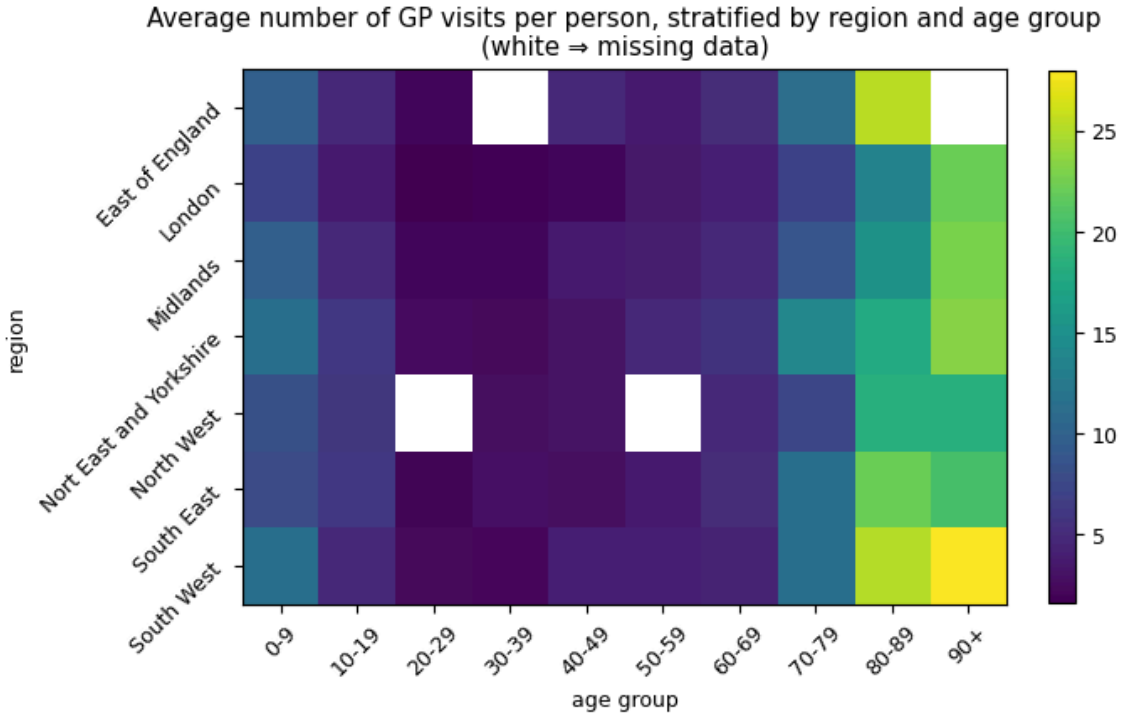
Data

The data set is in a 7 × 10 × 30 array for the 7 geographical regions, 10 age groups, and 30 samples per group.



Preprocess data

To preprocess the data, I flatten the raw dataset and stack the region and age group indices. The region_index array creates 300 repetitions for each of the 7 regions, while the age_index array assigns 30 repetitions for each of 10 age groups, repeated across all regions. This mean that for each region index, the age index will go from 0 to 9 (30 samples for each unique age index) and repeat the same 10 age categories when we move on to the next region index.

Each row in the output DataFrame corresponding to an individual with their age group, region and the count of number of times they attended an appointment with a GP during the past year. After removing 4 missing groups (each with 30 samples), the final sample size of the dataset is 1980 individuals, with the mean count of doctor visit of 8.385 and standard deviation of 10.953. The smallest count is 0 (as many people don't go to the doctor) and the largest count is 71.

	region_index	age_index	count_data
0	0	0	12.0
1	0	0	17.0
2	0	0	12.0
3	0	0	17.0
4	0	0	3.0

Complete pooling zero-inflated poisson model

I use a Zero-Inflated Poisson likelihood function for this data set because there are many more 0s in the data set than we would expect from a Poisson distribution since a lot of people never visit the doctor. In a complete pooling Zero-Inflated Poisson (ZIP) model, the lambda (Poisson rate) and theta (zero-inflation probability) parameters are shared across all observations, rather than varying by group or individual observation. The model assumes that all data points are generated homogeneously from the same Zero-Inflated Poisson (ZIP) distribution with these global parameters, without accounting for any group-level differences.

By explicitly modeling these two sources of zeros, the ZIP model avoids over-penalizing λ when excess zeros are observed, leading to a more accurate representation of the count-generating process. These are two processes to handle two sources of 0:

- **Zero-Inflation Process**: With probability θ, the count is zero due to structural factors unrelated to the Poisson process (e.g., lack of participation, absence of an event).
- **Poisson Process**: With probability $1 - \theta$, the count is drawn from a Poisson distribution, which can also produce zeros naturally when λ is low.

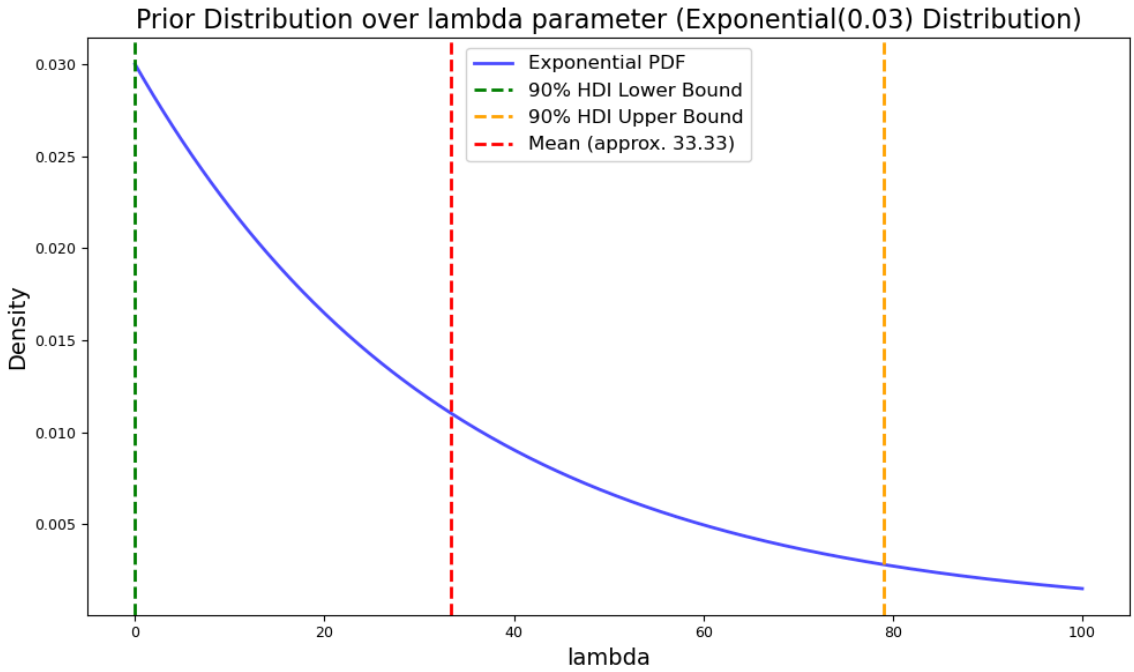
Set up prior

Poisson Rate (λ):

$$\lambda \sim \text{Exponential}(\text{rate} = 0.03)$$

The prior reflects the belief that the average count rate is small and strictly positive. An Exponential prior is chosen because it is non-negative and places more probability on

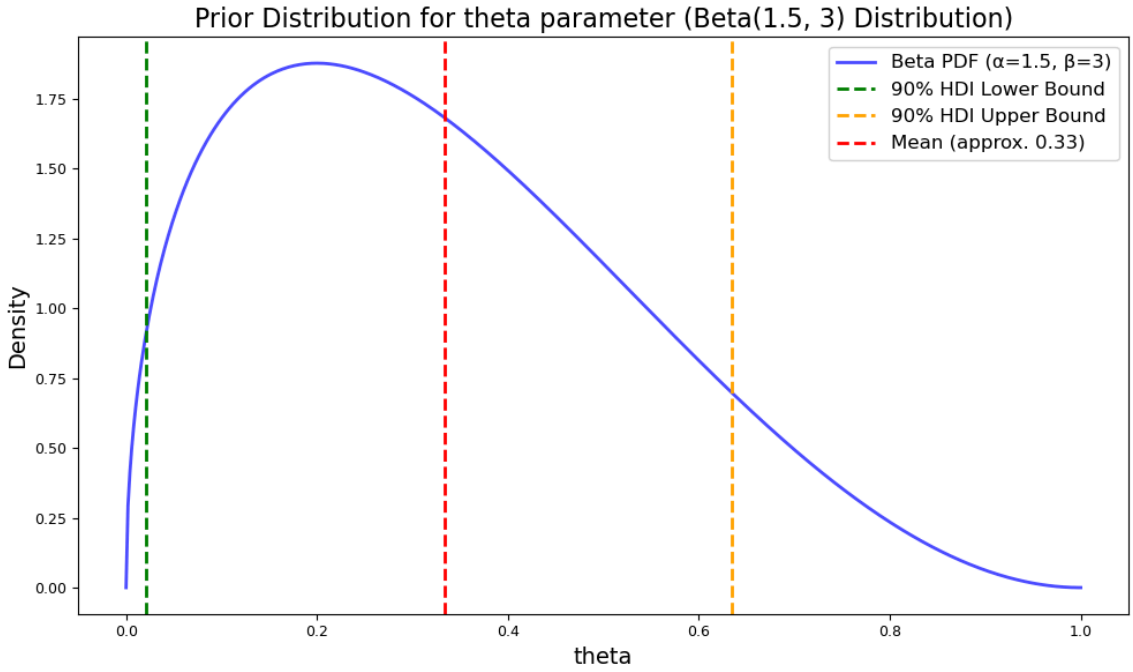
smaller values, consistent with expectations for sparse count data. When the rate is 0.03, the prior distribution allows the mean Poisson count rate to be centered around 33.33 but can go up to 80 for 90% of the time and more rarely, over 100. This range is wide enough since it is the Poisson mean count rate and not the actual count values, which can go up very large values even with smaller mean like 33.33.



