**Supplementary materials**

**1. Bayesian workflow of Bayesian zero-inflated negative binomial regression model**

**Step 1: check data**

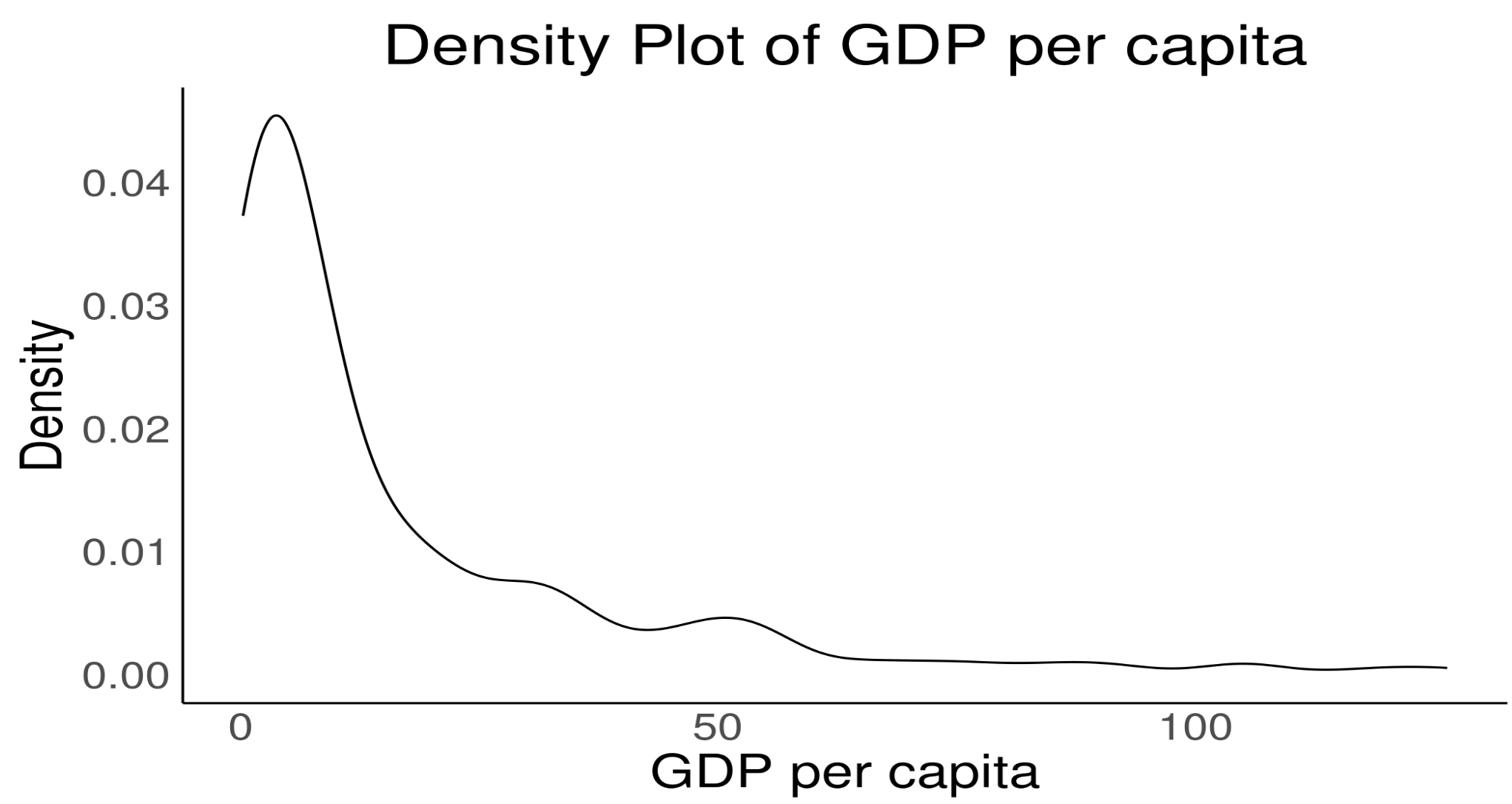
GDP per capita is the independent variable and is continuous data, however, it is not normally distributed, as evidenced by visual inspection (see Figure S1) and the Shapiro-Wilk test (*p* < .001). To address this, we applied a logarithmic transformation to the GDP per capita data.

