

A Bayesian Hierarchical Modeling Study on Factors Influencing Male Fetal Y-Chromosome Concentration

Jinchao Gai*

School of Mathematics and Statistics, Shandong University, Weihai 264209, China.

* Corresponding author: Jinchao Gai (Email: 202300820030@mail.sdu.edu.cn)

Abstract: This study develops a Bayesian hierarchical model to analyze factors influencing male fetal Y-chromosome concentration (Y%), crucial for Non-Invasive Prenatal Testing (NIPT) accuracy. Initially, Spearman's correlation analysis identified 13 significant indicators, including gestational week (positive correlation) and maternal BMI/weight (negative correlation). Given Y%'s non-negative, right-skewed, and long-tailed distribution, distribution fitting confirmed its alignment with a Rice distribution. A Bayesian hierarchical model was constructed incorporating these indicators as fixed effects, along with individual maternal random effects to account for heterogeneity. Parameters were estimated via MCMC sampling (NUTS algorithm, 300,000 iterations), with all chains converging effectively ($R < 1.01$). Results confirmed that gestational week positively affects Y%, while maternal BMI and weight negatively impact it, consistent with physiological patterns. This model provides a robust framework for optimizing NIPT timing and reliability.

Keywords: Spearman's Rank Correlation Test, Bayesian Hierarchical Model, Rice Distribution.

1. Introduction

Non-invasive prenatal testing (NIPT) has become a critical tool for detecting fetal chromosomal abnormalities by analyzing cell-free DNA in maternal blood, with its accuracy heavily dependent on the concentration of the fetal Y-chromosome (Y%). A key challenge in NIPT is ensuring reliability[1-2], particularly when testing is conducted too early or too late, as low Y% levels may lead to misdiagnosis while increasing clinical risks and costs. This study aims to address these issues by quantitatively analyzing the relationship between Y% and influencing factors such as gestational week and maternal BMI, thereby providing insights for optimizing testing protocols. The research innovatively integrates a Bayesian hierarchical framework with a Rice distribution assumption, which effectively captures the non-negative, right-skewed, and long-tailed characteristics of Y% data. By incorporating medical prior knowledge and individual random effects, the model accounts for heterogeneity across samples, enhancing both interpretability and robustness. Methodologically, the study first employs Spearman's rank correlation tests to identify significant monotonic relationships, confirming positive correlations between Y% and gestational week and negative correlations with maternal BMI and weight[3-4].

The distribution fitting results validated the Rice distribution as an appropriate model for Y%, supported by a Kolmogorov-Smirnov test p-value of 0.58. Subsequently, a Bayesian hierarchical model was constructed with 13 significant indicators as fixed effects and maternal individual random effects, estimated through MCMC sampling (NUTS algorithm) with 300,000 iterations.

Convergence diagnostics confirmed model stability ($R < 1.01$), and posterior analysis revealed that gestational week, maternal weight, and BMI are dominant factors affecting Y%, aligning with established physiological patterns. This study not only advances the statistical modeling of NIPT-related parameters but also offers practical guidance for clinical timing and risk assessment.

2. Model Development and Solution

2.1. Spearman Correlation Test

The data is from <https://www.mcm.edu.cn/>. First, a normality test was conducted on the data, and most of the data failed to meet the normality assumption. Therefore, the Spearman rank correlation coefficient test—which does not require a normality assumption—was used to qualitatively observe the correlation between the Y-chromosome concentration and other indicators[5].

Further analysis of the distribution of Y-chromosome concentration revealed that it exhibits non-negative, right-skewed, and long-tailed characteristics, consistent with the Rice distribution. Based on this, combined with medical background knowledge, explanatory variables with strong significance were selected, and individual random effects were considered to establish a Bayesian hierarchical model based on the Rice distribution. Markov Chain Monte Carlo (MCMC) sampling technique was used for parameter estimation, and the significance of the model and the effect size of each indicator were analyzed through credible intervals and other methods[6-7].

