & Reparameterization

While sampling from the posterior, we still get a few divergences (16 in total), which occurs when the posterior space has sharp peaks that do not allow PyMC's Nu-U-Turn Sampler to effectively explore it well with a single step size. This is likely because the scale of the adaptive priors depends on their hyperpriors. For instance, if $\sigma_t$ is very small, say $10^{-5}$, then the value of $\beta_j$ is forced to be extremely close to its mean of 0, while it can be spread much further away for larger values of $\sigma_t$. This creates shapes similar to the Devil's Funnel, which makes areas extremely difficult to explore. To solve this sampling problem, we can transform the sampling space in a way that avoids these extreme peaks. For instance, instead of sampling directly from $\text{Normal}(0, \sigma_t)$, we can sample from a well-behaved standard normal $\text{Normal}(0, 1)$ and rescale by the value of $\sigma_t$ (if the mean was different from zero, it would also be added). This is a computational modification in the code (snippet below) that does not change the definition of the model.

$$\beta_{REP_j} \sim \text{Normal}(0, 1) \tag{4}$$

$$\beta_j = \beta_{REP_j}\, \sigma_t \tag{5}$$

and

$$\gamma_{REP_m} \sim \text{Normal}(0, 1) \tag{6}$$

$$\gamma_m = \gamma_{REP_m}\, \sigma_d \tag{7}$$

```
beta_rep = pm.Normal('beta_rep', mu=0, sigma=1, shape=data_enc['team'].nunique())
beta = pm.Deterministic('beta', beta_rep * sigma_team) # adaptive prior 1

gamma_rep = pm.Normal('gamma_rep', mu=0, sigma=1, shape=data_enc['day'].nunique())
gamma = pm.Deterministic('gamma', gamma_rep * sigma_day) # adaptive prior 2
```

We can now move on to taking samples from the reparameterized posterior distribution. To check sampler convergence, rank plots were created for all parameters (which can be found in the code appendix because of the large number of parameters). All of them show uniformly distributed ranks across all chains and parameters. As suggested by the strong values of $\hat{R} = 1.0$ as well as the very large effective sample size of around 2,000 for all the parameters, the chains seem to explore the posterior space and mix with each other well.

## 3.3. Posterior

Now that we have computed the posterior values of each of the $\gamma$ and $\beta$ parameters, we can see the variance between the days of the week ($\gamma$-values) and the different teams ($\beta$-values). Figure 8 shows how the expected parameter values (plotted as points) among the teams is a lot larger than the variance among the days of the week. This means that we would expect the presence of a specific team to have larger influence on the final predicted attendance value, which also means a *smaller* shrinkage effect on the team than on the day of the week on which the game takes place. This is because the attendance-variance among the teams is so large that they The right plot also supports this idea as the variance among the team parameters is a lot more spread out in contrast to that of the weekdays, which has a sharp peak at 0.

