MINI Project

Shuo Han

2023-02-28

Introduction

Dataset

We consider a study investigated whether juvenile delinquency among males is related to brith order and 1137 boys are seperated into two groups with four categroies.

DATAset	Oldest	In-Between	Youngest	Only child	Total
Most delinquency	127 (26.9%)	123 (37%)	93 (37.1%)	17 (20.7%)	360 (31.7%)
Least delinquency	345 (73.7%)	209 (63%)	158 (62.9%)	65 (79.3%)	777 (68.3%)
Total	472	332	251	82	1137

H0 : No relationship between delinquency and birth order (1)

H1 : There is relationship between delinquency and birth order (2)

Analysis

Such Hypothesis (1) and (2) can be written in terms of odds ratios with following proporsition:

Proposition 1: Under product-multinomial sampling $p_{1j} = ... = p_{Ij} > 0$ for all $j = 1, ..., J$ if and only if

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{j'i}} = 1$$

for all $i, i' = 1, ..., I, and, j, j' = 1, ..., J$

Proposition 2: Under product-multinomial sampling $0 < p_{ij} = p_{i\cdot}p_{\cdot j}$ for all $j = 1, ..., J$ and all $i = i, ...I$ if and only if

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{j'i}} = 1$$

for all $i, i' = 1, ..., I, and, j, j' = 1, ..., J$

Proposition 3:

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{j'i}} = 1$$

for all $i, i' = 1, ..., I, and, j, j' = 1, ..., J$ if and only if $p_{11}p_{ij}/p_{1j}p_{i1} = 1$ for all $i. = 1, j. = 1$

Startup

Then we continue with our data, From proposition , it is sufficient to examine:

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = 127(209)/123(345) = 0.6254978$$

$$SE = \sqrt{1/n_{11} + 1/n_{22} + 1/n_{12} + 1/n_{21}} = 0.1539069$$

$$\frac{n_{11}n_{23}}{n_{13}n_{21}} = 127(158)/93(345) = 0.6254013$$

$$SE = \sqrt{1/n_{11} + 1/n_{23} + 1/n_{13} + 1/n_{21}} = 0.1668963$$

$$\frac{n_{11}n_{24}}{n_{14}n_{21}} = 127 * (65)/17 * (345) = 1.407502$$

$$SE = \sqrt{1/n_{11} + 1/n_{24} + 1/n_{14} + 1/n_{21}} = 0.2915145$$

Restate the hypothesis

Then H0 (1) can be expressed as :

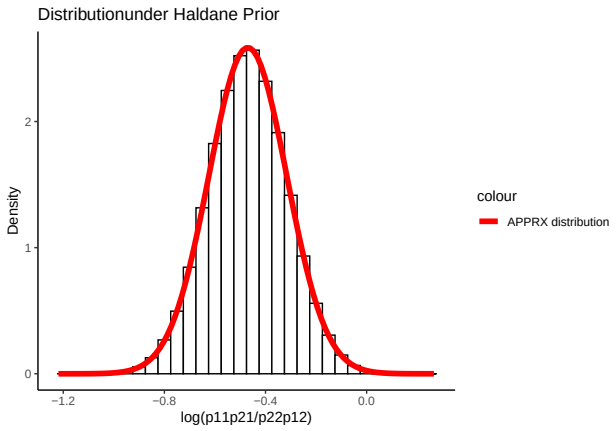
H0 : $log(\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{j'i}}) = 0$ **for all** $i, i' = 1, ..., I, and, j, j' = 1, ..., J$

