BDA Project: Analizing Default

Andres Potapczynski (ap3635), Jongwoo Choi (jc4816), Yi Chen (yc3356)

12/10/2018

```
library(rstan)
library(tidyverse)
library(arm)
library(ggplot2)
library(gridExtra)
library(bayesplot)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
knitr::opts_chunk$set(echo = TRUE)
```

```
file <-  '../DBs/core.txt'
file_model_logistic <- '../Analysis/log_reg_v02.stan'
file_model_base <- '../Analysis/binomial_spatial_base.stan'
file_model_binomial <- '../Analysis/binomial_spatial_01.stan'
file_model_binomial_ext <- '../Analysis/binomial_spatial_02.stan'
```

Setting the preamble

```
data <- read_delim(file = file, delim = '|')

# Sample the data
pct = 1
# pct = 0.1
# pct = 0.01
set.seed(seed = 42)
sample_size = round(pct * nrow(data))
sample <- sample(x = nrow(data), size = sample_size, replace = F)
data = data[sample, ]

## Selecting the relevant columns for the analysis
data_sub <- data %>% dplyr::select(
  state,
  city,
  county,
  zip,
  asset_market_value,
  mar_2_app,
  appraisal_value,
  app_2_inc,
  client_income,
  mar_2_inc,
  age,
  sex_F,
  condition_U,
```

```
  y)
summary(data_sub)
```

```
##     state               city              county               zip
##  Length:30499       Length:30499       Length:30499       Min.   : 1000
##  Class :character   Class :character   Class :character   1st Qu.:32680
##  Mode  :character   Mode  :character   Mode  :character   Median :55295
##                                                           Mean   :54236
##                                                           3rd Qu.:76148
##                                                           Max.   :99900
##  asset_market_value   mar_2_app       appraisal_value     app_2_inc
##  Min.   : 120000    Min.   : 0.9184   Min.   :  79521   Min.   :0.004633
##  1st Qu.: 350000    1st Qu.: 1.1259   1st Qu.: 302908   1st Qu.:0.088792
##  Median : 398000    Median : 1.2209   Median : 320052   Median :0.100663
##  Mean   : 491311    Mean   : 1.3378   Mean   : 371064   Mean   :0.098457
##  3rd Qu.: 463000    3rd Qu.: 1.3620   3rd Qu.: 348289   3rd Qu.:0.111070
##  Max.   :4519000    Max.   :21.8468   Max.   :1654602   Max.   :0.367728
##  client_income      mar_2_inc            age            sex_F
##  Min.   : 143.9   Min.   :0.02738   Min.   :18.00   Min.   :0.0000
##  1st Qu.: 284.3   1st Qu.:0.10934   1st Qu.:27.00   1st Qu.:0.0000
##  Median : 310.7   Median :0.12381   Median :32.00   Median :0.0000
##  Mean   : 409.8   Mean   :0.12856   Mean   :34.29   Mean   :0.3082
##  3rd Qu.: 341.7   3rd Qu.:0.13978   3rd Qu.:40.00   3rd Qu.:1.0000
##  Max.   :1887.2   Max.   :1.09440   Max.   :65.00   Max.   :1.0000
##   condition_U           y
##  Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000
##  Mean   :0.3954   Mean   :0.06403
##  3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000
```

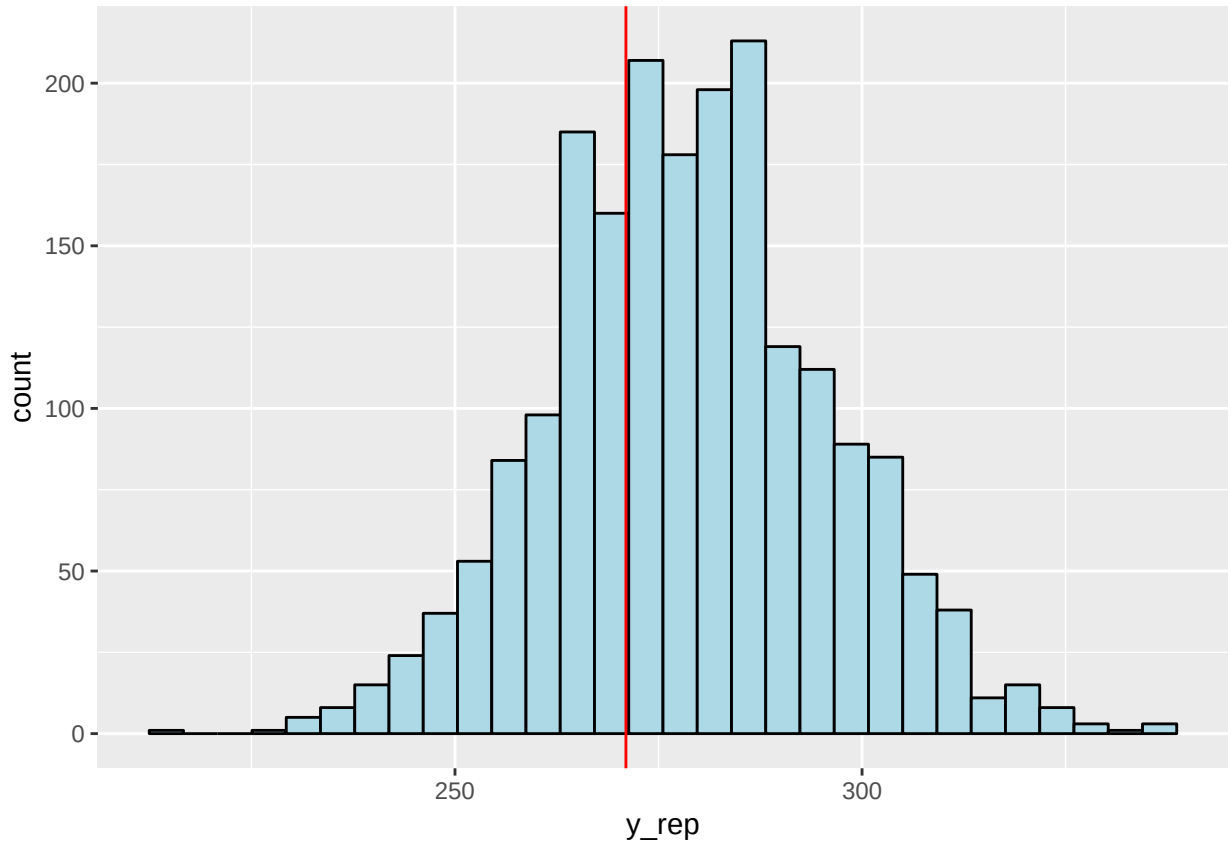
```
geo <- data_sub %>%
  group_by(state) %>%
  summarize(market_mean = mean(asset_market_value),
            appraisal_mean = mean(appraisal_value),
            income_mean = mean(client_income),
            mar_2_inc_mean = mean(mar_2_inc),
            app_2_inc_mean = mean(app_2_inc),
            mar_2_app_mean = mean(mar_2_app),
            age_mean = mean(age),
            y_sum = sum(y),
            state_n = n()) %>%
  ungroup()
```

```
## Rescaling
inputs <- geo %>%