Zero-Inflation Probability (θ):

$$\theta \sim \text{Beta}(\alpha = 1.5, \beta = 3)$$

A Beta distribution is used as it is well-suited for probabilities, with support on the interval ([0, 1]). The chosen parameters ($\alpha = 1.5$), $\beta = 3$) results in a prior distribution slightly skew to the right and have a large 90% interval, reflecting a prior belief that structural zeros are moderately likely but not dominant. For each individual, their expected probability to belong to the group that never see a doctor is 0.33.



Likelihood

The ZIP likelihood accounts for both sources of zeros and non-zero counts:

$$P(y = 0) = \theta + (1 - \theta) \cdot e^{-\lambda}$$

$$P(y = k) = (1 - \theta) \cdot \frac{\lambda^k e^{-\lambda}}{k!}, \quad k > 0$$

- When y = 0, it can result from the structural zero process (with probability θ) or the Poisson process (with probability $(1 - \theta)e^{-\lambda}$).
- When y > 0, counts are generated only from the Poisson distribution.

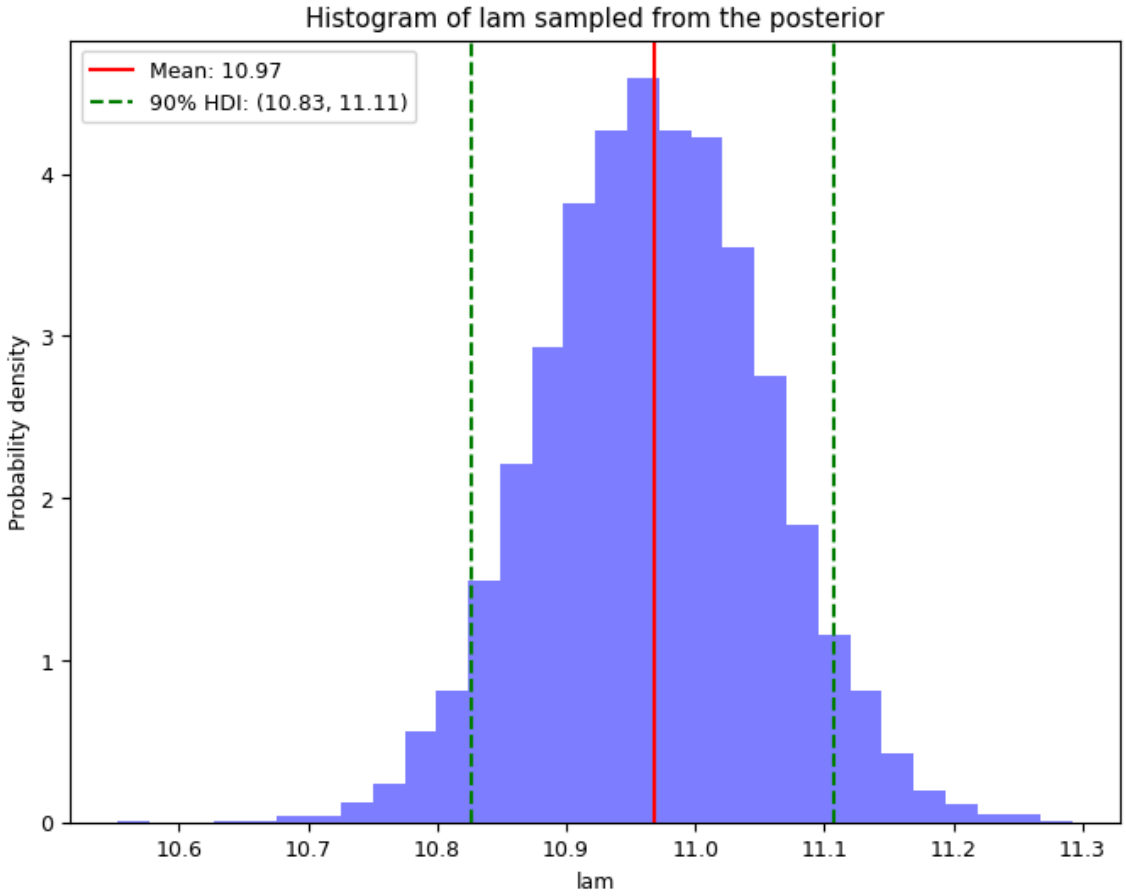
Construct model

I define a complete_pooling_model using pm.ZeroINflatedPoisson to model the likelihood. The sampler shows good convergence with r_hat = 1, uniform rank distribution and a normally distributef pair plot for the λ and θ parameters.

Posterior distribution

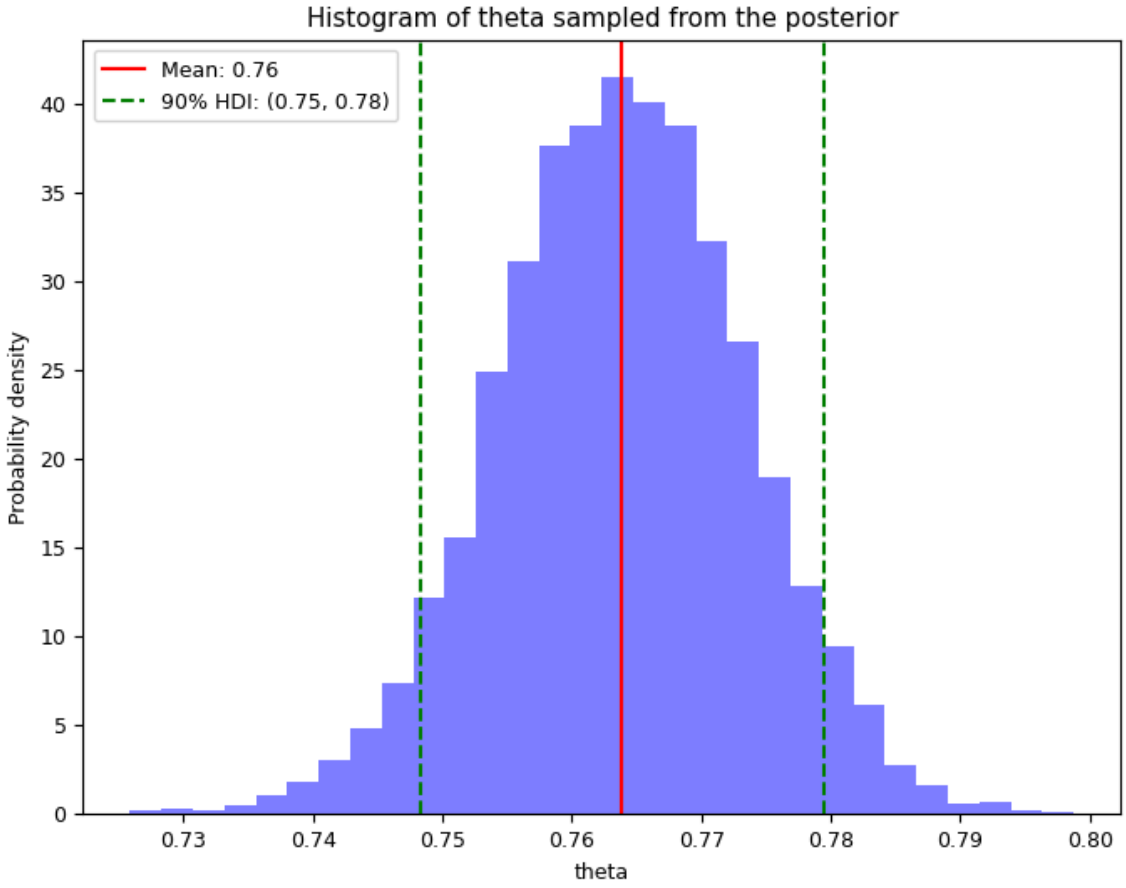
Lambda (Poisson rate/ Mean Poisson count)

A posterior mean of 10.97 suggests that, on average, the count of doctor visits is expected to be around 10.97, given the data. The 90% HDI (10.83, 11.11) is quite narrow, indicating a high level of certainty in this estimate.



Theta (zero inflation probability)

From the posterior plot of theta shown below, we can see that the data influence the prior significant because the posterior mean shift to 0.76 with a very narrow 90% HDI interval from 0.75 to 0.78. This narrow interval suggests a high level of certainty about the estimated value of theta, reflecting the belief that the zero-inflation probability (the proportion of structural zeros) is between 75% and 78%. This implies that approximately 75% to 78% of the population likely belongs to a group that never sees a doctor.



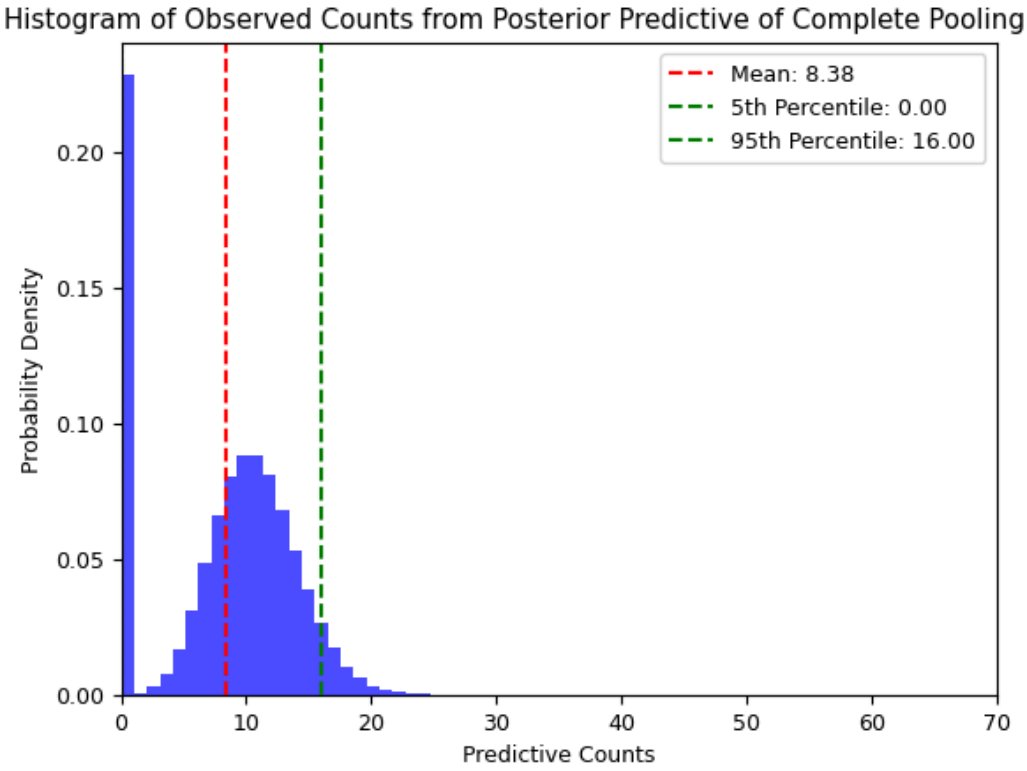
Posterior predictive check

To assess the model's ability to predict new observations, I created a posterior predictive distribution that first sample a set of parameters (lambda and theta) from their posterior distributions, and then generate new values from the Zero-Inflated Poisson (ZIP) distribution using those set of parameters.