The Spearman rank correlation coefficient is a non-parametric statistic used to measure the strength and direction of a monotonic relationship between two variables. Compared with the Pearson correlation test, it does not rely on the distribution assumption of variables. Instead, it characterizes the correlation pattern of variables by constructing rank statistics, which can better capture non-linear relationships between variables. The formula for calculating the rank statistic is as follows:

$$r_s = \left[\frac{1}{n} \sum_{i=1}^n r_i R_i - \left(\frac{n+1}{2} \right)^2 \right] / \left(\frac{n^2-1}{12} \right) \quad (1)$$

Where $r_1, r_2, \dots, r_n$ are the rank statistics derived from $X_1, X_2, \dots, X_n$, and $R_1, R_2, \dots, R_n$ are the rank statistics derived from $Y_1, Y_2, \dots, Y_n$.

Through the Spearman correlation test, the correlation characteristics between various variables can be qualitatively understood, providing guidance for the subsequent

establishment of a relationship model. Correlation between Y-Chromosome concentration and other indicators is shown in table 1.

Table 1: Correlation between Y-Chromosome Concentration and Other Indicators

	Age	Height	Weight
Y%	-0.1172	-0.1011	-0.1670
p-value	0.0001	0.0009	0.0000
	Gestational Week of Test	Maternal BMI	Raw Read Count
Y%	0.0803	-0.1550	-0.0860
p-value	0.0082	0.0000	0.0046
	Proportion Mapped to Reference Genome	Proportion of Duplicate Reads	Number of Uniquely Mapped Reads
Y%	-0.1594	0.1058	-0.0694
p-value	0.0000	0.0005	0.0224
	GC Content	Z-Value of Chromosome 13	Z-Value of Chromosome 18
Y%	-0.0166	-0.0766	-0.1782
p-value	0.5844	0.0117	0.0000
	Z-Value of Chromosome 21	Z-Value of X-Chromosome	Z-Value of Y-Chromosome
Y%	0.0335	-0.0559	0.1135
p-value	0.2704	0.0661	0.0002
	X-Chromosome Concentration	GC Content of Chromosome 13	GC Content of Chromosome 18
Y%	0.4744	-0.0372	-0.0499
p-value	0.0000	0.2213	0.1007
	GC Content of Chromosome 21	Proportion of Filtered Reads	Number of Pregnancies
Y%	-0.0107	0.0860	-0.0072
p-value	0.7253	0.0046	0.8137

Note: Data marked in red indicates a p-value less than the significance level of 0.01

At a significance level of $\alpha = 0.01$, the marked data in the table shows a certain correlation with the Y-chromosome concentration. Among them, indicators such as the gestational week of the test and the Z-value of the Y-chromosome have a positive correlation with it; indicators such as maternal BMI, weight, and GC content have a negative correlation with it.

2.2. Bayesian Hierarchical Model

To establish a relationship model between Y-chromosome concentration and other indicators, this study adopted the Bayesian method and incorporated medical prior knowledge to obtain more effective estimates. This decision was based on the following considerations: significant individual differences exist in variables such as maternal age, BMI, and gestational week of the test; multiple test records are available for the same pregnant woman; the sample size of the test data is small with high heterogeneity; and the accuracy and validity of the data are difficult to judge.

2.2.1. Prior Distribution of Y-Chromosome Concentration

To better perform inference using the Bayesian method, it is necessary to characterize the prior distribution of the fetal Y-chromosome concentration. For this purpose, the FITTER library was used to iterate through possible distributions and fit their parameters. Finally, it was found that the distribution of Y-chromosome concentration follows the Rice distribution. Its Cumulative Distribution Function (CDF) is:

$$f(y|v, \sigma) = \frac{y}{\sigma^2} \exp\left(-\frac{y^2+v^2}{2\sigma^2}\right) I_0\left(\frac{yv}{\sigma^2}\right), y \geq 0 \quad (2)$$

Where: y : Observed value (non-negative); $v \geq 0$: Location parameter; $\sigma > 0$: Scale parameter; $I_0(\cdot)$: Modified Bessel Function of the First Kind (order 0), defined as:

$$I_0(z) = \sum_{k=0}^{\infty} \frac{1}{(k!)^2} \left(\frac{z}{2}\right)^{2k} \quad (3)$$