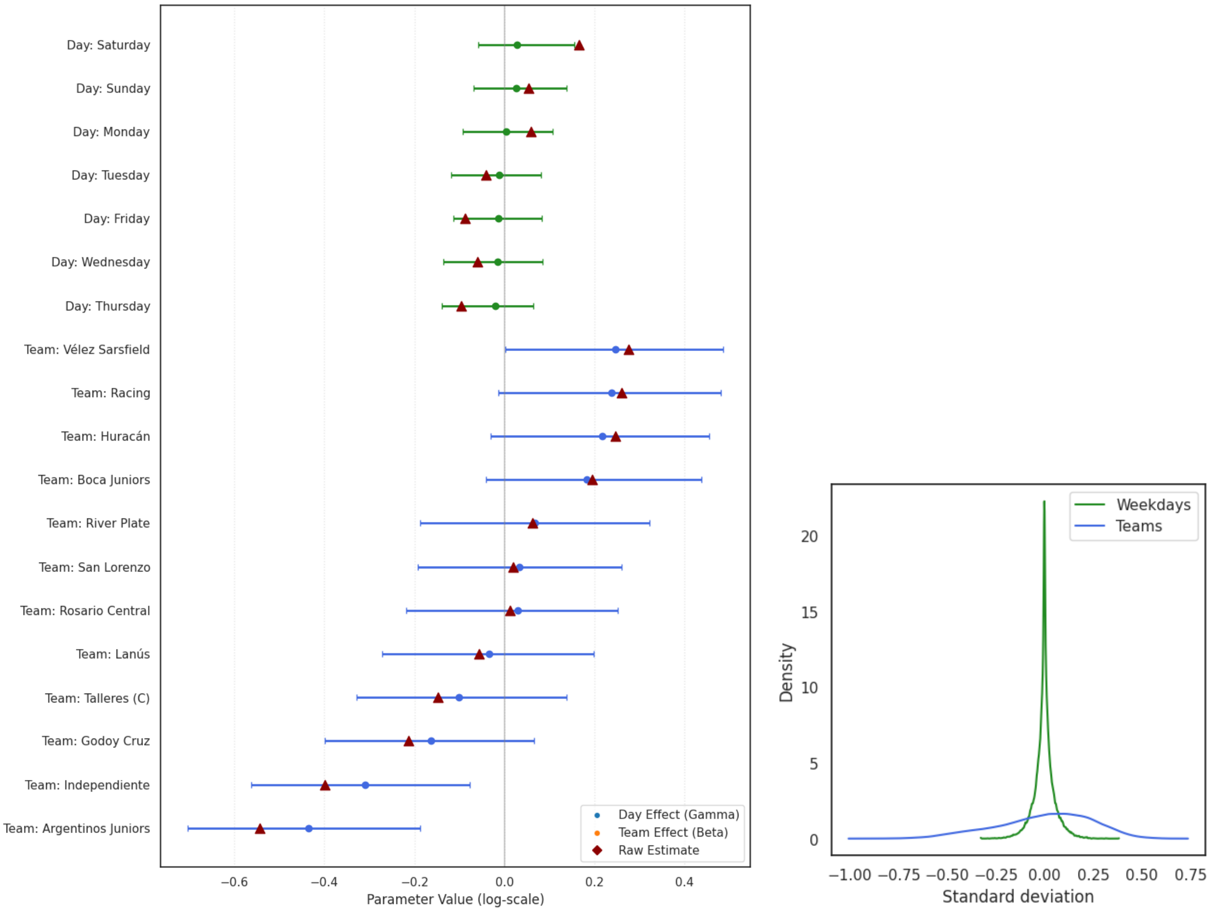

**Figure 8:** Left: Posterior distribution of the $\gamma$ and $\beta$ parameters for each cluster. The days and teams are ranked based on their posterior values. The observed average was plotted in red to observe pooling shrinkage. Right: Distribution of standard deviations of the $\gamma$ and $\beta$ values.

On the plot, the 95% HDI plots for the variance of the $\gamma$-parameters all overlap with 0, indicating that we cannot be certain about the significance of this variation at this confidence level. For the HDI plots of the $\beta$-values, the intervals of $\gamma_3$ (Team ”Independiente”) and $\gamma_6$ (Team ”Argentinos Juniors”) now do not overlap with the line anymore, suggesting a strong and significant variation of these groups away from the mean. This tells us that the partial pooling was likely a wise decision as the groups do indeed seem to vary.

We also obtain posterior values for $b$ (intercept) and $\alpha$. This tells us that the mean baseline attendance is $e^{\bar{\beta}} = e^{10.055} \approx 23272$ and the mean ’noise’ parameter is $\bar{\alpha} = 5.006$ which implies that the Gamma-Poisson was a good choice compared to the Poisson model as the model correctly models the overdispersion. However, a fairer comparison would be to compare the posterior predictive plot for each day and team.

| | mean | sd | hdi_3% | hdi_97% | mcse_mean | mcse_sd | ess_bulk | ess_tail |
|---|---|---|---|---|---|---|---|---|
| intercept | 10.055 | 0.087 | 9.893 | 10.220 | 0.002 | 0.002 | 1392.0 | 1762.0 |
| alpha | 5.006 | 0.474 | 4.152 | 5.870 | 0.006 | 0.008 | 5526.0 | 3060.0 |

**Table 2:** Posterior values for $b$ and $\alpha$.

This time, we can clearly see that the posterior predictive plots are different for each team or day which is a ’deliberate’ advantage of this model over the complete pooling one. At the same time, we also see how the shape of the posterior predictive plots varies more depending on the team than on the day of the week.