While the mean of the posterior predictive distribution closely match the dataset mean, the maximum value is lower than the dataset's (31 vs 71) and the stadard deviation is also lower (5.484 vs 10.953).

This suggests that while the model is capable of reproducing the general magnitude of counts (as indicated by the similar means), it is underestimating the spread and the extreme values (the larger counts in the observed data), particularly in the upper tail of the distribution, where higher values (like 71) appear less frequently in the posterior predictive samples.

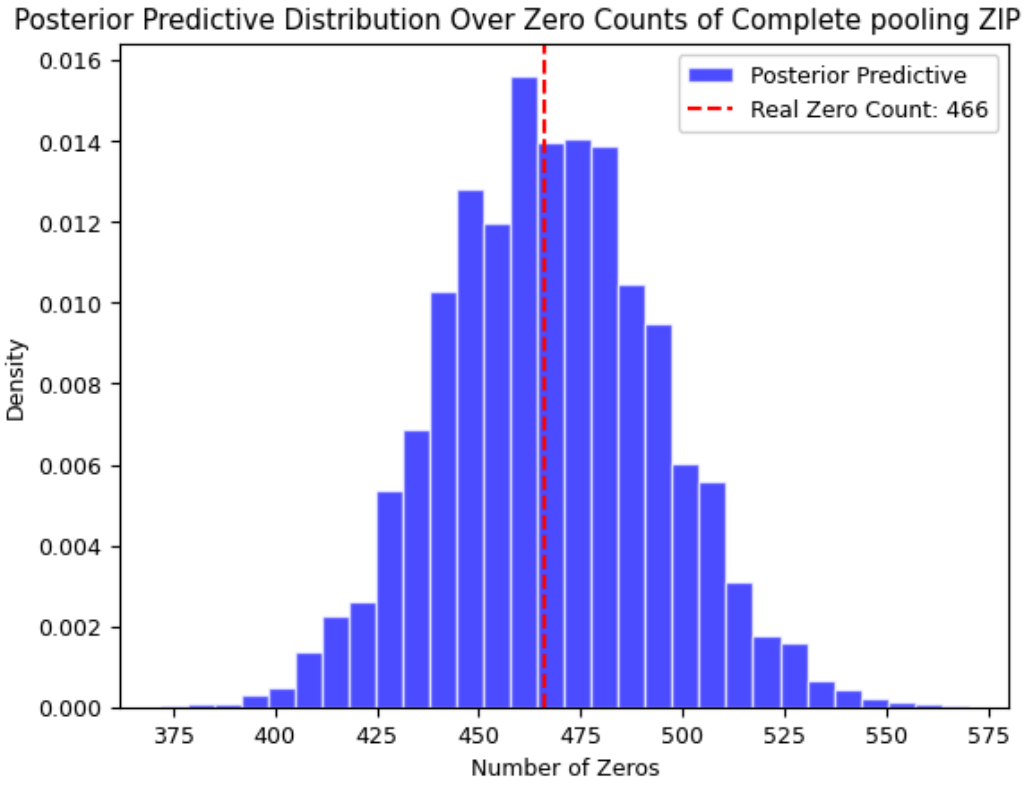
This is highly due to the fact that complete pooling uses a single set of parameters for all observations, it forces the model to fit all data points using the same Poisson and zero-inflation process. The model doesn't account for potential group-level differences accross regions (where one might have more hospitals/ health issues) and accross ages (where the elderly groups tend to visit doctos more often). The one-size-fits-all distribution will favor the majority group, thus putting very low probability mass at the larger tail (extreme values) of the distrbution.



Histogram over the number of zeros according to the posterior-predictive distribution:

However, the posterior predictive of the complete pooling model does not have trouble with estimating the proportion of zero count, evidenced from the mean of the zero-count distribution closely matches the real zero count in the dataset. This is because zero counts are primarily driven by the zero-inflation mechanism (theta), which the model can estimate accurately, even when all observations share the same value for theta).

Additionally, we can see that the Zero inflated Posson model allows for large probability mass for zero count values both larger and smaller than the real zero count, indicating that it is not penaltizing large zero data as in the normal Poisson model.



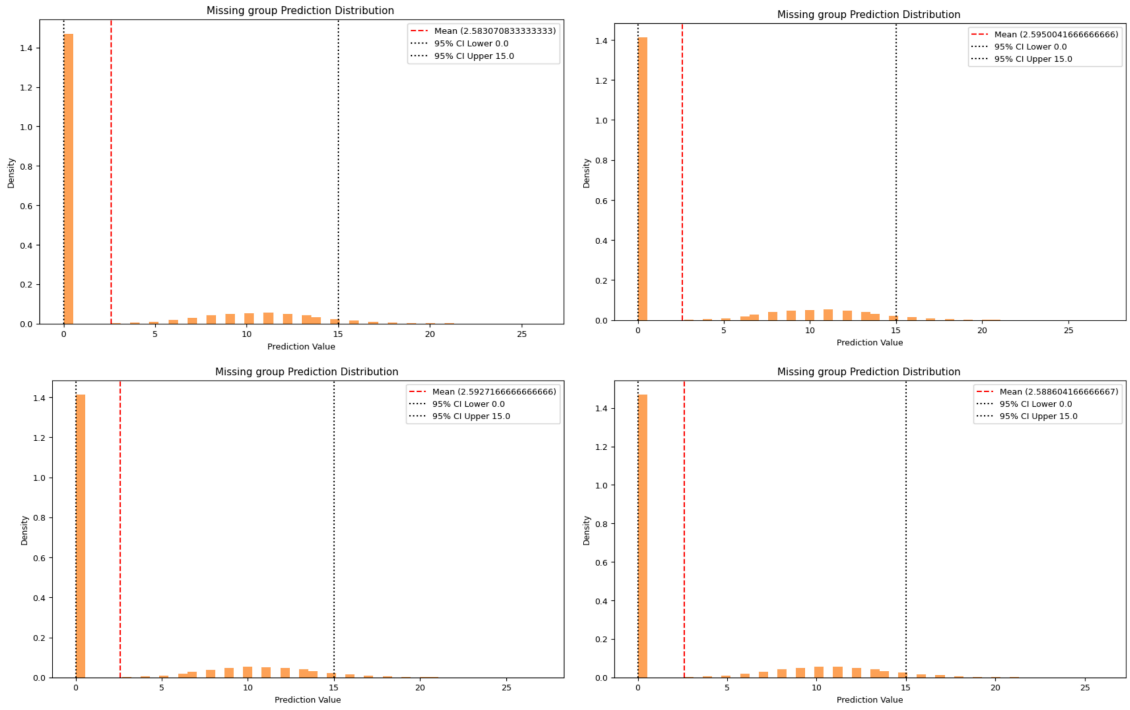
Prediction for missing groups

Since the count for all individuals are determined by the same set of population parameters (lambda, theta)and the only variation comes from the data itself, there is no distinction between groups

-> The process of generating prediction for each missing group will be the same, resulting in similar distribution for each group

Strategy: Predictions for these groups are generated based on the **posterior samples** of λ and θ that were learned from the original dataset

- **Zero-Inflated Poisson Sampling**: For each individual, generate multiple predictions by sampling from the **Zero-Inflated Poisson (ZIP) distribution** using the posterior samples of λ and θ. First we determine whether the count is zero (via a Bernoulli trial using θ) or a sample from the Poisson distribution. If sampled from the Poisson distribution, we will generate the count of the individual using a random variable from a Poisson distribution with the rate parameter being the λ posterior).
- **Repeat for Multiple Samples**: This process is repeated for all pairs of posterior samples to account for the uncertainty in the parameter values. Each individual's count is predicted once for each posterior sample.
  - For each group, the final predictions will have shape `(num_people * num_samples)`, where `num_people` is the number of people in the missing group (e.g., 30) and `num_samples` is the number of posterior samples (e.g., 8000). I visualize the prediction for each missing groups by generating the histogram of all predictive posterior samples of all individuals in the group.



The 4 histograms above represent the prediction for 4 missing groups. Indeed there is no significant difference between 4 groups. For example, for missing group 4, the mean of missing group prediction is around 2.6, the 95% of the probability mass lie below count value of 15, the standard deviation is 4.93 and the maximum prediction value is 28.

Notably, the mean of missing group prediction is around 2.58, which is smaller than the posterior predictive distribution's mean (8.38) and the standard deviation is also smaller (4.93 vs 5.484). This is because in posterior predictive check, we are sampling from the posterior distribution of the parameters, and then generating predictions that are conditioned on the observed data. However, when predicting for missing groups, you're generating predictions for a group that has no observed data. As a result, the predictions for the missing group are based purely on the posterior distribution of the parameters. As shown above, the variation in posterior distributions of both lambda and theta is small -> when we do not show it any data that can convince that the estimates might be higher, the prediction of missing groups shows much lower mean and variation.