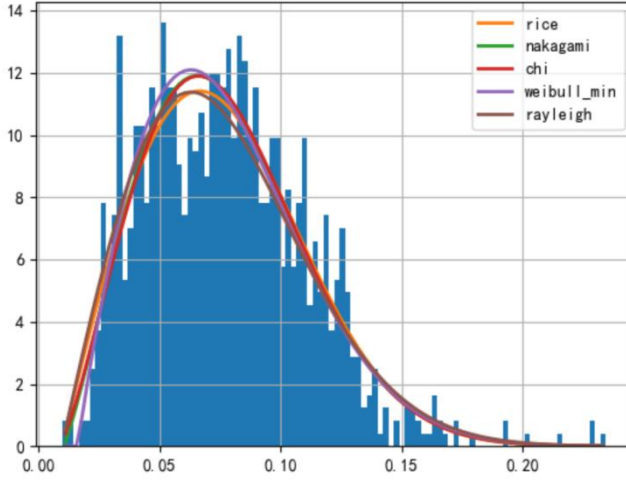


Figure 1: Distribution Fitting Effect

According to figure 1 and figure 2, this distribution conforms to the non-negative, right-skewed, and long-tailed characteristics of the data. Moreover, the p-value of the Kolmogorov-Smirnov (KS) statistic (ks-pvalue = 0.58) is significantly greater than the significance level $\alpha = 0.05$ (under the null hypothesis H_0 : the data does not conform to the Rice distribution). Therefore, the fitting effect is satisfactory.

2.2.2. Model Establishment

Based on the Spearman correlation coefficient analysis and existing research results (used as prior knowledge), this study selected 13 indicators with significant Spearman coefficients (including gestational week, BMI, etc.) as explanatory variables for the Bayesian mixed hierarchical model. Meanwhile, according to Assumption 3.2, individual random effects of different pregnant women were considered, and these effects were assumed to come from a normal population, i.e.:

$$u_{s[i]} \sim \mathcal{N}(0, \sigma_u^2), s = 1, 2, \dots, S \quad (4)$$

Where $u_{s[i]}$ represents the random effect of the group $s[i]$ to which the i -th sample belongs, used to capture between-group heterogeneity (e.g., differences in equipment, hospitals, test batches, etc.).

It is assumed that the i -th observed value y_i follows the Rice distribution, and its probability distribution is:

$$y_i \sim \text{Rice}(v_i, \sigma), i = 1, 2, \dots, N \quad (5)$$

The location parameter v_i is determined by both fixed effects and random effects, and its absolute value is taken to ensure non-negativity:

$$v_i = |\alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_{s[i]}| \quad (6)$$

Where β_k represents the regression coefficient of the

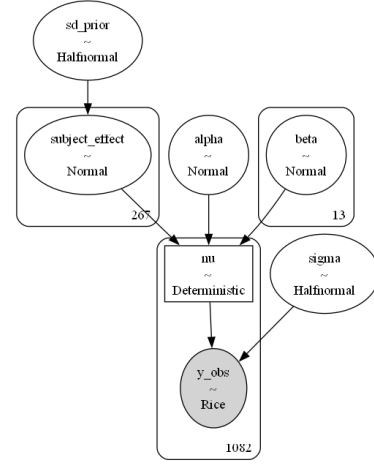


Figure 2: Bayesian Inference Structure Diagram

explanatory variable x_{ik} .

2.2.3. Prior Parameter Setting

A hierarchical Bayesian framework was adopted, and weak-information priors were set for all parameters:

$$\alpha \sim \mathcal{N}(0, 4), \sigma_u \sim \text{Half-Normal}(1), \sigma \sim \text{Half-Normal} \quad (7)$$

Notes:

- Normal priors were used for α and β_k to achieve parameter shrinkage and prevent overfitting;
- Half-normal distributions were used for σ_u and σ to reflect the non-negativity of noise parameters;

2.2.4. MCMC Sampling

In Bayesian inference, it is usually difficult to obtain an exact posterior distribution. Therefore, the MCMC sampling technique was used to draw samples from one or more Markov chains that converged to a stationary distribution, and then estimate the model parameters.