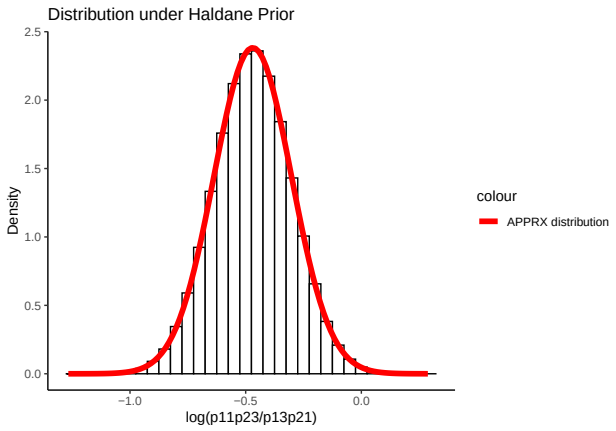
Bayesian Method – Haldane Prior

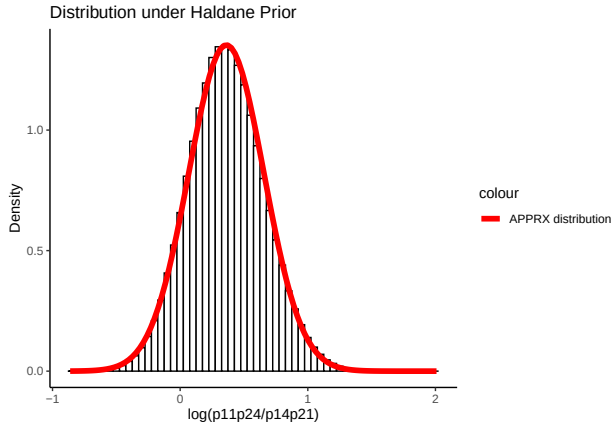
Prior: $P(p1.) \sim Dir(0, 0, 0, 0)$ and $P(p2.) \sim Dir(0, 0, 0, 0)$

Likelihood Function: $L(p1.|data) \propto \prod_{i=1}^{4} p_{1i}^{n1i}$ and $L(p2.|data) \propto \prod_{i=1}^{4} p_{2i}^{n2i}$

Then posterior can be given as : $P(p1.|data) \propto Dir(n11, n12, n13, n14)$ and $P(p2.|data) \propto Dir(n21, n22, n23, n24)$







From the plot we can see the normal approximations can approximate the posterior distribution of the log odds quite well. So we can also generate samples from the normal distribution to approximate log odd ratios in th future.

Results of Haldane Prior

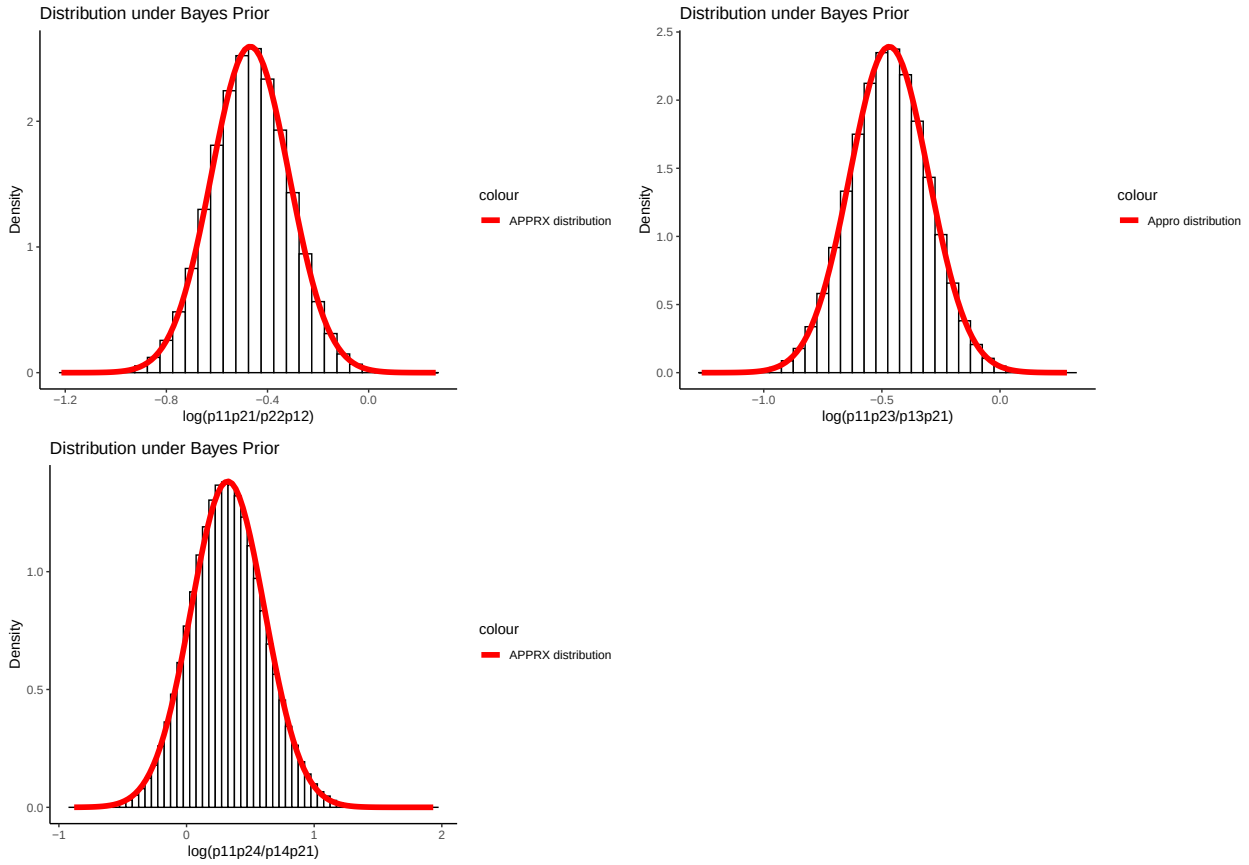
To this end, we use the large sample together with simulation from Haldane Prior calulated standard errors for the log odds ratios. Testing whether the log odds ratios are different from zero, we get the confidence interval for n = 10000, 1e5 and 1e6.