Partial pooling zero-inflated model

Model overview

Partial pooling model allow groups (age groups and regions) to have their own specific effects while borrowing strength from the overall population-level information (global effect), particularly when the data from some groups might be sparse.

1. **Global Hyperprior**:
    - The **global intercept** (`lambda_bar`) represents a **population-level effect** that is shared across all age and region groups. It informs the model about the central tendency of the rate for all observations, regardless of age or region.

- It is modeled as a **normal distribution** with mean 0 and standard deviation 1.5:

```
lambda_bar = pm.Normal('lambda_bar', mu=0, sigma=1.5)
```

2. **Group-Specific Priors**:
   - The model has **group-specific priors** for age and region effects:
     - `sigma_age` and `sigma_region` are both Exponential distributions with the rate = 1 to control the deviations of each age group and region from the global effect. The flexibility of Exponential (1) allows for any deviation of some age/ region group from the global baseline `lambda_bar`. Each cluster variable needs its own standard deviation parameter to adapt to different amount of pooling towards the global bas
     - I reparameterize the prior using 2 standard normal distribution `std_age` and `std_region` since the sampler was struggling with divergence.

```
sigma_age, sigma_region = pm.Exponential('sigma_age', lam=1)
```

   - The group-specific effects (`lambda_age_effects` and `lambda_region_effects`) are determined as:

```
lambda_age_effects = pm.Deterministic('lambda_age_effects', std_age * sigma_age)
lambda_region_effects = pm.Deterministic('lambda_region_effects', std_region * sigma_region)
```

3. **Combination of Global and Group-Specific Priors to Create the Poisson Rate**:

   - The **total effect** for each observation (`lambda_total`) is the sum of the global effect (`lambda_bar`) and the group-specific effects (`lambda_age_effects[age_index]` and `lambda_region_effects[region_index]`). :

```
lambda_total = pm.Deterministic('lambda_total', lambda_bar + lambda_age_effects[age_index] + lambda_region_effects[region_index])
```

   - All `lambda_bar` , `lambda_age_effects` and `lambda_region_effects` are centered at 0 but with different spreads around the global mean (lambda_bar). This uniform centering of all global and group effects around 0 ensure that before information updating, the age and region effects are assumed to have no overall bias in any direction and that after updating, the varying effects will correctly reflect shrinkage of the group-specific effects (lambda_age_effects and lambda_region_effects) towards the global mean (lambda_bar) (depending on how informative the data from that particular age group or region is it).

   - The **Poisson rate** (`mu`) is then transformed to a positive value by applying the exponential function to ensure positive support for mean count:

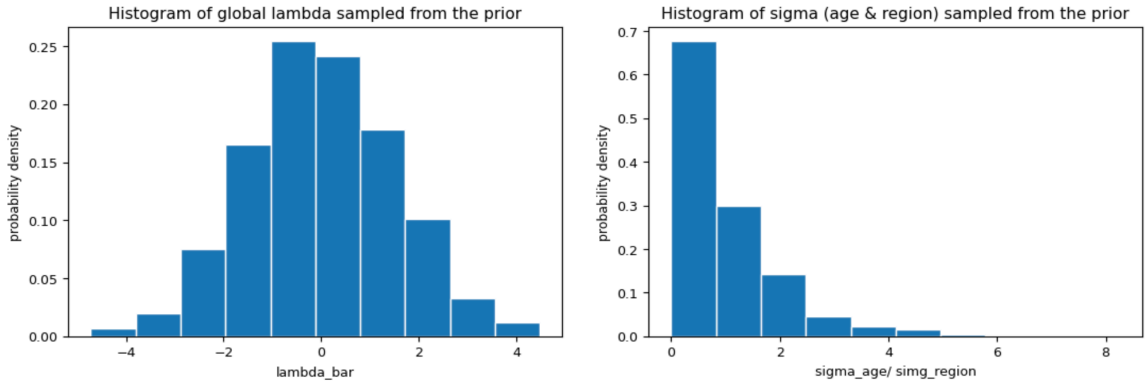
```
mu = pm.Deterministic('mu', pm.math.exp(lambda_total))
```

4. **Summary:**

- The **global prior** (`lambda_bar`) represents a baseline effect that applies across all groups (age and region).

- The **group-specific priors** (for age and region) allow each group to have its own effect, but the size of these effects is shrunk toward the global prior by the hierarchical structure.
- The **total effect** (`lambda_total`) for each observation is the sum of the global effect and the group-specific deviations, and this determines the Poisson rate (`mu`).
- **Zero-inflation** is modeled separately, influencing the probability that an observation is zero.
- The model influences the **posterior distribution** of the parameters by combining the global and group-specific information, and this posterior is used to make predictions for both the count data and the zero-inflation probability.

Set up prior

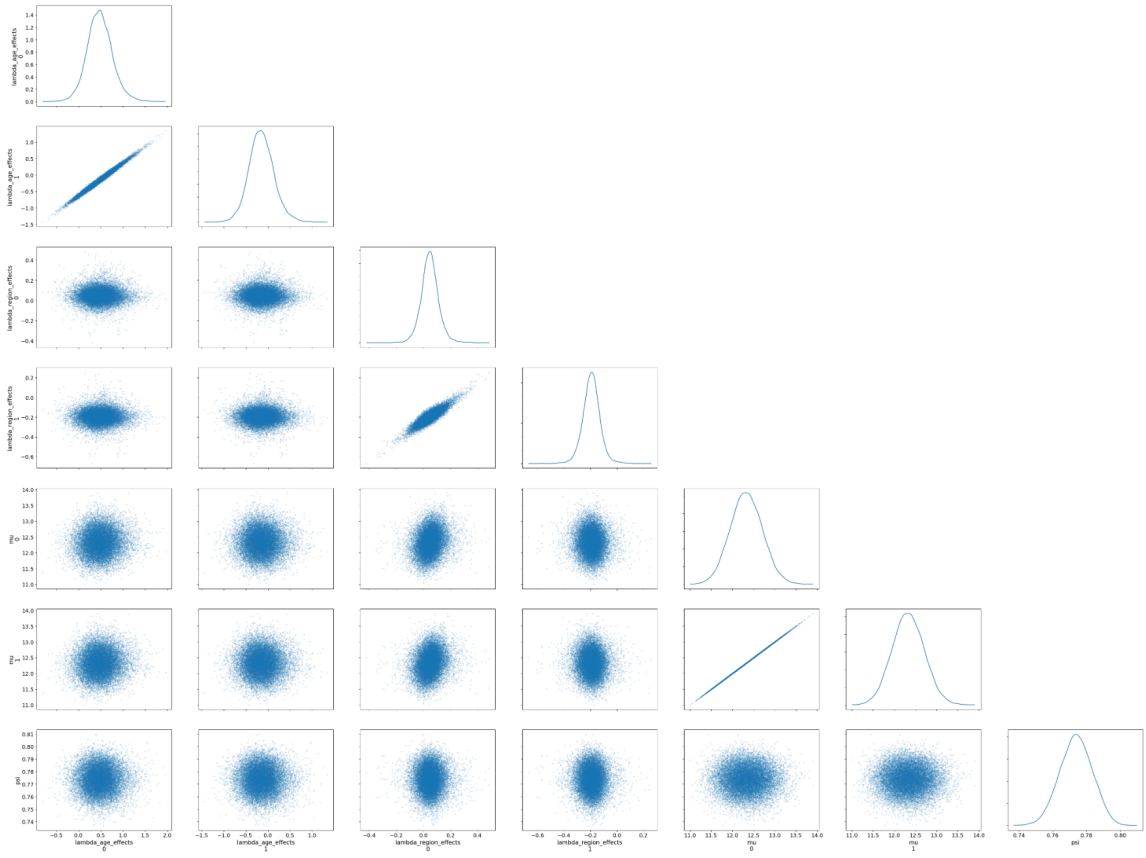


The prior `lambda_bar` $\sim \text{Normal}(\text{mu} = 0, \text{std} = 1.5)$ is initiallyset to 0, meaning we are assuming no strong prior belief about the baseline effect of all individuals (before accounting for group-level effects). `sigma_age` and `sigma_region` $\sim \text{Exponential}(\text{rate} = 1)$ to have positive support and is assumed to cluster around the global mean but also allows for the possibility of relatively large values.

The range are kept small because the Poisson rate (`mu`) is the exponential transformation of `lambda_total` , which is the sum of all priors above -> transform small range to very large range that is flexible enough to account for all possible values of `mu`