Figure S1. Density plot of GDP per capita

The sample size is the dependent variable, which is a simple count data that contains many zeros, and non-zero data have over-dispersion. To diagnose overdispersion of count data, we used the model comparison leaf one out cross validation (LOO-CV) method to compare the zero-inflation Poisson model with the zero-inflation negative binomial regression (Winter & Bürkner, 2021). The results show that the zero inflation Poisson model fits worse (*elpd-diff* = -5872.9, *se-diff* = 3133.1), indicating that the data are overdispersion. Therefore, Bayesian zero-inflated negative binomial regression is appropriate to process the data.

**Step 2: Check model parameters**

For zero-inflated negative binomial Regression:

Including 3 parameters:

*μ* is the mean of the negative binomial regression part;

*ϕ* is the shape parameter of the negative binomial regression part;

*p* is the probability of zero.

**Step 3: Set priors**

**(1) The shape parameter *ϕ***

For the shape parameters in Bayesian negative binomial regression, an inv\_gamma(0.4, 0.3) prior is appropriate, as it can cover a broader range (Bürkner, 2024; https://github.com/paul-buerkner/brms/issues/1614). Therefore, we will also use an inv\_gamma (0.4, 0.3) prior for the shape parameters in our Bayesian zero-inflated negative binomial regression model.

**(2) The intercept *β0*** **of**

Based on the self-compiled big team science data, the total sample size, after filtering according to our operational definition, ranges from 1,100,000 to 1,200,000. To ensure that the intercept *β₀* can cover all possible ranges from 1 to 1,200,000, we consider two priors: a more diffuse prior, *β₀* ~ Normal (0, 4.67), and a more concentrated prior, *β₀* ~ Normal (6, 2.67).

Using the properties of the log-normal distribution, the *β₀* ~ Normal (0, 4.67) prior covers a range from about 8.23e-07 () to 1,214,691 (), while the β₀ ~ Normal (6, 2.67) prior ranges from about 0.13 () to 1,214,691 (). Both priors effectively cover the target range of 1 to 1,200,000.

We have made a visual comparison between the intercept priors *β₀* ~ Normal (6, 2.67) and β₀ ~ Normal (0, 4.67) (see figure below). Notably, the prior *β₀* ~ Normal (0, 4.67) is more extreme, with a median of e⁰ = 1, whereas the median for *β₀* ~ Normal (6, 2.67) is e⁶ = 403.

Given that one big team science study sampled from 187 countries, and that there are approximately 230 countries and regions worldwide, it is reasonable to conclude that all big team science samples we ultimately collected represent the majority of countries worldwide. As the participants in most big team science studies come from dozens of countries, countries representing the median sample size in the final data are likely to appear less frequently in big team science. For this median, sample sizes in the range of a few hundred are appropriate. Therefore, *β₀* ~ Normal (6, 2.67) is the more appropriate prior.