odds ratio	log(odds)	CI(n=10000)	n=100000	n = 1000000
0.625	-0.4692	[-0.7713601 -0.1731035]	[-0.77136,-0.1701]	[-0.2157858 0.9835972]
0.625	-0.4693	[-0.7974688 -0.1452485]	[-0.7974688,-0.1452485]	[-0.7972792 -0.1411580]
1.408	0.3418	[-0.2157858 0.9835972]	[-0.1991654, 0.9639039]	[-0.1980471 0.9586218]

Under Haldene prior, we can observe that 0 is not included in all of the 95% confidence intervals of log odds ratios, only included in the third ratio. This indicates that the rows and columns of the table are not independent, and we should conclude that there is relationship between delinquency and birth order.

Increasing N will make the histogram more closely resemble a normal distribution, but this will not substantially alter the estimates of means and standard deviations.

Bayesian Method – Bayes Prior



From the plot we can see the normal approximations can approximate the posterior distribution of the log odds quite well. So we can also generate samples from the normal distribution to approximate log odd ratios in th future.

Results of Bayes Prior

To this end, we use the large sample together with simulation from Bayes Prior calulated standard errors for the log odds ratios. Testing whether the log odds ratios are different from zero,we get the confidence interval for n = 10000, 1e5 and 1e6:

odds ratio	log(odds)	CI(n=10000)	n=100000	n = 1000000
0.625	-0.4692	[-0.7714922 -0.1717882]	[-0.7710784 -0.1685761]	[-0.7701072 -0.1667967]
0.625	-0.4693	[-0.7941300 -0.1450292]	[-0.7962336 -0.1432821]	[-0.7953949 -0.1420597]
1.408	0.3418	[-0.2157858 0.9835972]	[-0.2258783 0.9119627]	[[-0.2250953 0.9062865]

Under Bayes prior, we can observe that 0 is not included in all of the 95% confidence intervals of log odds ratios, only included in the third ratio. This indicates that the rows and columns of the table are not independent, and we should conclude that there is relationship between delinquency and birth order.

Increasing N will make the histogram more closely resemble a normal distribution, but this will not substantially alter the estimates of means and standard deviations.

Conclusion

In this project, I calculated the 95% credible interval for each distinct log(odds ratio) from the simulated values, for each of the two priors. I then check if 0 is in each credible interval by trying n=10,000, 100,000, and 1,000,000 simulations. Also, I plot the density histogram of the simulated ln(odds ratio) along with the Frequentist normal approximation. And under both priors, I can observe that 0 is not included in all of the 95% confidence intervals of log odds ratios. This indicates that the rows and columns of the table are not independent and we should conclude that there is relationship between delinquency and birth order.

Appendix

Code chunk1:

```
library(ggplot2)
# Define vectors Y1 and Y2
Y1 <- c(127, 123, 93, 17)
Y2 <- c(345, 209, 158, 65)
# Define Haldane prior vector Here we use beta
priors <- rep(0, 4)
# priors <- rep(1, 4)
# Calculate new vectors Y11 and Y21
Y11 <- Y1 + priors
Y21 <- Y2 + priors
# Set the number of simulations to run
nsim <- 1000000
# nsim = 1000
# nsim = 10000
# Set the seed for reproducibility
set.seed(1234)
# Create a matrix to store log odd ratios
log_ratio <- matrix(0, nrow = 3, ncol = nsim)
# Run a loop to simulate data and calculate log odd ratios
for (i in 1:nsim) {
  # Simulate beta distributions for Y11 and Y21
  p1 <- rbeta(1, Y11[1], Y11[2] + Y11[3] + Y11[4])
  q1 <- (1 - p1) * rbeta(1, Y11[2], Y11[3] + Y11[4])
  r1 <- (1 - p1 - q1) * rbeta(1, Y11[3], Y11[4])
  s1 <- 1 - p1 - q1 - r1
  p2 <- rbeta(1, Y21[1], Y21[2] + Y21[3] + Y21[4])
  q2 <- (1 - p2) * rbeta(1, Y21[2], Y21[3] + Y21[4])
  r2 <- (1 - p2 - q2) * rbeta(1, Y21[3], Y21[4])
  s2 <- 1 - p2 - q2 - r2
  # Calculate log odd ratios and store in log_ratio matrix
  log_ratio[1, i] <- log((p1/q1)/(p2/q2))
  log_ratio[2, i] <- log((p1/r1)/(p2/r2))
  log_ratio[3, i] <- log((p1/s1)/(p2/s2))
}
# calculate means and standard deviations
means <- rowMeans(log_ratio)
sds <- apply(log_ratio, 1, sd)
# set index k and create sequence xx
k <- 1
```

```r
xx <- seq(min(log_ratio[k, ]), max(log_ratio[k, ]), length.out = 101)
# create a data frame with the log ratio values
df <- data.frame(log_ratio = log_ratio[k, ])
# create a ggplot histogram
ggplot(df, aes(x = log_ratio)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.05, color = "black", fill = "white") +
  labs(x = "log(p11p21/p22p12)", y = "Density",
       title = "Distributionunder Haldane Prior") +
  stat_function(fun = dnorm, args = list(mean = means[k], sd = sds[k]),
                aes(color = "APPRX distribution"), size = 2) +
  scale_color_manual(values = "red") +
  theme_classic()
# calculate quantiles
CI1 = quantile(log_ratio[k, ], prob = c(0.025, 0.975))

# set index k and create sequence xx
k <- 2
xx <- seq(min(log_ratio[k, ]), max(log_ratio[k, ]), length.out = 101)
# create a data frame with the log ratio values
df <- data.frame(log_ratio = log_ratio[k, ])
# create a ggplot histogram
ggplot(df, aes(x = log_ratio)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.05, color = "black", fill = "white") +
  labs(x = "log(p11p23/p13p21)", y = "Density",
       title = "Distribution under Haldane Prior") +
  stat_function(fun = dnorm, args = list(mean = means[k], sd = sds[k]),
                aes(color = "APPRX distribution"), size = 2) +
  scale_color_manual(values = "red") +
  theme_classic()
# calculate quantiles
CI2 = quantile(log_ratio[k, ], prob = c(0.025, 0.975))
# set index k and create sequence xx
k <- 3
xx <- seq(min(log_ratio[k, ]), max(log_ratio[k, ]), length.out = 101)
# create a data frame with the log ratio values
df <- data.frame(log_ratio = log_ratio[k, ])
# create a ggplot histogram
ggplot(df, aes(x = log_ratio)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.05, color = "black", fill = "white") +
  labs(x = "log(p11p24/p14p21)", y = "Density",
       title = "Distribution under Haldane Prior") +
  stat_function(fun = dnorm, args = list(mean = means[k], sd = sds[k]),
                aes(color = "APPRX distribution"), size = 2) +
  scale_color_manual(values = "red") +
  theme_classic()
# calculate quantiles
CI3 = quantile(log_ratio[k, ], prob = c(0.025, 0.975))
```