Sampling diagnostic

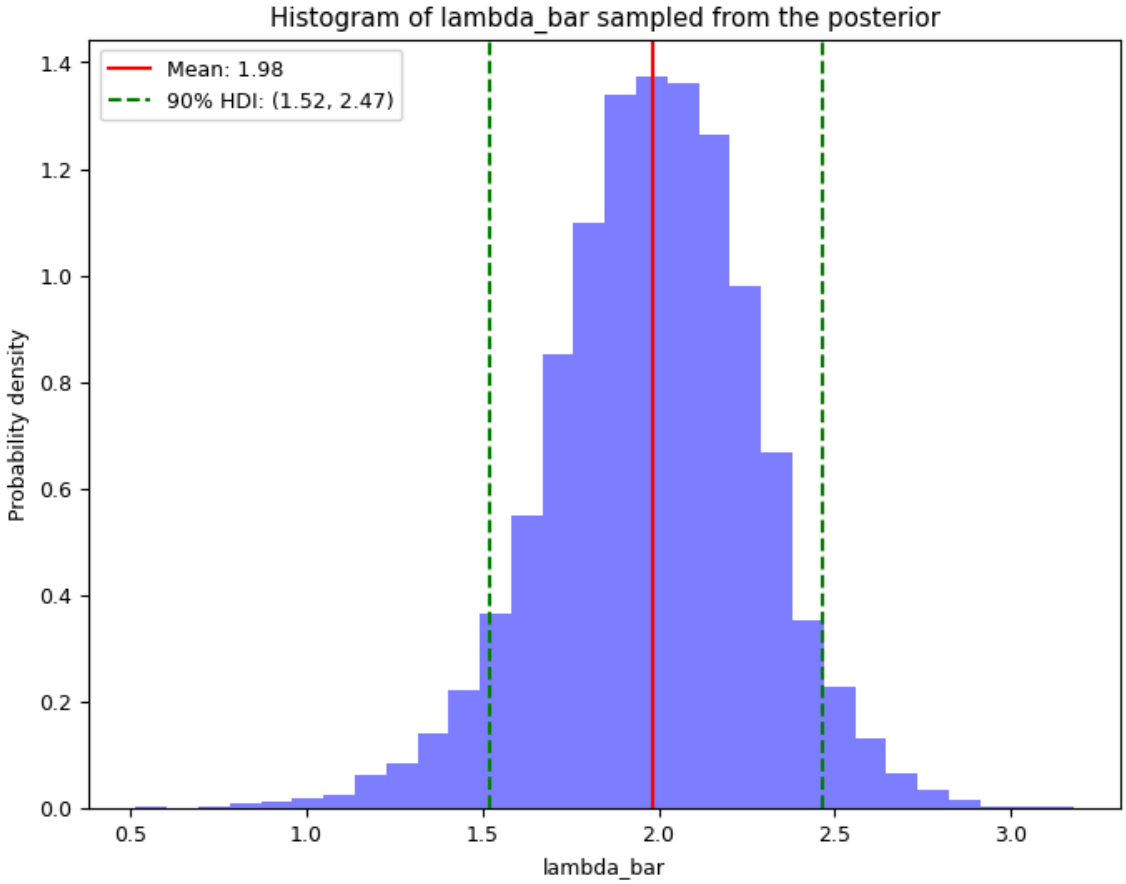
I check the sampler for r_hat = 1, uniform rank plot and normal distribution of pairwise parameters. Note that there is strong positive relationship between the `lambda_age_effect` , `lambda_region_effect` and poisson mean `mu` for observations of the group. This is notmal because while each observation are drawn independently, they are highly related because they share the same underlying group-level effect.



Posterior distributions

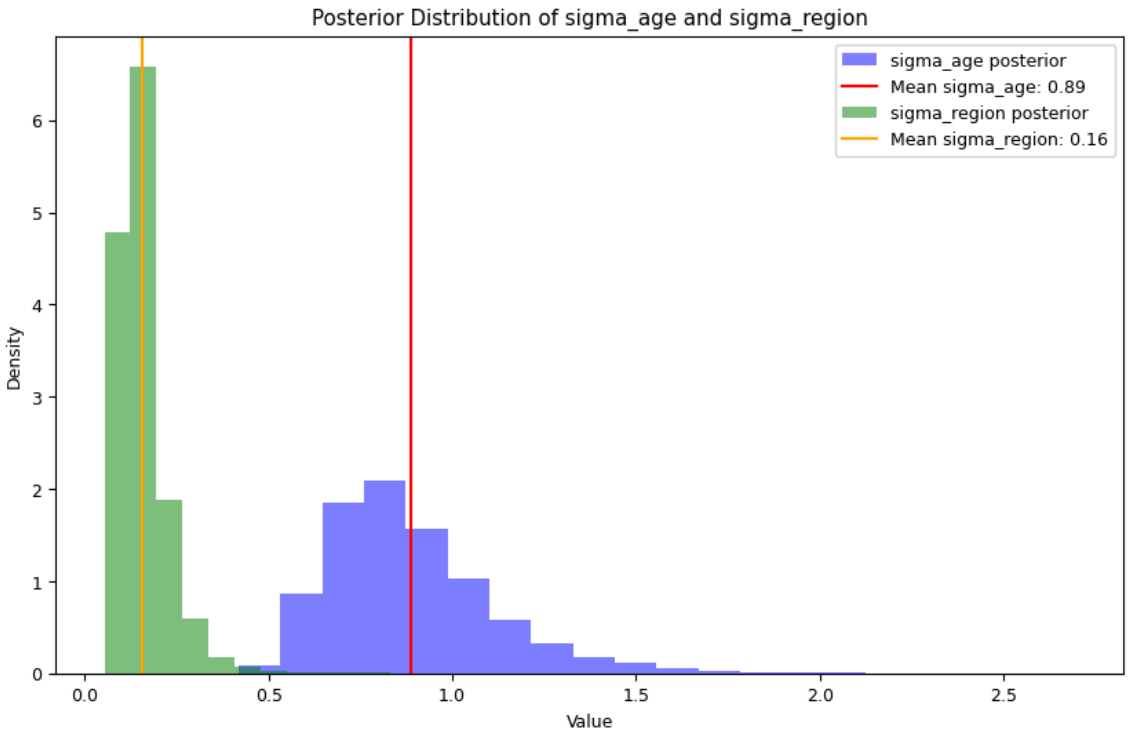
Lambda_bar: global hyperprior

Based on the data, the model has learned that the central tendency for the global intercept is likely around 1.98, indicating that the baseline effect (before considering age and region effects) is higher than the prior expectation of 0.



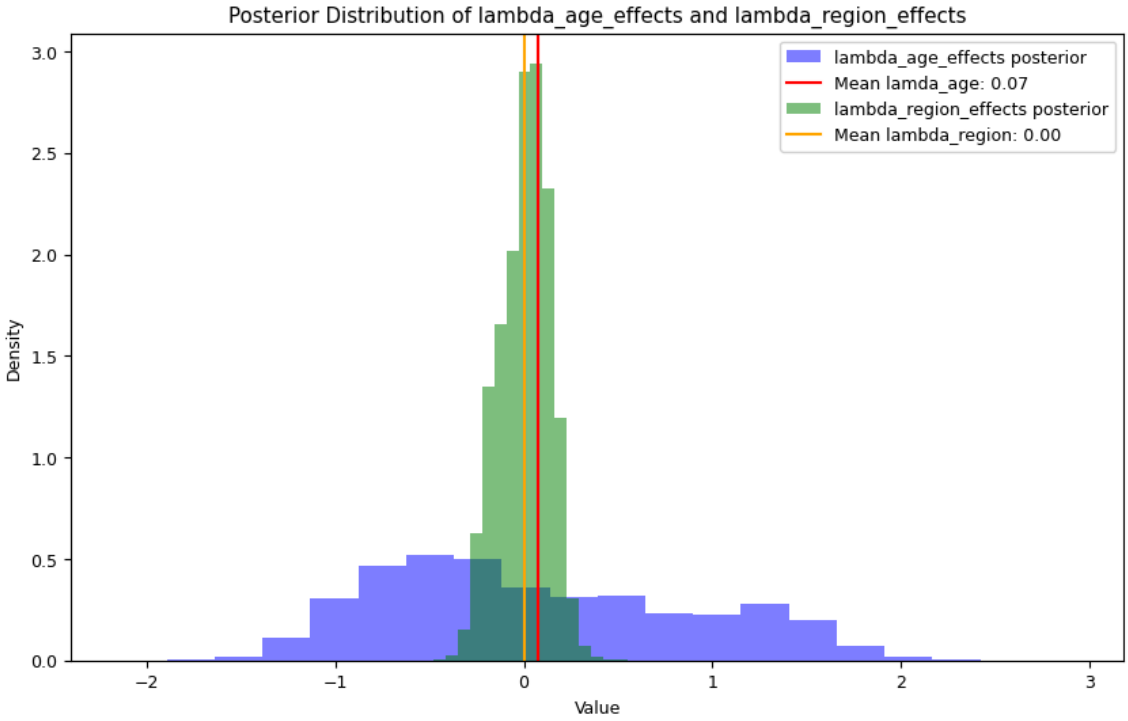
Sigma_age and sigma_region

The posterior distribution shows that the mean of sigma_age is much higher (0.89) than sigma_region posteriors (0.16) with high certainty (since there is little overlap). This shows that the estimated variation across age group is much larger, resulting in less shrinkage towards 0 than the region group (that it may contribute minimally to the prediciton model)



Lambda_age and lambda_region

Corresponding to sigma, the spread of lambda_age effects is much larger, meaning we have higher variation accross age groups and region groups. Additionally, the mean of lambda_region posterior is still the same as its prior's mean (0) indicating that seeing the data does not have updating-belief effects on the contribution of the region group.



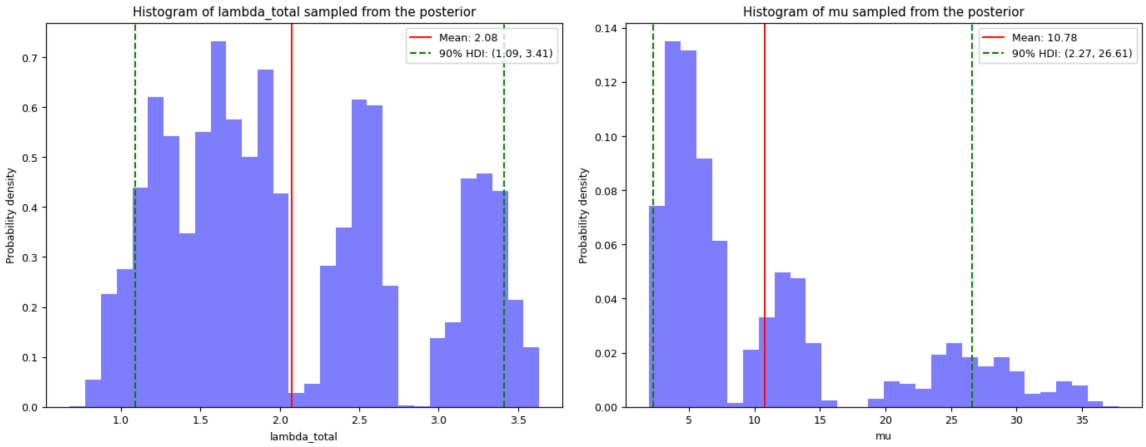
Lambda_total and mu

The posterior distribution of lambda_total and mu have multimodality, meaning that the posterior distribution suggests multiple plausible values for the parameter, each

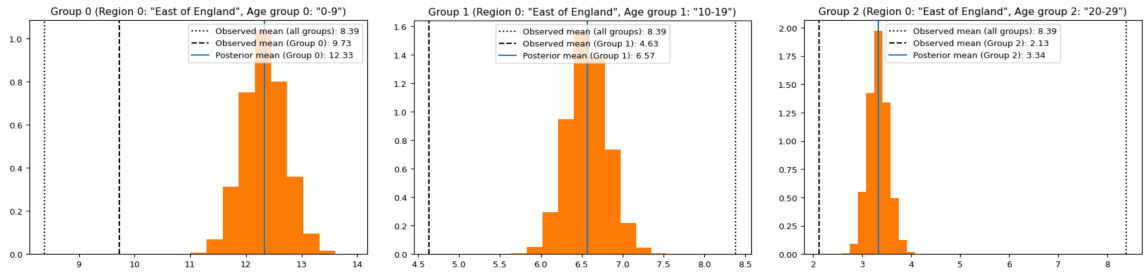
corresponding to different regions or types of behavior in the data. This indicates that the Poisson count rate are different accross different groups, correctly showcasing the strength of multilevel model.