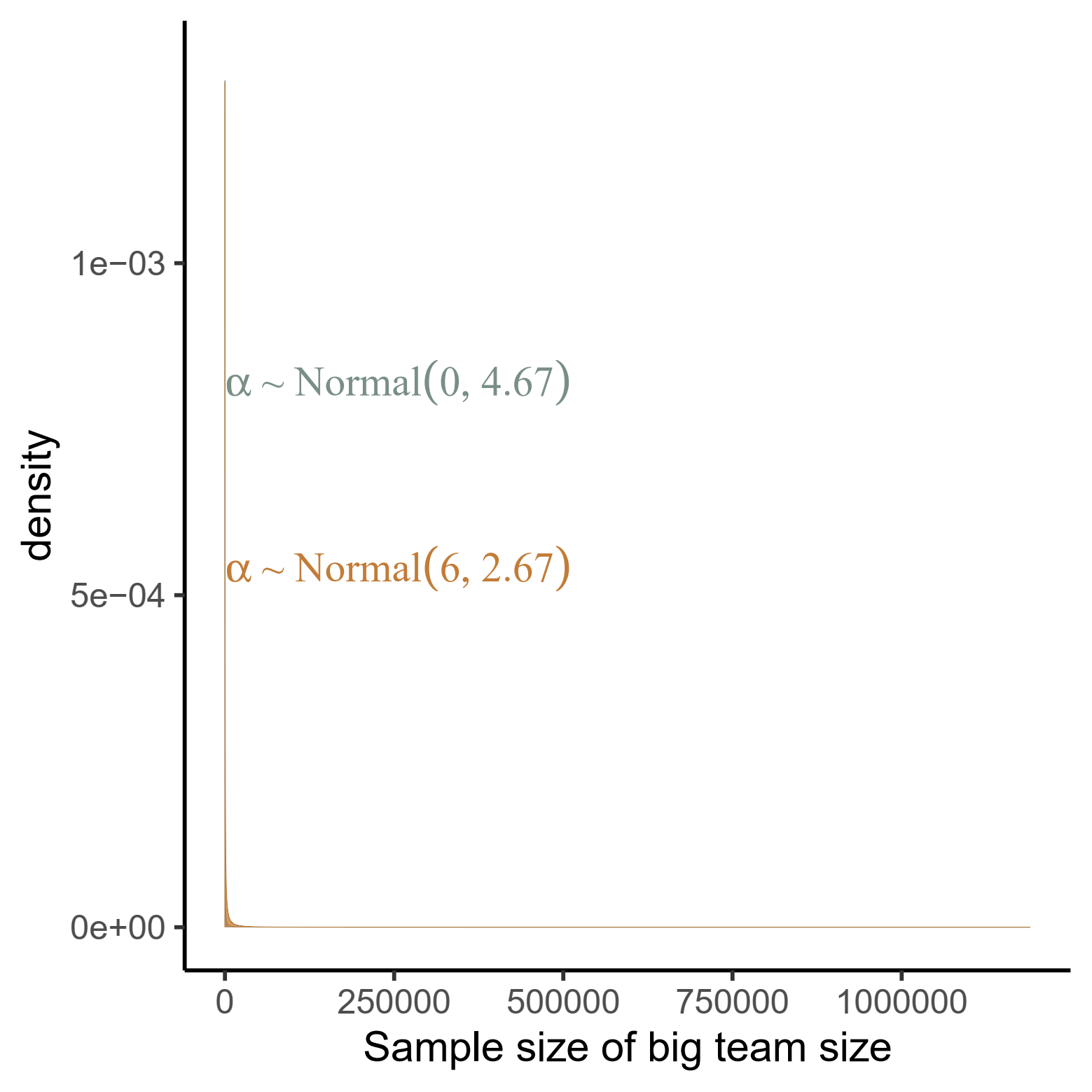