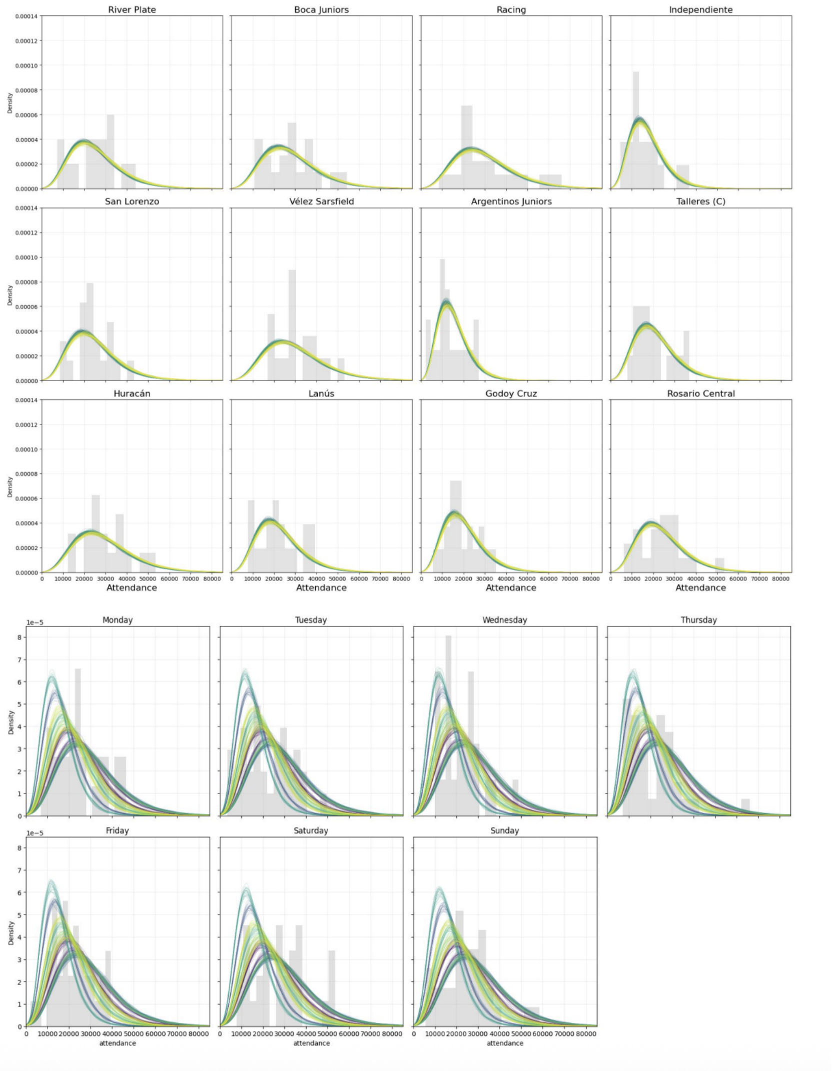

**Figure 9:** Posterior predictive plots for each team and day with 175 posterior predictive samples on each plot. The different colors represent the samples from different days for the top plot, and different teams on the bottom plot.

# 4. Model Comparison

To determine which of the models has the highest predictive out-of-sample accuracy, we need to compute the expected log-pointwise predictive density (ELPD). Ultimately, this aims to show by how much adding out-of-sample data would affect the deviance of the model. If out-of-sample data introduces a lot of deviance, our model is likely overfitting to the data. Eventually, this shows how well the model can make predictions for new, unseen data–which is our goal for the missing game data here. One method for this is computing the log-pointwise-predictive density: this is done by considering the log-likelihood of each data point $p(y_i|\theta)$ over *all* posterior draws of the parameters. Finally, the log-estimates for each data point are averaged to obtain the LPPD. However, this is often a poor estimate for out-of sample data as it is computed using only the known data. Using LOO-CV (Leave-One-Out Cross Validation) is a better approach as it simulates out-of-sample predictions, which is also used in this assignment.

Pareto-Smoothed Importance Sampling (PSIS) was preferred for several advantages. Firstly, it adjusts the weights of highly unlikely observations that are far away from the 'bulk' of the other observations, leading to more stable predictions. Additionally, it provides a warning when the parameter $k$ of the Pareto distribution exceeds a certain threshold (usually 0.7), implying that importance sampling did not work reliably. None of these warnings occurred in this model comparison. Using PSIS-LOO, we can now compare the ELPD for the two models, where higher values show a larger predictive accuracy. The resulting ELPD values, as well as their difference, are plotted below to get an intuition about how these models compare. It is not surprising that the partial pooling model performs better in this comparison

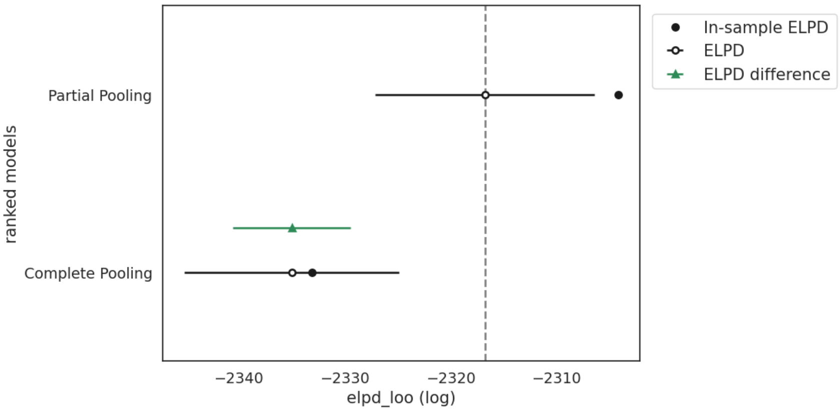

**Figure 10:** Model Comparison using PSIS-LOO

| | rank | elpd_loo | p_loo | elpd_diff | weight | se | dse | warning |
|---|---|---|---|---|---|---|---|---|
| Partial Pooling | 0 | -2316.825409 | 12.567463 | 0.000000 | 1.0 | 10.320541 | 0.000000 | False |
| Complete Pooling | 1 | -2335.049570 | 1.887400 | 18.224161 | 0.0 | 10.144371 | 5.582008 | False |