For the established Bayesian hierarchical model, programming was performed using the PyMC library, and the No-U-Turn Sampler (NUTS) algorithm was used for sampling. Given the high complexity of the model, 10,000 warm-up iterations and 300,000 sampling iterations were conducted to ensure that the MC chains converged to a stationary state.

2.3. Model Evaluation

2.3.1. Convergence Diagnosis

Currently, there is no simple and effective method for diagnosing the convergence of sampling using the MCMC method. By observing the trace plots of the MC chains, they showed a fluctuation pattern around the mean, with no unusual or non-stationary patterns. Additionally, according to the Gelman-Rubin diagnostic method, the $\hat{R}$ values of all parameters were calculated and found to be strictly less than 1.01, indicating that the MC chains had converged. Trace plots of parameters are shown in figure 3.

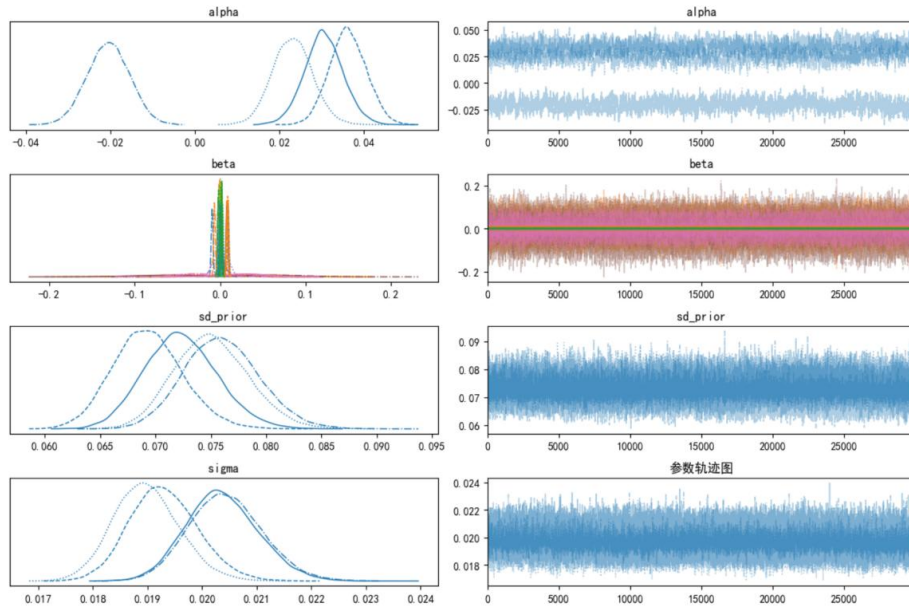


Figure 3: Trace Plots of Parameters (Alpha, Beta, SD_Prior, Sigma)

2.3.2. Significance Test

Significance testing under the frequentist framework can only provide a binary conclusion of "significant" or "not significant", and the magnitude of the p-value cannot represent the actual strength of the effect. Under the Bayesian framework, the best point estimate is given by the posterior mean, and the credible interval of the posterior distribution is also provided. When evaluating significance, it is necessary to check the inclusion relationship between zero and the 95% Highest Density Interval (HDI). If the 95% HDI includes 0, it indicates insufficient evidence; otherwise, it can be considered that there is significant evidence. Furthermore, the

probability of parameters crossing zero and the width of the 95% HDI were calculated to evaluate the significance and effect size of each parameter (see Table 2).

Maternal weight, BMI, and gestational week of the test all had strong effects on the Y-chromosome concentration. When weight or BMI increased, the Y-chromosome concentration decreased, which is consistent with the conclusions of relevant studies; at the same time, the Y-chromosome concentration increased with the increase in gestational week, which conforms to normal physiological laws. Parameter Estimate and Effect Size are shown in table 2.

Table 2: Parameter Estimates and Effect Sizes

Parameter	Mean	Effect Size
Weight	-1.105	0.213
Maternal BMI	-0.85	0.172
Random Effect	0.73	0.146
Intercept	-0.16	0.074
Gestational Week of Test	0.20	0.022
X-Chromosome Concentration	0.12	0.021
Age	0.01	0.020
Raw Read Count	0.01	0.008
Height	-0.01	0.007
Proportion of Filtered Reads	0	0.006
Number of Uniquely Mapped Reads	0	0.005
GC Content	0	0.005
Chromosome Z-Value	0	0.004
Proportion of Duplicate Reads	0	0.003
Proportion Mapped to Reference Genome	0	0.003

3. Conclusions

This study successfully constructed a Bayesian hierarchical model based on the Rice distribution to analyze the relationship between male fetal Y-chromosome concentration and multiple maternal indicators. Key influencing factors, such as gestational week, BMI, and weight, were identified through Spearman's correlation tests. Model evaluation confirmed the significant effects of maternal weight, BMI, and gestational week on Y-chromosome concentration, with Y% increasing as gestational week advances and decreasing as weight or BMI increases. These findings align with known physiological patterns and medical research. By adopting a Bayesian framework, this study effectively incorporated medical prior knowledge, enhancing the stability of parameter estimation and the rationality of the model. Based on the findings and limitations of this study, expanding the sample size and incorporating multi-center data would enhance the generalizability of the Bayesian hierarchical model. This could help validate the observed relationships across diverse populations and reduce potential biases from single-center studies. Second, exploring additional distributions beyond the Rice distribution—such as the Weibull or Gamma distributions—may provide further insights into the nature of Y-chromosome concentration data, especially if more complex skewness or kurtosis patterns are observed in larger datasets.

References

- [1] Zhang Liangliang, Zhuo Zhaozhen, Huang Shengwen, et al. Comparative Analysis of the Application Value of NIPT and NIPT-plus in Prenatal Screening [J]. *Guizhou Medicine*, 2025, 49(06): 942-944.
- [2] Zhang Zhenzhen, Chen Weiqing, Ding Ying. Application Value of Non-Invasive Prenatal Testing Combined with Beta-Subunit of Human Chorionic Gonadotropin and Serum Pregnancy-Associated Protein A in Predicting Fetal Chromosomal Abnormalities in Pregnant Women at 9–13+6 Weeks Gestation [J]. *Chinese Journal of Maternal and Child Health*, 2025, 40(12): 2165-2168.
- [3] Yu Dandan, Li Fengjin, Yao Xinyu, et al. Efficacy Analysis of NIPT in Detecting Common Fetal Chromosomal Aneuploidies Among Different High-Risk Populations [J]. *Chinese Journal of Family Planning and Obstetrics & Gynecology*, 2025, 17(03): 72-76+82.
- [4] Qian Yue, Jin Keqin, Hu Min, et al. Screening Value of Expanded Non-Invasive Prenatal Testing for Pregnant Women at High Risk for Down Syndrome [J]. *Zhejiang Clinical Medicine*, 2025, 27(02): 180-183.
- [5] Qian Wenjun, Zhang Li, Yang Tong, et al. Application and Efficacy Analysis of Non-Invasive Prenatal DNA Testing in Fetal Sex Chromosome Aneuploidy Screening [J]. *Chinese Journal of Maternal and Child Health*, 2025, 40(03): 389-393.
- [6] Liu Jianzhen, Chen Hongzhen, Meng Xiangrong, et al. Application of Non-Invasive Prenatal Screening in Chromosomal Aneuploidy Screening for Twin Pregnancies and Analysis of Fetal Cell-Free DNA Concentration [J]. *Chinese Journal of Prenatal Diagnosis (Electronic Edition)*, 2023, 15(04): 22-26.
- [7] Wang Xuechun. Effects of Different Concentrations of Colchicine on Root Morphological Characteristics and Chromosomes of Fava Beans [J]. *Horticulture and Seedlings*, 2023, 43(01): 25-27.