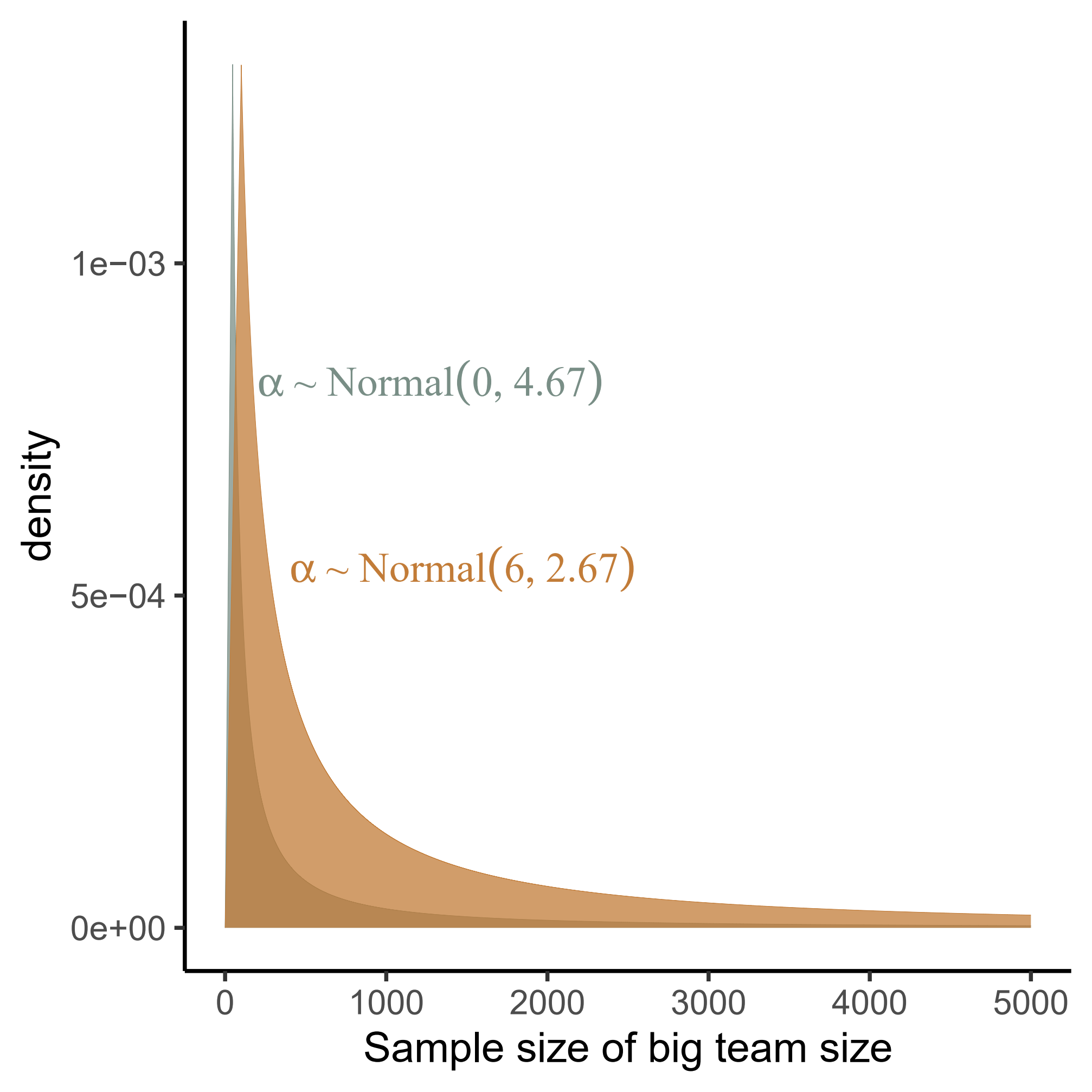
 