In comparison to the complete pooling model, while the mean of the parameter `mu` is roughly similar between this partial pooling model (10.78) and the complete pooling model (10.97), the variance of the parameter `mu` is significantly larger (79.06 vs 0.0074)

-> The partial pooling model allows different groups to have unique deviation away from the global mean, which results in less shrinkage and more variation in the data. Thus, it is better at accounting for extreme count values.



Plot posterior distribution of some groups [`mu`]

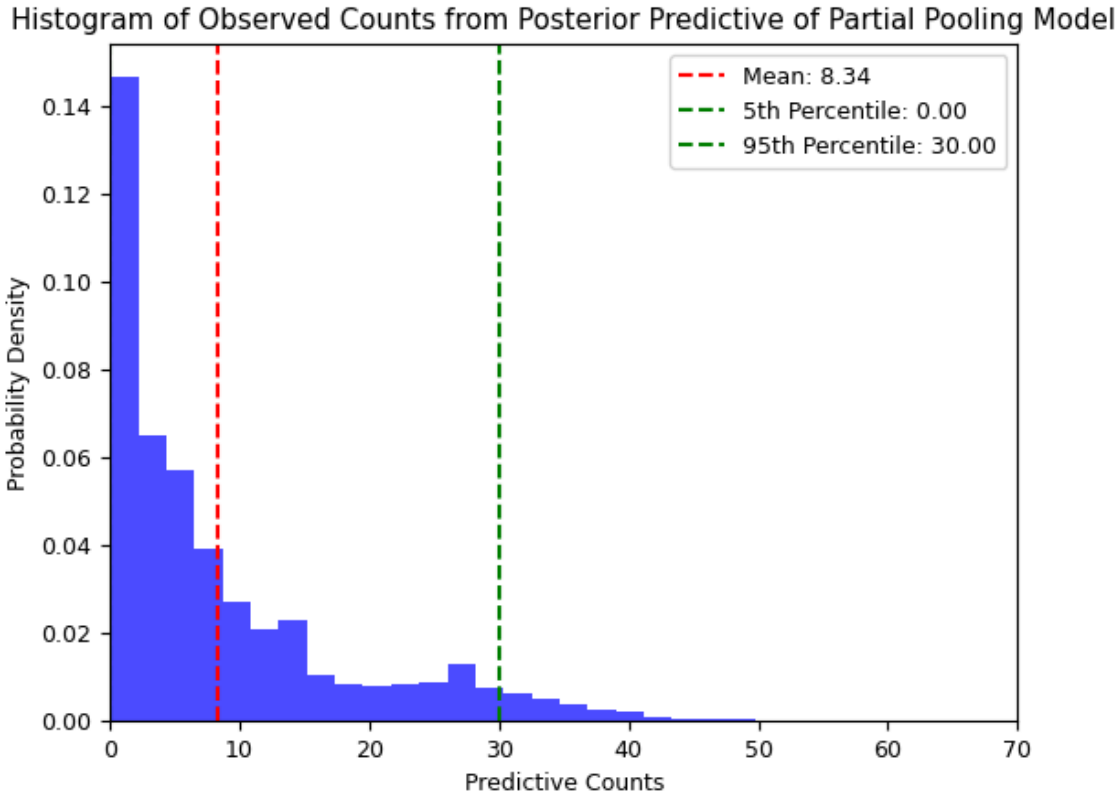


From the examples of 3 different groups, we can see that the posterior mean of the Poisson count rate adapts to where the observed mean is relative to the global mean. In group 0, the posterior mean is significantly larger than in group 3 because the observed mean in group 0 is larger than the global observed mean, while the observed mean in group 2 is smaller than the global observed mean.

-> This shows that the model allows different variation across groups.

Posterior predictive

The posterior predictive check shows greater fit in the partial pooling model. While the mean is roughly the same as in complete pooling model, the maximum value of posterior predictive samples (65) has cover a greater range and closely match the real data. The 95th percentile (30) is larger than complete pooling model (15), as well as the Standard deviation of (9.478 vs 5.484)



The posterior predictive distribution over zero count is similar to that of complete pooling model, sinct both are Zero inflated poisson model that handles 0 count in the same way.

Predict for missing groups

When predicting for a new, unseen group (e.g., a new age group or region) that was not part of the data used to construct the model, we need to rely on the global hyperpriors and the posterior distribution of the global parameters (`lambda_bar` , `sigma_age` and `sigma_region`) and not use group-specific adapting posteriors because these priors (`lamda_age_effects` , `lamda_region_effects` , `lambda_total` and `mu`) are specific to the training data, which did not include missing data.

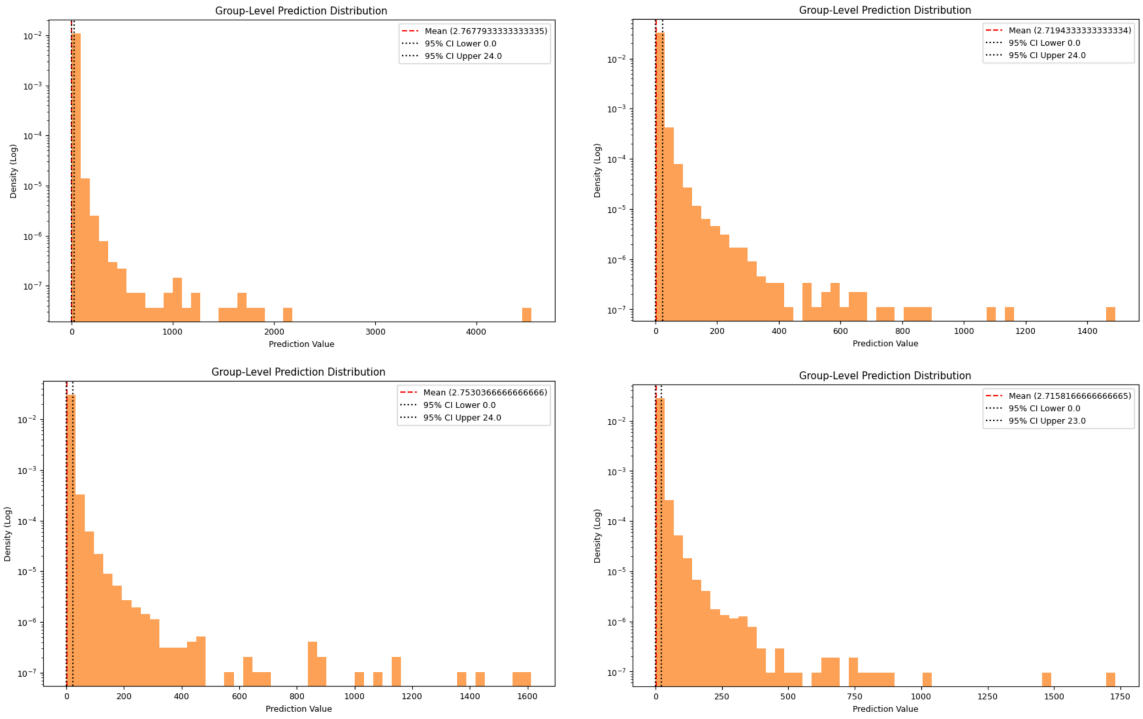
Workflow

1. **Sample from the Posterior for Global Parameters**: Use the posterior samples of `lambda_bar` , `sigma_age` , and `sigma_region` .
2. **Simulate Group-Specific Effects**: For the new age group and region:
   - Draw `std_age_new ~ Normal(0, 1)` and scale it using posterior samples of `sigma_age` :
     $\text{lambda\_age\_effect} = \text{std\_age\_new} \times \text{sigma\_age}$
   - Similarly, draw `std_region_new ~ Normal(0, 1)` and scale it using posterior samples of `sigma_region` :
     $\text{lambda\_region\_effect} = \text{std\_region\_new} \times \text{sigma\_region}$

3. **Combine Effects**: Combine these effects with `lambda_bar` to calculate `lambda_total`:
   $\lambda_{\text{total}} = \lambda_{\text{bar}} + \lambda_{\text{age\_effect}} + \lambda_{\text{region\_effect}}$
4. **Transform to Poisson Rate**: Transform the total effect to the Poisson rate:
   $\mu = \exp(\lambda_{\text{total}})$
5. **Zero-Inflation Probability**: Use the posterior samples of `psi` to model zero-inflation.
6. **Simulate Counts**: Use `zero_inflated_poisson` function with the new `mu` and `psi` to simulate counts for the new group.

Key assumption:

The variability (sigma_age, sigma_region) is consistent across all groups, including the new ones. The new group's standardized effects (std_age, std_region) are drawn from the same priors as the groups in the training data.