**Table 3:** LOO comparison

as it deliberately uses the variability between days of the week, and football teams, to estimate the attendance for a given game. We already saw the improvement in the posterior predctive plots above. Using the standard error of the difference (‘dse‘), we can also conclude that the partial model is indeed better: for a 99% estimate, which corresponds to a $z$-score of 2.576, the difference in ELPD does not cross the ’no-difference’ line:

$$18.224161 - 2.576 \cdot 5.582008 \approx 3.84 > 0. \tag{8}$$

Hence, at this confidence level, it is not possible that the ELPD of the complete pooling model is as high as the ELPD for the partial pooling model.

# 5. Prediction of the missing points

Given that the model comparison provides strong evidence in favor of the partial pooling model, we will use it instead of the complete pooling model to predict the missing values. There are 22 missing data points in total. As we are making **within-cluster** predictions, we can use the posterior values for the intercept $b$, the team parameter $\beta_j$ and the weekday parameter $\gamma_m$ to create a distribution over the possible predicted values.

Therefore, we can use the values sampled in the 4 chains to calculate the mean and a density interval for the predicted attendees. For that, we use

$$(b^{(i)}, \alpha^{(i)}, \beta_t^{(i)}, \gamma_m^{(i)}) \text{ for } i = 1, ..., 4000 \tag{9}$$

and draw from the Gamma-Poisson likelihood for each set of values. This gives us a distribution over the predicted attendance values for each missing game (Table 4). Refer to the ”Summary of Findings” for visualizations that show and explain the plausibility of these predicted values.

**Word Count:** 2,887

| Team | Day | Mean Attendance | 95% Interval |
|---|---|---|---|
| River Plate | Thursday | 24616 | 7920–54130 |
| River Plate | Tuesday | 24732 | 7404–52054 |
| River Plate | Friday | 24881 | 7973–51837 |
| River Plate | Monday | 25091 | 8036–52702 |
| River Plate | Tuesday | 25038 | 7414–52098 |
| Boca Juniors | Wednesday | 27952 | 9127–58429 |
| Racing | Tuesday | 29245 | 9315–61571 |
| Racing | Sunday | 30489 | 9735–63175 |
| Racing | Tuesday | 29660 | 9653–63720 |
| Independiente | Tuesday | 17034 | 5451–35622 |
| Independiente | Thursday | 17101 | 5366–36278 |
| Vélez Sarsfield | Wednesday | 29594 | 9051–62453 |
| Vélez Sarsfield | Saturday | 30579 | 9637–64956 |
| Vélez Sarsfield | Monday | 30012 | 9446–62099 |
| Argentinos Juniors | Tuesday | 15105 | 4714–31305 |
| Argentinos Juniors | Saturday | 15732 | 4837–32968 |
| Talleres (C) | Saturday | 21591 | 6674–46523 |
| Huracán | Wednesday | 28312 | 9101–59198 |
| Huracán | Friday | 28783 | 9176–60780 |
| Huracán | Wednesday | 28710 | 8826–60295 |

**Table 4:** Predicted mean attendance and 95% intervals for missing games.

## 6. Acknowledgments

This is a good place to acknowledge Richard McElreath and his great book that we use as a class resource. It has been great to work with his variety of visualizations, as well as Prof Scheffler's translations into PyMC.

Secondly, I would like to praise the "Very Normal" channel on YouTube which has provided additional explanations along this assignment process.

**Using Class Code for this Assignment**:

Class Code was taken and adapted from multiple sessions. I put effort into improving visualizations and clarity rather than just copy-pasting code.

- Code from Session 17 breakout (for prediction) was modified to create the predictions but mostly to derive the transformation of the posterior chain data.

**Using my LBA Assignment**: The "Model Comparison" section has a lot of similarities with the LBA assignment as I used it as a starting point to justify the choice of comparison technique. The approach in both assignments is also very similar, so referring to the LBA was helpful in crafting this section.

## 7. AI Statement

AI was used solely to convert the PyMC output into LATEX-tables. The ChatGPT conversation can be accessed here: https://chatgpt.com/share/69222e76-7c8c-8012-8a32-fbb172f1b1d2