Figure S2. Prior prediction checking of *β₀*. The right panel shows the same data as the left panel, but with the range of the x-axis restricted to better highlight the differences between the two distributions.

**(3) The intercept γ0 of**

The probability of 0 in our data may be mostly concentrated between 0.1-0.9, and after logit transformation it is (-2.2, 2.2). We set γ0 ~ Normal (0, 1), which corresponds exactly to this range.

**(4) The slope *β1* ofand** **the slope *γ1*of**

For the slope parameters *β1* and *γ1*, we use the changing slope to plot the prior prediction distribution and then select the appropriate prior (For detailed information, please refer to https://github.com/Chuan-Peng-Lab/BTS\_Sample\_Stage1\_RR/tree/master/4\_Analyses/4\_2\_Exploratory\_analysis/Bayesian\_Zero-inflated\_negative\_binomial\_Regression).

The prior selection can meet two criteria (as shown below): (1) it is within the possible range of values (there are values on the x-axis, and the sample size can take the larger value of 1,200,000); (2) the change in the data is not extreme.

Based on the above criteria, the priors we finally chose were *β1* ~ Normal (0, 0.1) & *γ1* ~ Normal (0, 0.1).

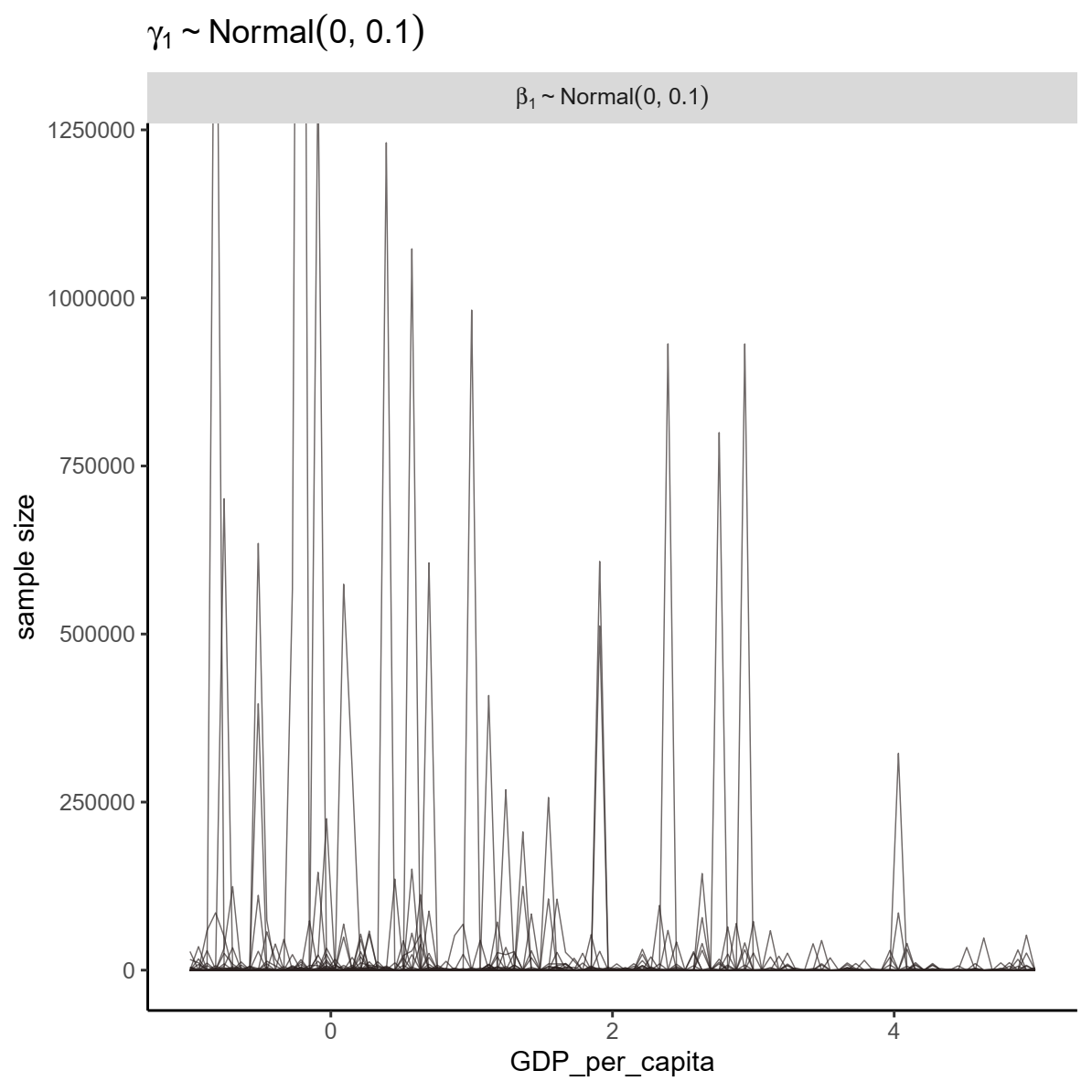
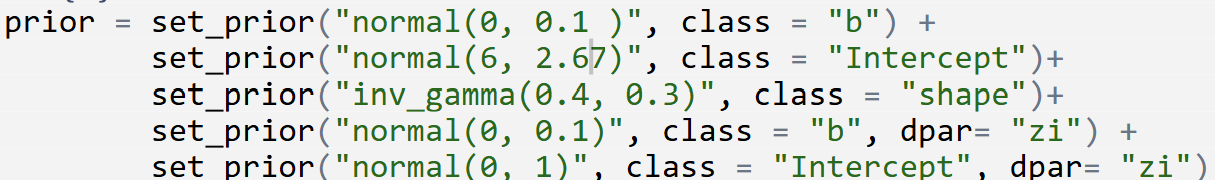


Figure S3. Prior prediction checking of *β₀* & *γ1*.

In summary, the priors we use are as follows:



Next, we will use this prior for further prior prediction checks.

**Step 4:** **Prior prediction checking**

The visualization of prior predictive checks is presented in Figure S4. In the left panel we can see that the sample size can accommodate values up to 1,200,000. In the right panel, the actual data fall within the range of the prior distribution, demonstrating the appropriateness of the prior.

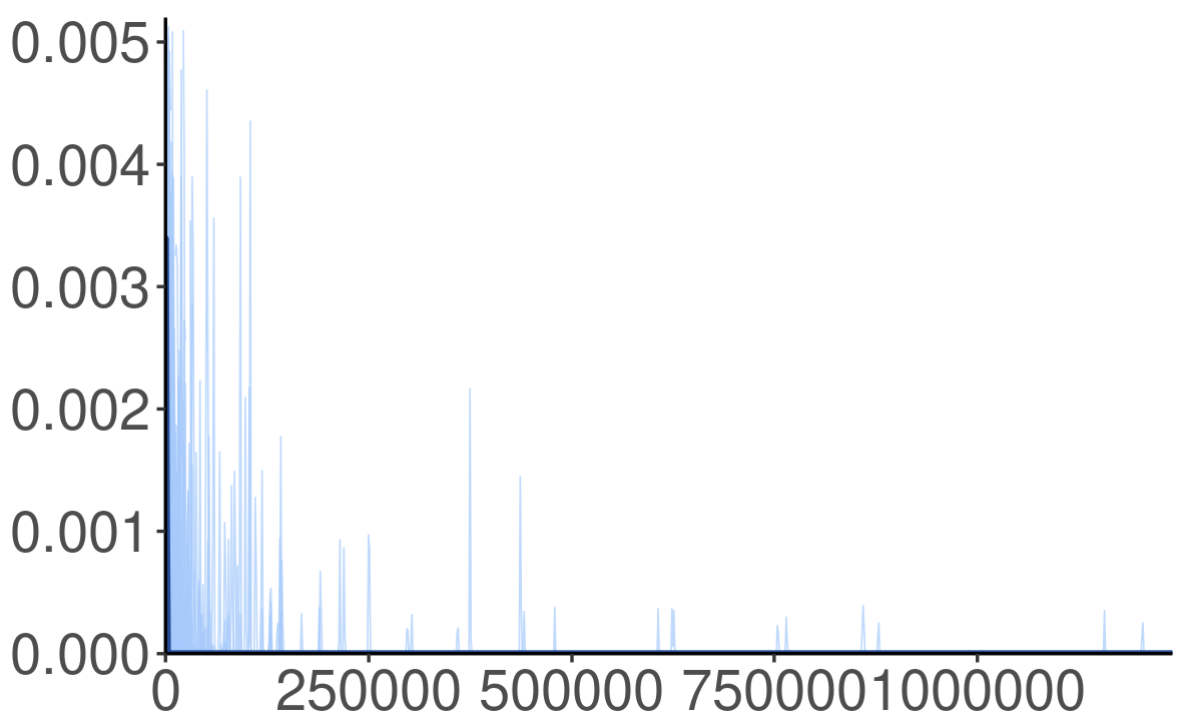
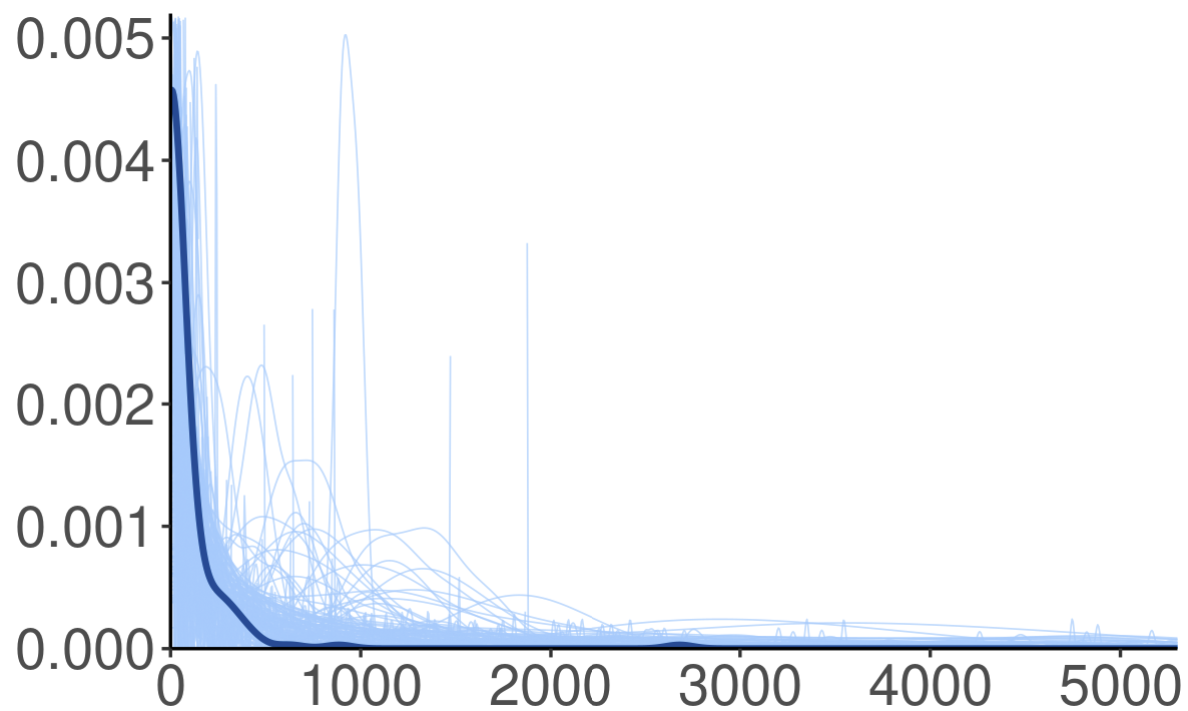
 

Figure S4. Prior prediction checking that includes all parameters. The right panel shows the same data as the left panel, but with the range of the x-axis restricted to better highlight actual data.

**Step 5: Model operation results**

Interpreting results:

(1) Intercept = 5.43 means that when log (GDP per capita) = 0, i.e. GDP per capita = 1 ($1000), the model predicts a sample size of exp (5.15) = 95.58.

(2) log\_GDP\_per\_capita = 0.34 means that when GDP per capital increases by 1%, the model’s predicted sample size increases by 0.34%.

(3) Zi\_intercept = 1.38 means that when log (GDP per capita) = 0, i.e. GDP per capita=1 ($1000), the probability of the model predicting a sample size of 0 is plogis (1.38) = 79.90%.

(4) Zi\_log\_GDP\_per\_capita = -0.09 means that when GDP per capita increases by 1%, the probability of the model predicting a sample size of 0 decreases by 0.08%.

(5) Shape = 1.33 indicates excessive dispersion of the sample size data. Based on the variance formula below, the closer the shape parameter is to positive infinity, the smaller the data dispersion, and the closer the shape parameter is to 0, the larger the data dispersion.

**Step 6: Posterior prediction checking**

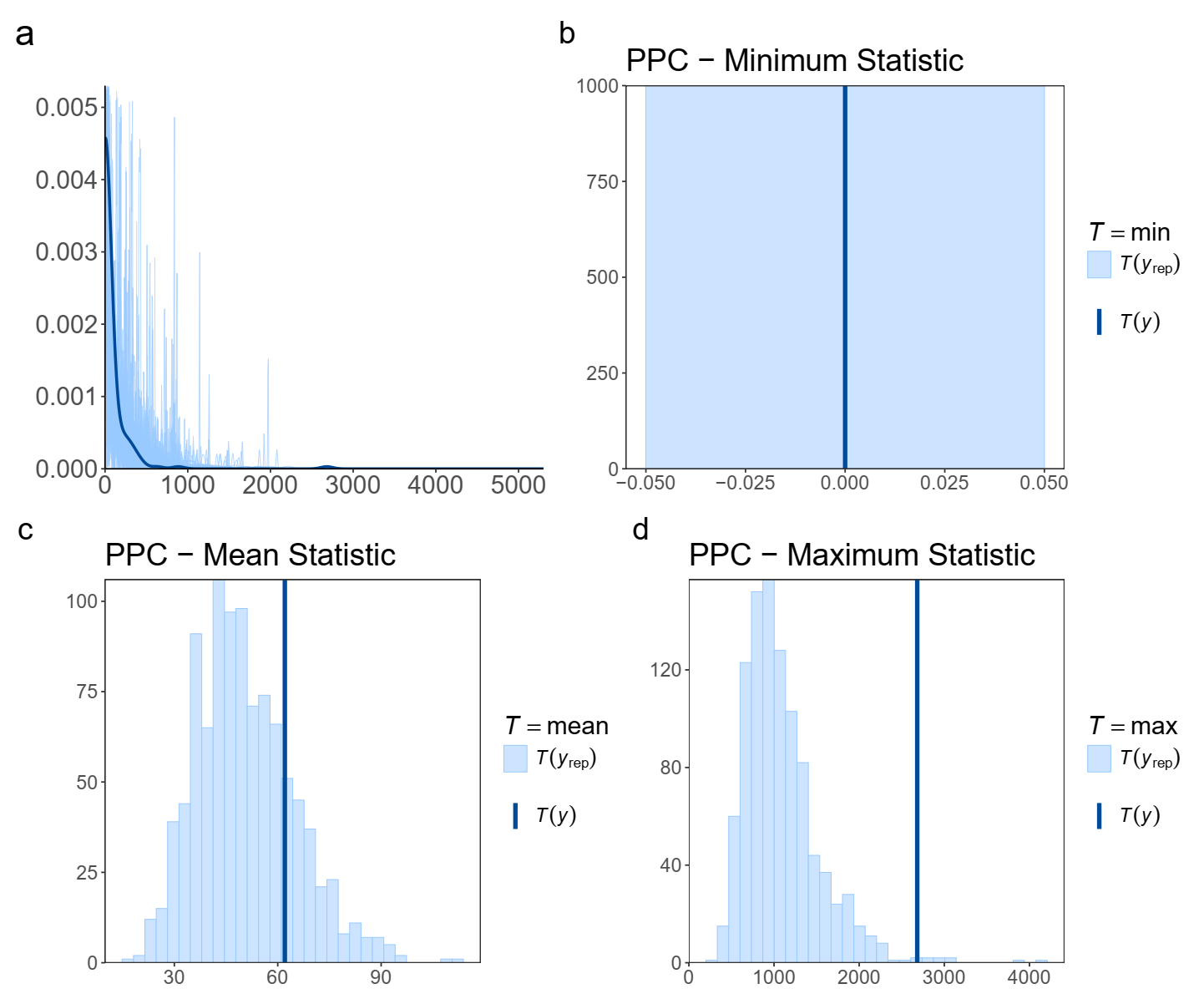
First, as shown in panel (a) of Figure S5, the actual data are well contained within the posterior predictive distribution. In addition, panel (b) shows that the minimum value of the actual data, zero, is exactly at the center of the posterior predictive distribution. Panel (d) shows that although the maximum value of 2,682 is an outlier (since all other values are below 1,000) and does not fall entirely at the center of the posterior predictive distribution, the distribution sufficiently covers this extreme value, making it acceptable. In other words, the posterior predictive distribution adequately covers the range of values in the actual data. Panel (c) shows that the mean of the actual data is approximately at the center of the posterior predictive distribution, indicating that the distribution adequately represents the central tendency of the actual data. Overall, this suggests that our model effectively describes the actual data.

Figure S5. Prior prediction checking

In summary, the prior we have set is appropriate based on current knowledge.