The histogram of missing group prediction shows that while the mean prediction is not really different from the complete pooling model, the variation is greatly different. For example group 4, The 95% confidence interval is 23 compared to 16, the Standard Deviation is 11.72 compared to 4.93 and the maximum predicted value goes up to 1731 (group 1's prediction even goes to 4000)

This is due to the fact that the partial pooling model allows large variation between the groups. Therefore, when predicting for an unseen group, we have large uncertainty in whether the new group belong to a zero visit, occasionally visit or frequently visit group, thus resulting in a wider range of predicted values. When predicting for a new group, the model essentially samples even from the very long tail of the posterior distributions of lambda_age_effects and lambda_region_effects without the constraint of observed data. Additionally, since we don't have the data and its variance, the model cannot apply

shrinkage to the group-specific effects and keep predicted values like counts relatively closer to the global mean.

In conclusion:

When predicting unseen data, the complete pooling model generates predictions with limited variance because it assumes all groups are identical and estimates a single global mean. In contrast, the partial pooling model generates predictions with excessive variance due to its ability to account for group-level differences and its lack of direct observations for the new group.

The difference in variance exist in both in-group variance and between-group variance:

Complete Pooling ZIP Variance: [24.2426909 24.30822421 24.32439529 24.31546597]
Partial Pooling ZIP Variance: [174.97338649 5399.33472894 167.57372259 150.34890919]

- In-group variance: The 4 missing group variance are around 24 which is significantly smaller than that of partial pooling variance
- Between group variance: Since partial pooling model allows each group to vary differently, we have great difference in the prediction between groups. While the variance of prediction accross 4 groups in complete pooling mode is consistent (around 24), there is much variability in the variance of partial pooling prediction (can range from as low as 150 to as high as 5399)

The choice of which model to prefer depends on the trade-off between bias and variance, as well as the context of the analysis:

Complete Pooling Model:

- Best suited for: Scenarios where there is strong prior belief or evidence that the groups are similar, or when the data is too sparse to estimate group-level effects reliably.
- Strength: Low variance in predictions, but high bias if the groups are actually different.
- Limitation: Fails to capture meaningful differences between groups, leading to underestimation of variability.

Partial Pooling Model:

- Best suited for: Scenarios where group-level differences are expected, and the goal is to balance between capturing these differences and maintaining some degree of shrinkage to a global mean.
- Strength: Can model between-group variability while regularizing predictions through shrinkage.
- Limitation: For unseen groups, the lack of data can lead to overly uncertain predictions with potentially extreme values.

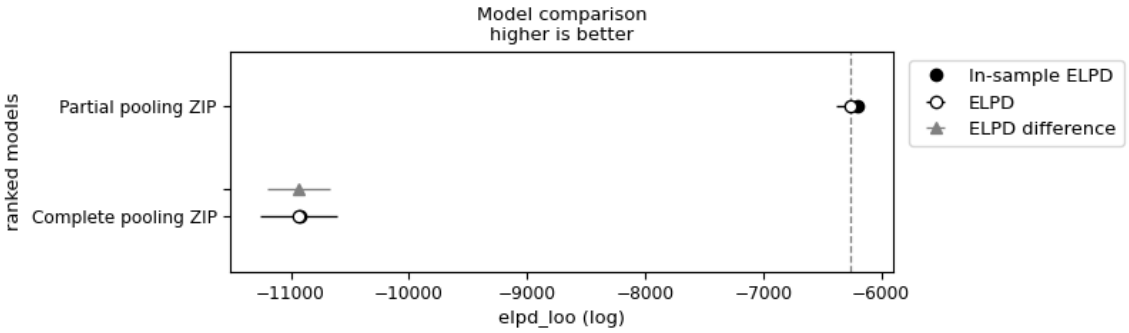
Model comparison for predicting existing groups

PSIS-LOO (Pareto Smoothed Importance Sampling Leave-One-Out) cross-validation is a method used for assessing model predictive accuracy. Overall, ELPD is average log-likelihood of all samples. Higher value of ELPD indicates larger average accuracy, and thus, better ability to fit new, unseen data.

From the plot below, the Partial pooling ZIP is more effective, achieving the higher ELPD of -6265 with a weight of 0.87. In contrast, the Complete pooling Model exhibits the lower ELPD of -10933 and a lower weight 0.12, highlighting its inadequacy for this dataset.

The ELPD difference between the Partial pooling ZIP model and the Complete pooling is 4668, which mean that it is highly likely that the observed data is under the Partial pooling ZIP model compared to the Complete pooling Model. Visually, the ELPD difference error bars do not overlap with the best ELPD obtained by the Partial pooling ZIP model.

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Partial pooling ZIP	0	-6265.638422	59.997924	0.000000	0.875981	115.067685	0.000000	False	log
Complete pooling ZIP	1	-10933.800987	12.992884	4668.162565	0.124019	321.520298	266.693689	False	log



Optional stretch goal

In a Poisson distribution, the mean and variance are equal. However, in the dataset, the variance is significantly larger than the mean, a phenomenon called overdispersion. Even when only counting non-zero observations, the mean of the non-zero dataset is 10.966 while the variance is 128, 58.

If the variance of the count data exceeds what the Poisson process can accommodate, the ZIP model may produce:

- Poor fits to non-zero counts.
- Overconfident predictions for extreme values.

The Negative Binomial distribution generalizes the Poisson by introducing an additional parameter to model overdispersion. Combining with a zero-inflated process, we have ZINB model - Zero inflated negative binomial model that addresses both excess zeros and overdispersion. The ZINB model allows the variance to exceed the mean, making it more robust for datasets with high variability in counts. By capturing overdispersion, the ZINB model reduces the likelihood of overconfident predictions and better matches the observed count distributions.

It combines:

- A binary process to account for structural zeros.
- When (1-psi), or when the individual does not belong to the "never go to doctor group", the count for that individual's visit count is sampled from a Negative Binomial distribution, which introduces a dispersion parameter alpha/ phi to account for extra variability.

Negative Binomial Component This component models the **count data** (including both zero and non-zero counts) with overdispersion:

- The NB distribution introduces a **dispersion parameter** α to account for variance that scales quadratically with the mean (μ) :
  $P(Y = y|\mu, \alpha) = \frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$ $

where α:** Dispersion parameter that adjusts variance $(\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\alpha})$.

Combined Likelihood The ZINB likelihood combines the zero-inflation and NB components:

$$P(Y = y) = \begin{cases} \psi + (1-\psi)\cdot \mathrm{NB}(0|\mu,\alpha), & \text{if } y = 0 \\ (1-\psi)\cdot \mathrm{NB}(y|\mu,\alpha), & \text{if } y > 0 \end{cases}$$

This architecture allows the ZINB model to handle:

- **Excess zeros** (via ψ).
- **Overdispersion in non-zero counts** (via α).
  - $\alpha \to 0$: Reduces the NB distribution to a Poisson distribution, minimizing variance.
  - Large α: Increases variance significantly, allowing the model to account for greater variability in the counts.

As a result:

- Predictions for unseen groups reflect the uncertainty in both:
  1. Whether a zero is structural or from the NB distribution ((\psi)).
  2. The expected rate (μ) and variability (α) for non-zero counts.
- This leads to more realistic predictions for data with excess zeros and overdispersed counts compared to simpler models like Poisson or Negative Binomial alone.

Construct model

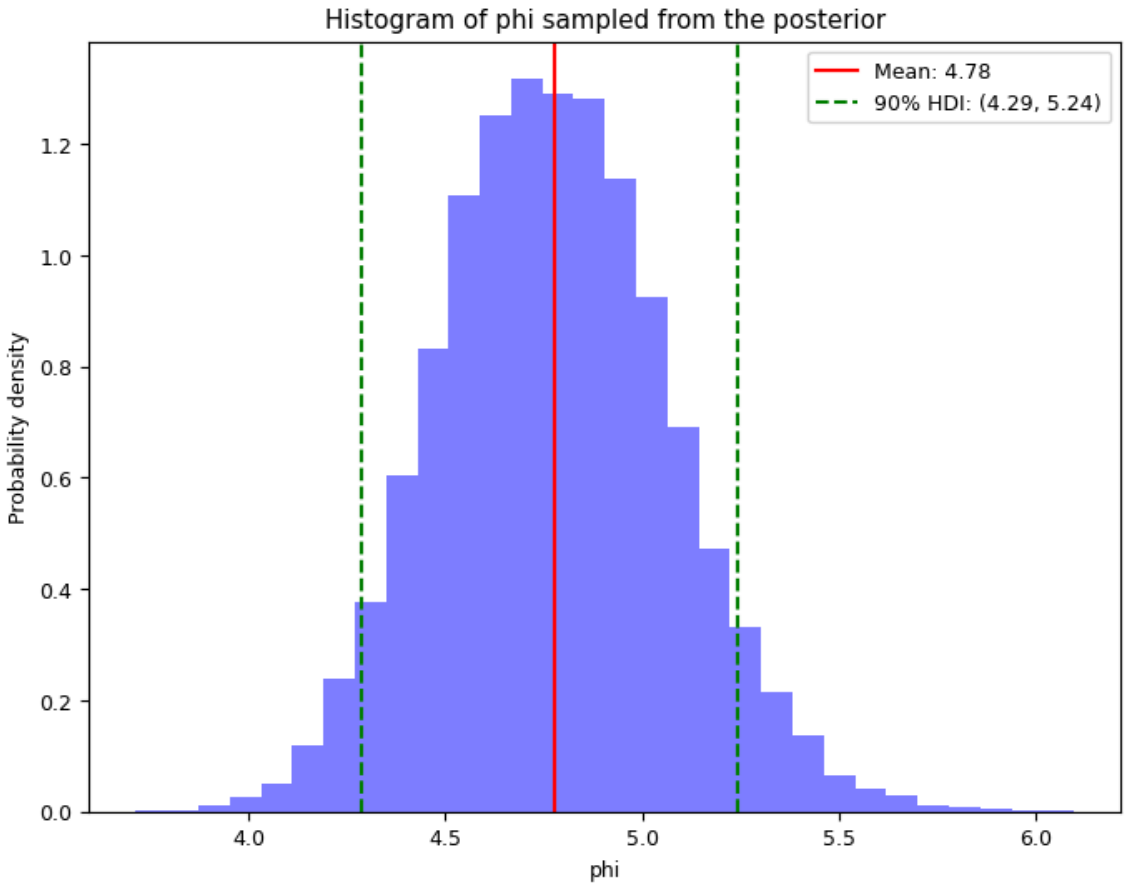
The partial pooling model remain the same in the ZINB model. The only thing changes is that I use the Zero Inflated Negative Binomial likelihood with its dispersion parameter phi.