  mutate(
    income_st = (income_mean - mean(income_mean)) / sd(income_mean),
    appraisal_st = (appraisal_mean - mean(appraisal_mean)) / sd(appraisal_mean),
    market_st = (market_mean - mean(market_mean)) / sd(market_mean),
    mar_2_inc_st = (mar_2_inc_mean - mean(mar_2_inc_mean)) / sd(mar_2_inc_mean),
    app_2_inc_st = (app_2_inc_mean - mean(app_2_inc_mean)) / sd(app_2_inc_mean),
    mar_2_app_st = (mar_2_app_mean - mean(mar_2_app_mean)) / sd(mar_2_app_mean),
```

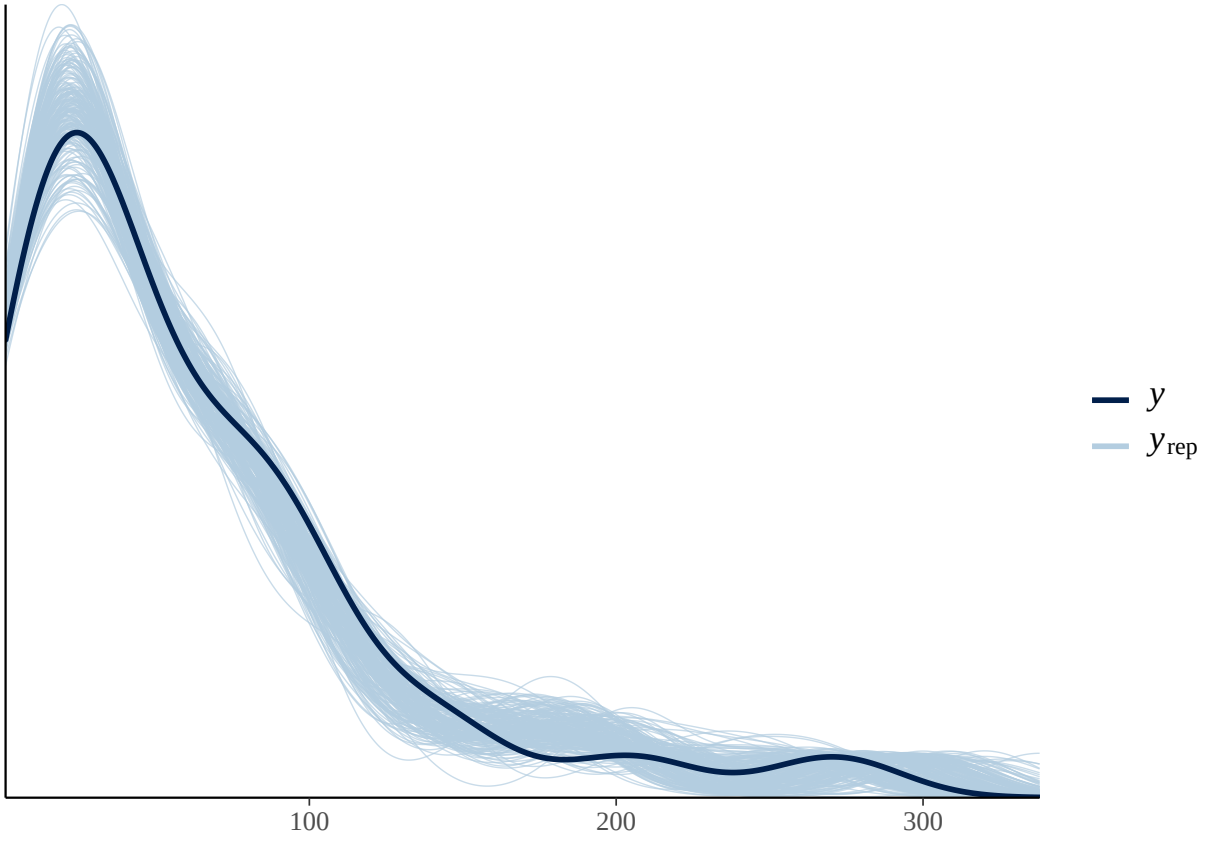
```
    age_st = (age_mean - mean(age_mean)) / sd(age_mean)
         ) %>%
  dplyr::select(
    income_st,
    mar_2_inc_st,
    appraisal_st,
    app_2_inc_st,
    mar_2_app_st,
    market_st,
    age_st,
    state_n,
    y_sum
  )

## Inputs for STAN
y = inputs$y_sum
Ns = inputs$state_n
X = inputs %>% dplyr::select(-y_sum, -state_n)
N = nrow(X)
D = ncol(X)

data_stan_noX = list(N=N, Ns=Ns, y=y)
data_stan = list(N=N, D=D, X=X, Ns=Ns, y=y)
```

```
comp_01 <- stan_model(file_model_base)
sm = sampling(comp_01, data=data_stan_noX, iter=1000, chains=4)
sims <- rstan::extract(sm)
```

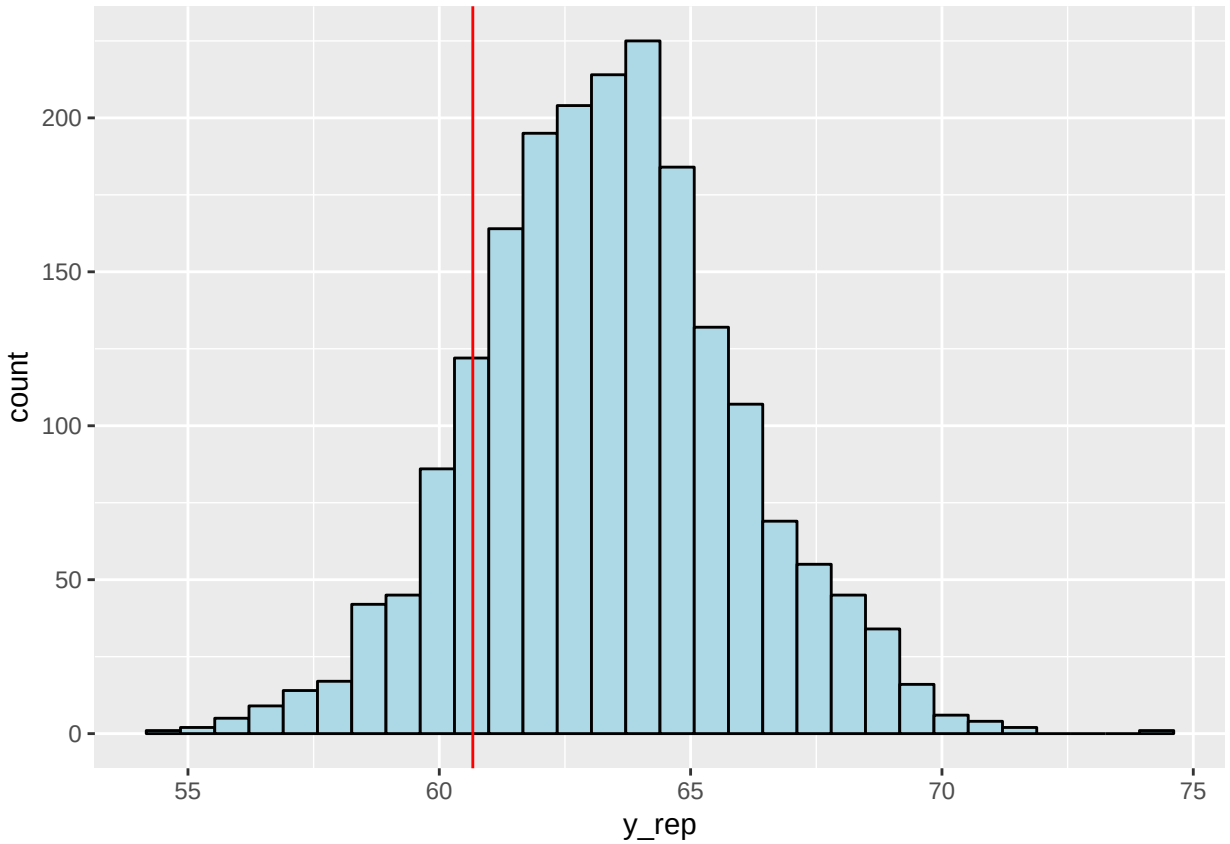
```
y_max <- apply(X = sims$y_rep, MARGIN = 1, FUN = max)
df <- data.frame(y_rep = y_max)
ggplot(df, aes(x=y_rep)) +
  geom_histogram(fill='lightblue',
                 color='black') +
  geom_vline(xintercept = max(y), color='red')
```



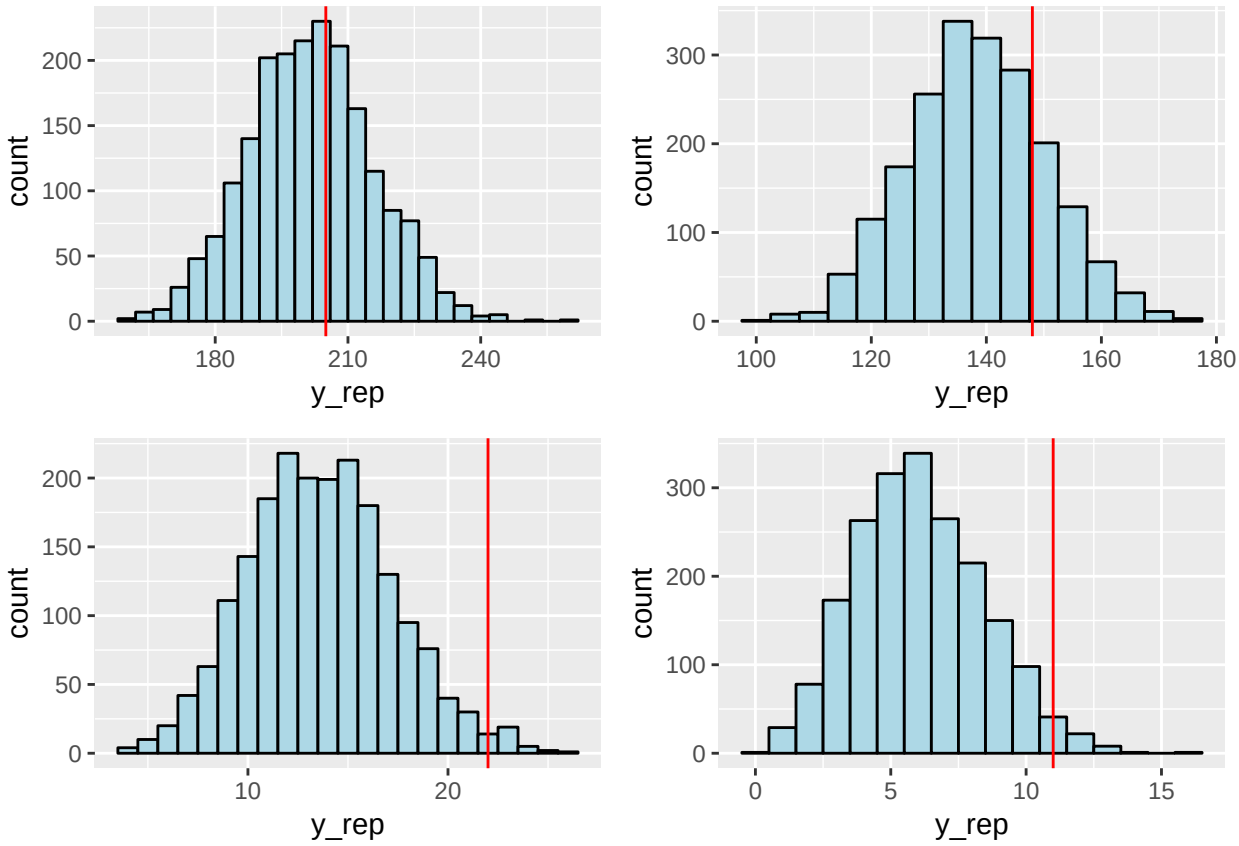

```
y_rep <- as.matrix(sm, pars = "y_rep")
ppc_dens_overlay(y = y, y_rep[1:200,])
```




```
df <- data.frame(y_rep = apply(X = sims$y_rep, MARGIN = 1, FUN = sd))
ggplot(df, aes(x=y_rep)) +
  geom_histogram(fill='lightblue',
                 color='black') +
  geom_vline(xintercept = sd(y), color='red')
```

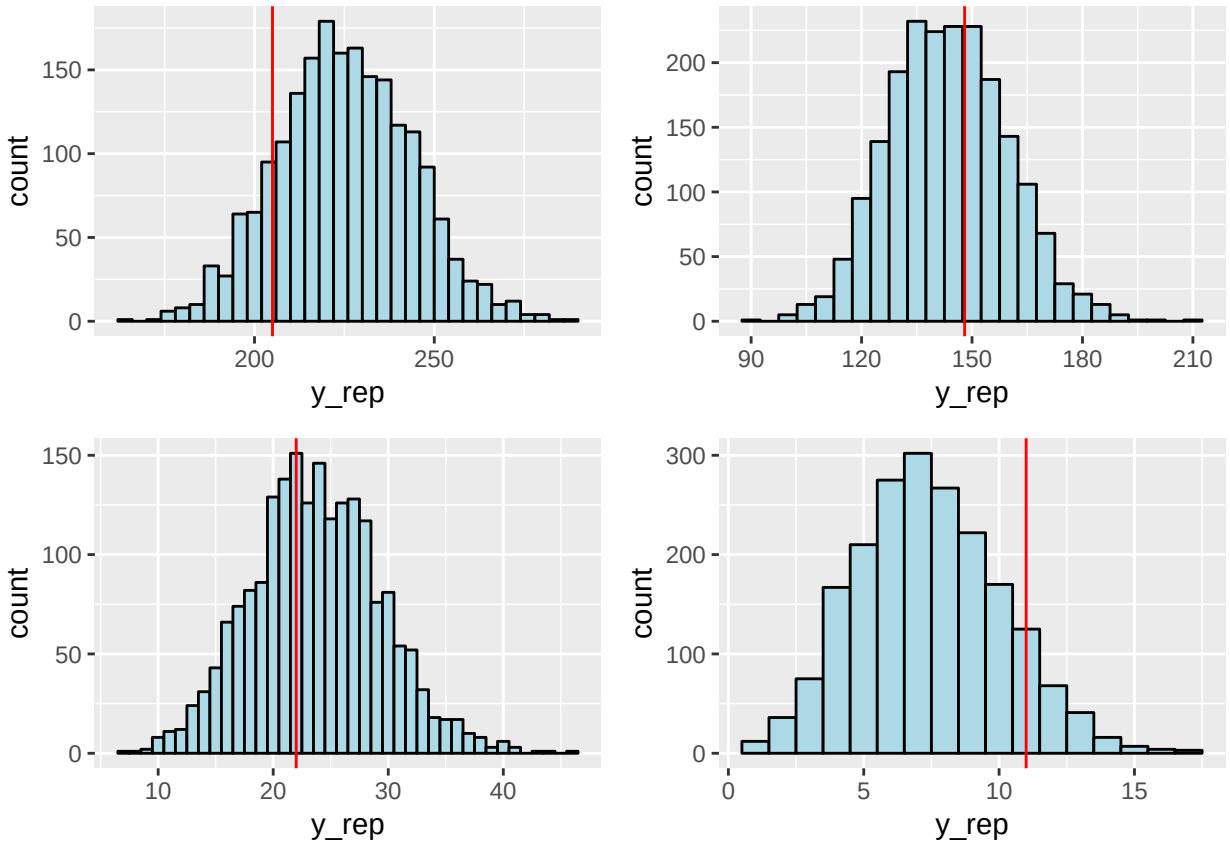



```
## Inference for Stan model: binomial_spatial_base.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##       mean se_mean sd 2.5%  50% 97.5% n_eff Rhat
## theta 0.06       0  0 0.06 0.06  0.07   677    1
##
## Samples were drawn using NUTS(diag_e) at Sun Nov 25 20:31:29 2018.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```



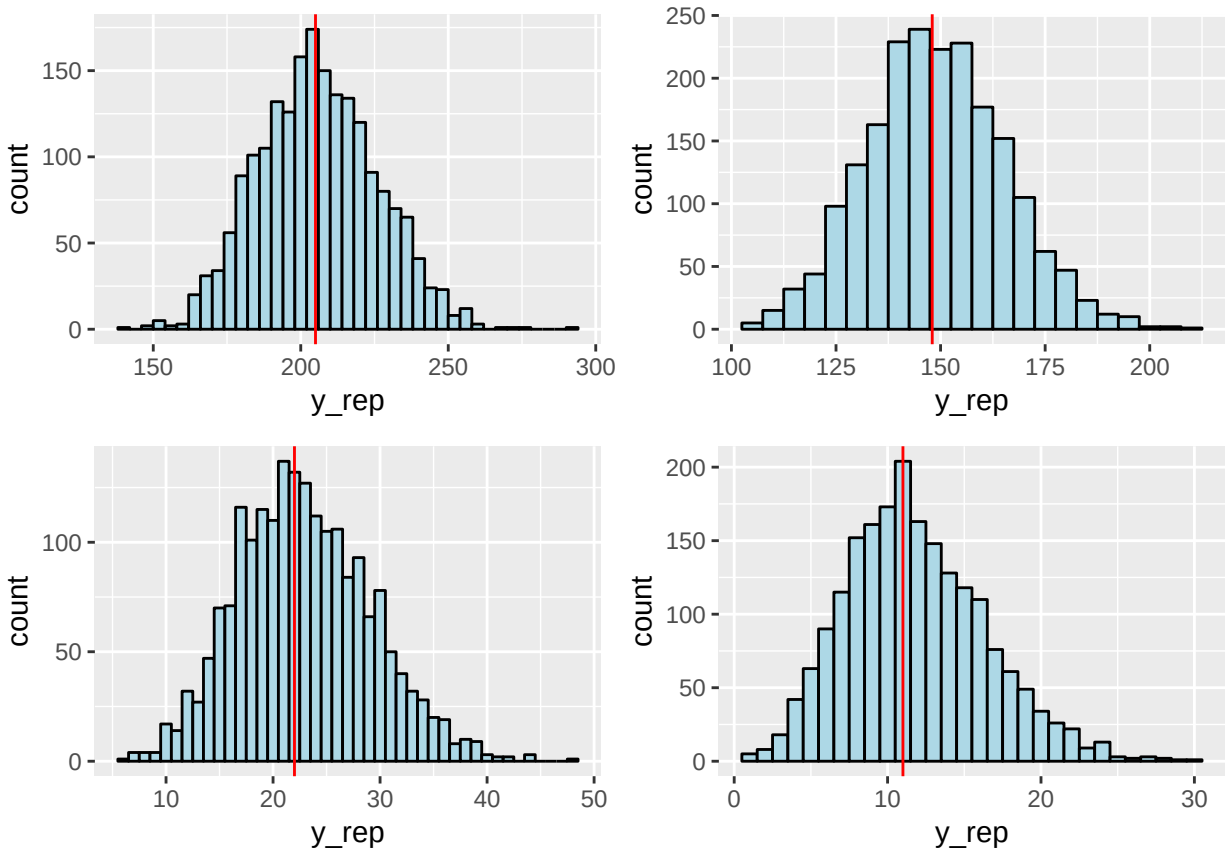

```
comp_02 <- stan_model(file_model_binomial)
sm = sampling(comp_02, data=data_stan, iter=1000, chains=4)
sims <- rstan::extract(sm)
```

The plots have improved




```
comp_03 <- stan_model(file_model_binomial_ext)
sm = sampling(comp_03, data=data_stan_noX, iter=1000, chains=4)
sims <- rstan::extract(sm)
```

Now the plots have improved even further



asdf