The prior for the dispersion parameter phi in the Negative Binomial (Gamma-Poisson) distribution reflect the idea that phi controls the variance in the model, specifically the level of overdispersion in the count data. When defining the prior for phi, I let the range of phi to be large enough to adapt to the case when the data exhibit little variance and Var (Y) would be approximately μ since $\frac{\mu^2}{\alpha}$ diminishes

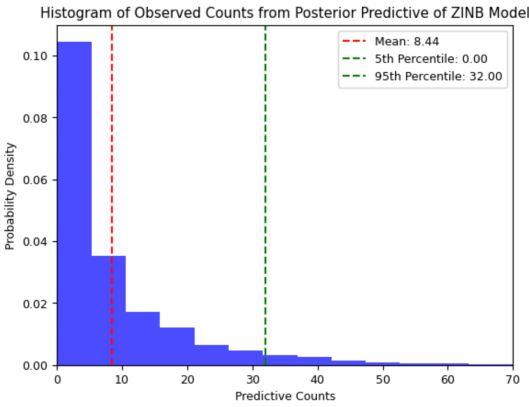
I choose phi ~ Exponential (0.01) so that the mean is 1/ 0.01 = 100 and even allow for greater values than that.

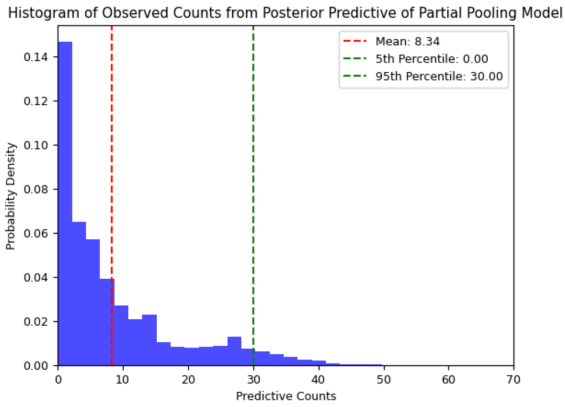
Posterior distribution

The posterior distribution of other parameters does not change significantly. For the posterior distribution of the dispersion parameter phi, we



Posterior predictive check





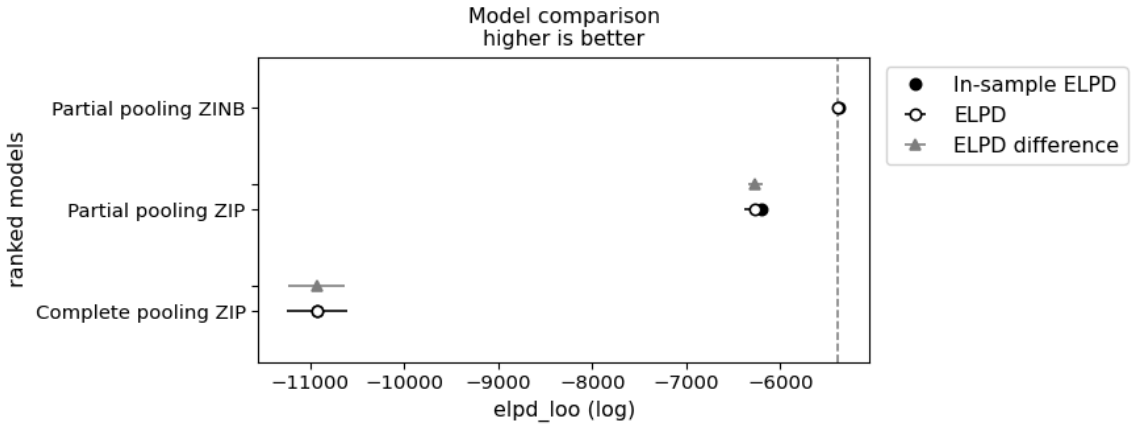
In ZINB model, the mean and 95th percentile has increased slightly from the partial pooling model. The maximum value of posterior predictive samples goes up to 158 while the partial pooling's maximum posterior predictive is 61. The standard deviation is also higher (11.08 vs 9.47).

This indicates that ZINB model is likely shifting its distribution to allow higher counts than what the partial pooling model would predict. It is allowing for greater variance in predictions and the accommodation of extreme values that the partial pooling model may not predict as accurately.

Model comparison

When predicting for existing groups: the ZINB model is most effective, achieving the highest ELPD of -5394 with a weight of 0.997. The ELPD difference between the ZINB model and the Partial pooling ZIP model is 59.99 with no overlap in the ELPD difference error bars of the Partial pooling ZIP model.

In conclusion, the ZINB model provides the best predictive performance (highest elpd_loo) and has the lowest uncertainty in its predictions, as reflected in the lower standard errors. This model balances individual group effects while also accounting for the zero-inflation in the data, making it both accurate and stable.



Predicting missing groups

The majority of the strategy for predicting missing groups stay the same, except for the step when we generate predictive count using the ZINB model.

After the zero-inflated generation process, for the observations that are not zero-inflated, we simulate counts from a **Negative Binomial distribution**. The Negative Binomial distribution is parameterized by two parameters:

1. **r** (shape parameter, related to dispersion) — This is given by the **phi** parameter.
2. **p** (probability of success) — This is derived from the mean **mu** and the shape parameter **r**. The relationship between **mu** and **phi** for the Negative Binomial is:
    - The **mean** (\mu) is related to the **shape** (r) and **probability of success** (p) as:

$$\mu = \frac{r(1-p)}{p}$$

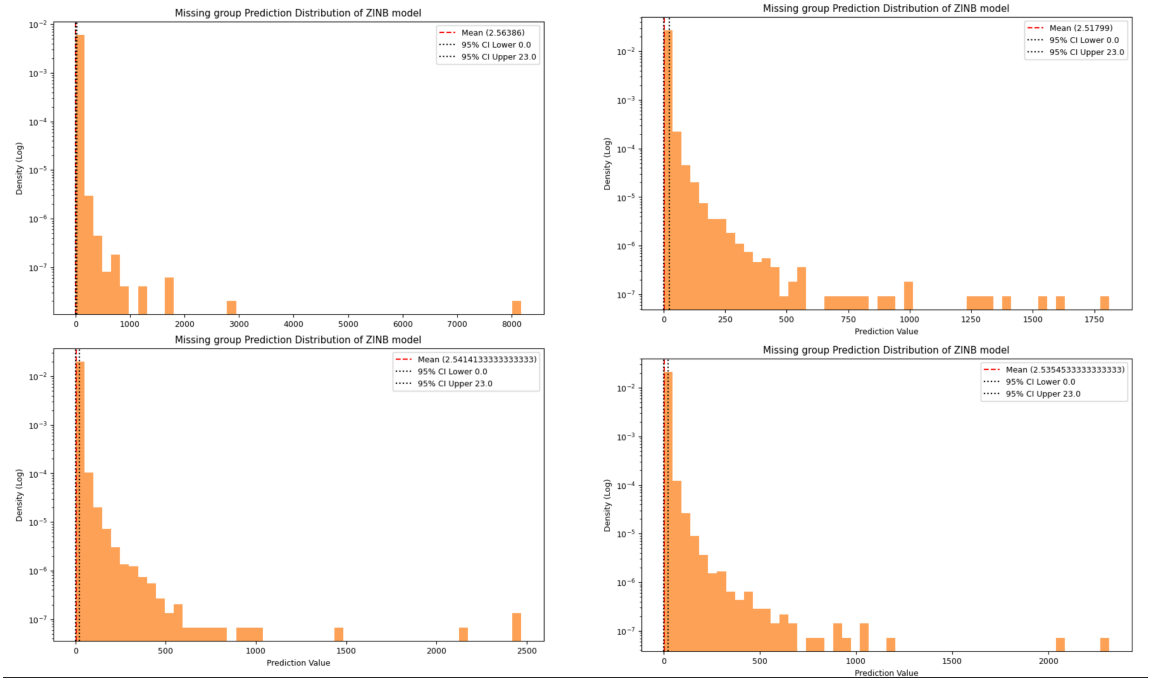
        - Solving for (p), we get:

$$p = \frac{\phi}{\mu + \phi}$$

We use the **Negative Binomial distribution** to generate counts as a random sample from the Negative binomial distribution with n=r, p=p, size=mu.size:

- `n=r` is the shape parameter of the Negative Binomial, which controls the overdispersion.
- `p=p` is the probability of success, controlling how likely it is for an event to occur.
- `size=mu.size` ensures that we generate counts for each observation.

The resulting `nb_counts` array contains simulated counts from the Negative Binomial distribution for the observations that were not zero-inflated.



For ZINB, since the Negative Binomial distribution allows for overdispersion, the predicted counts for a new observation show higher variability: more spread out counts and more frequent large counts than the ZIP model. It still generate excess large counts more than the real data due to the nature of predicting missing data.

Using ZINB for predicting unseen groups is especially suitable for problems when the variation of different groups are highly different.

AI statement: I use AI tool to help visualize different prior distribution and help me choose the suitable one. I also use it to explain the code in the class and explain the ZINB model and ask it to adapt ZINB model for the existing pipeline.

Reference: CS146 session 15, 16, 17

PYMC documentation of Zero Inflated Negative Binomial distribution
https://www.pymc.io/projects/docs/en/stable/api/distributions/generated/pymc.ZeroInflatedN

In [ ]: