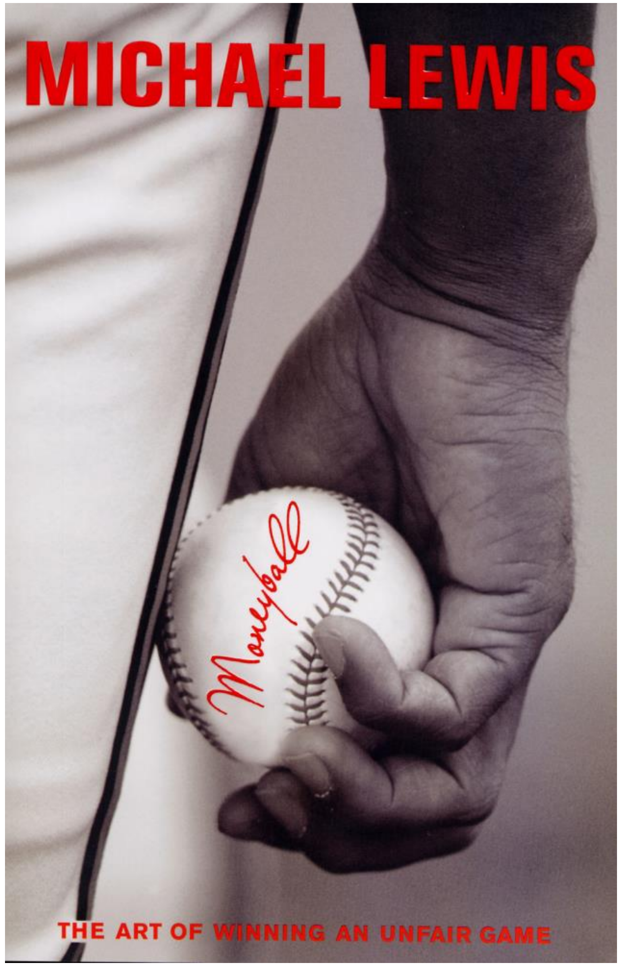
Predicting MLB player batting average (BA): A comparison between Bayesian and frequentist approaches

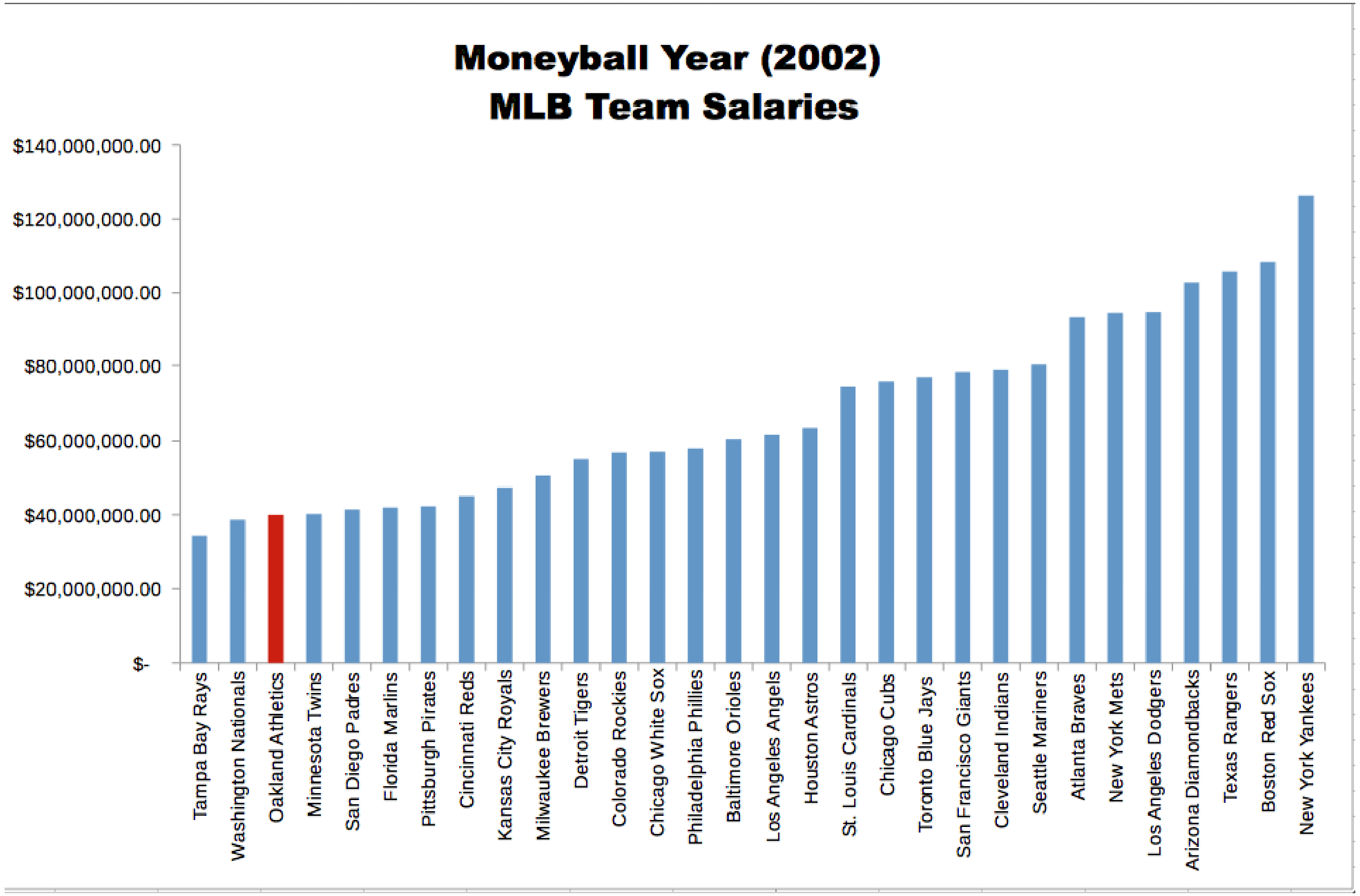
EN.553.732 Bayesian Statistics

Kenneth Feder, Joseph High, Joseph Yu

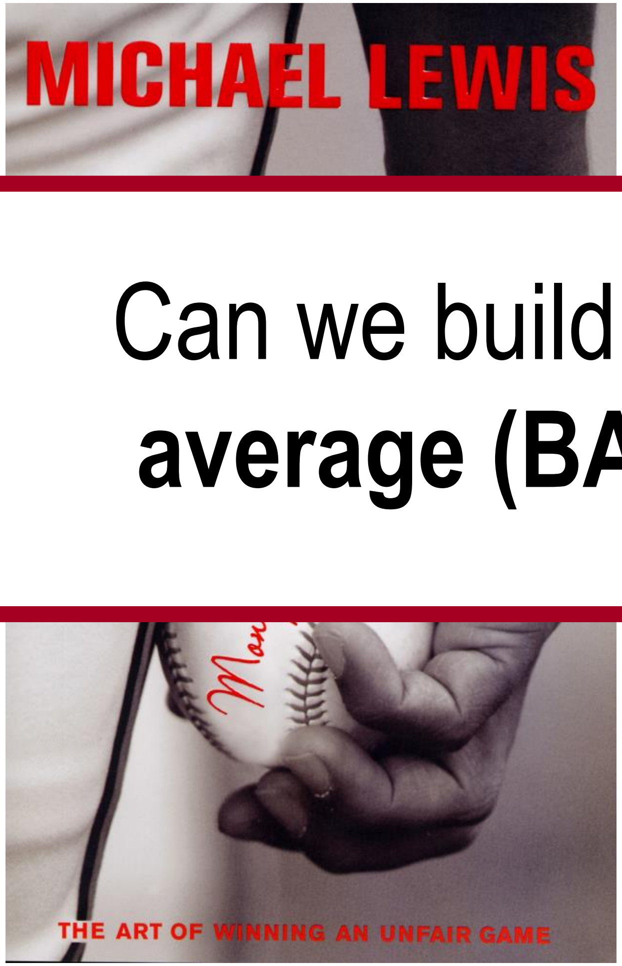
Dec 2nd, 2017

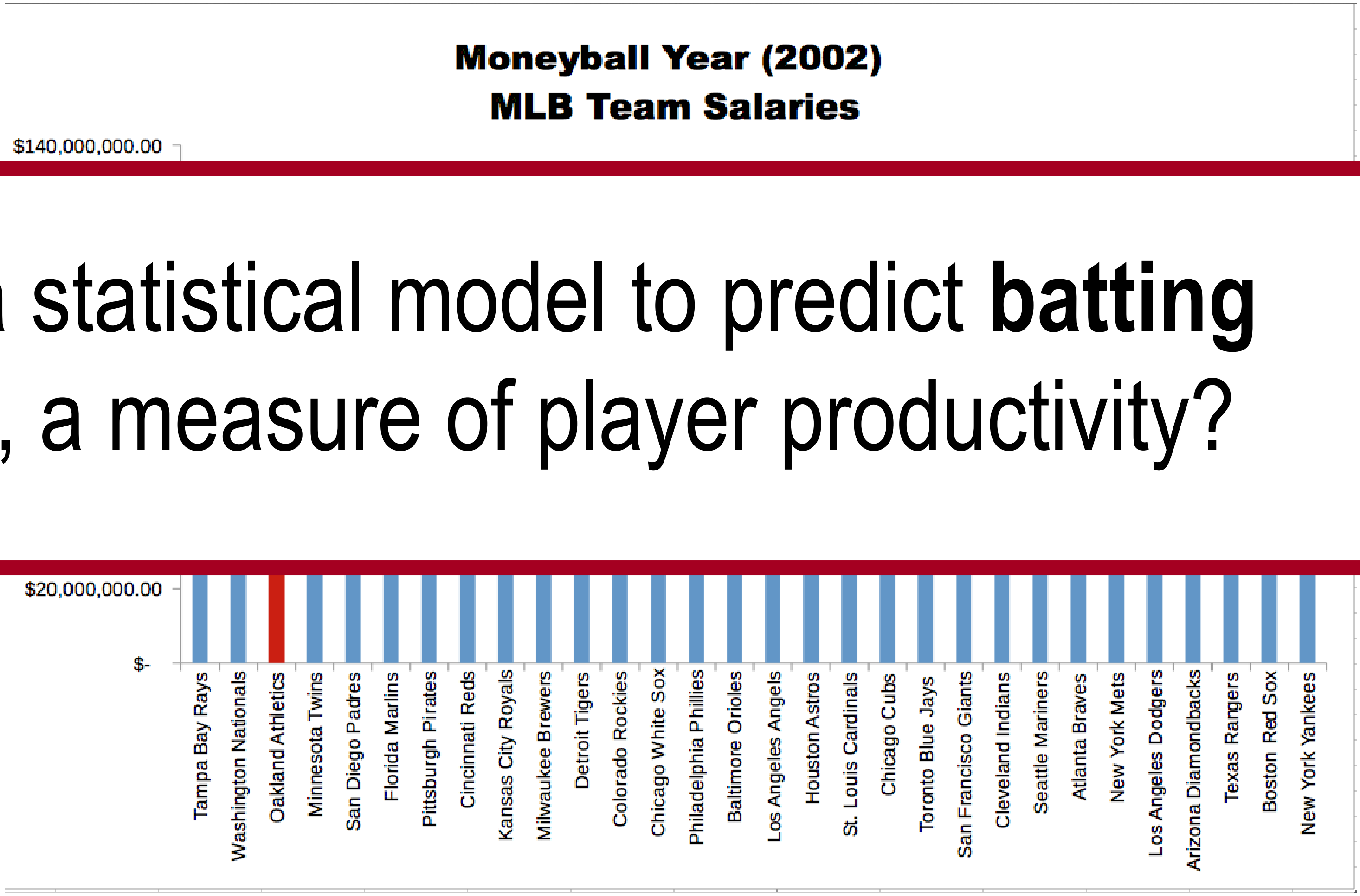
Oakland A's successfully leveraged sabermetrics to field a competitive MLB team





Oakland A's successfully leveraged sabermetrics to field a competitive MLB team





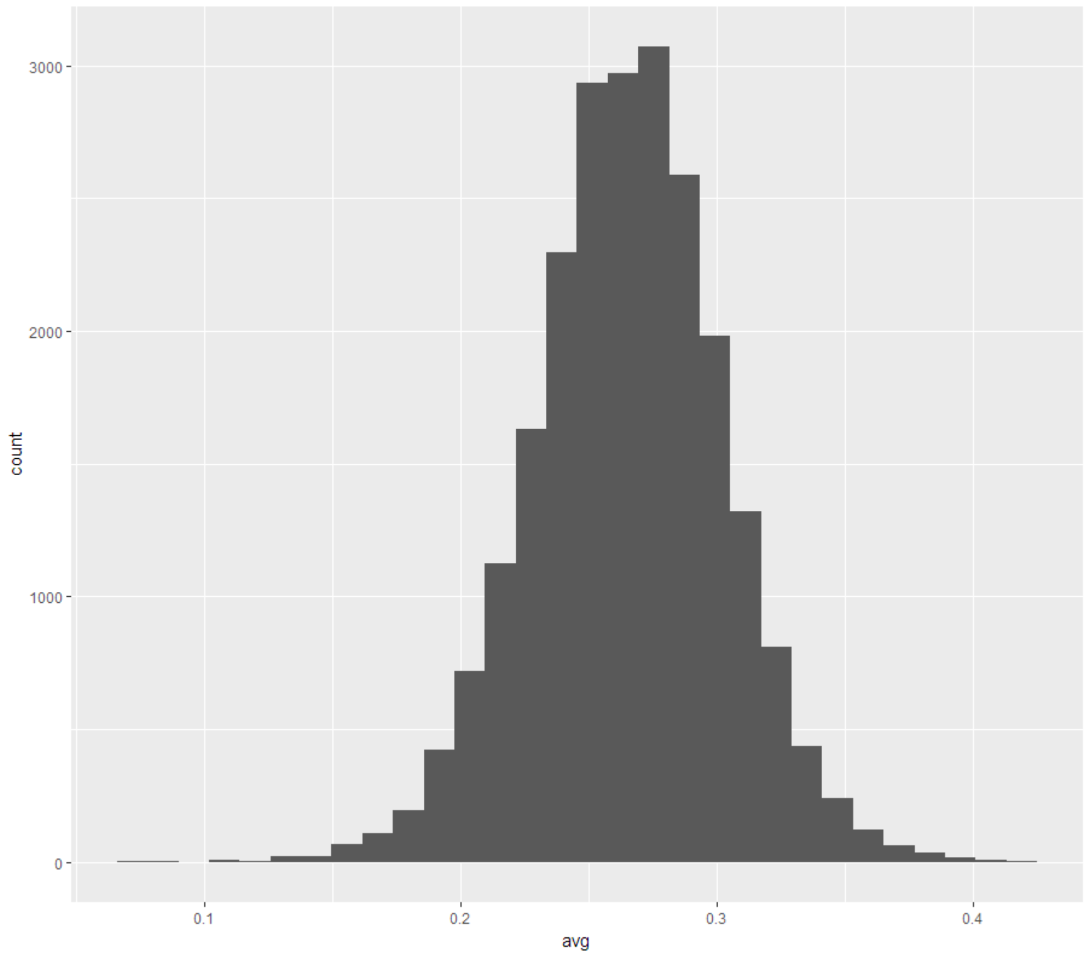
Can we build a statistical model to predict **batting average (BA)**, a measure of player productivity?

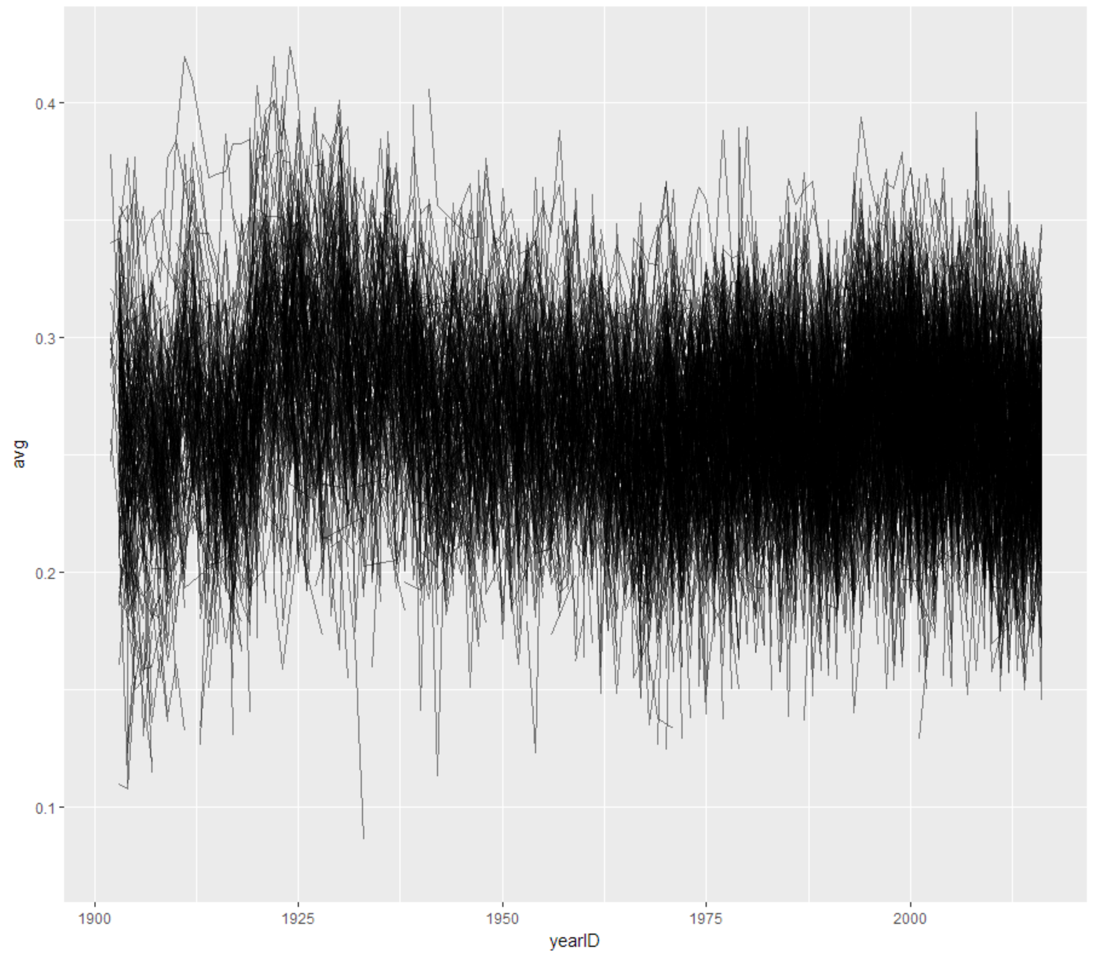
MLB player statistics and inclusion criteria

- Data acquired from Lahman Baseball archive
  - http://www.seanlahman.com/baseball-archive/statistics/

- MLB player data inclusion criteria:
  1. Played after 1901
  2. Had at least 100 at bats per season
  3. Played in either the National League or American League

Batting average data visualized





Multi-level Bayesian model was implemented

Outcome

B_{ij}: predicted batting average (BA) for a given player (i), year (j)

Predictors

B_{ij-1}: previous year BA
B_{ij-2}: prior year BA

age
height
weight
era
- lively ball (1920 -)
- expansion (1961 -)
- free agency (1977 -)
- steroids (1994 -)

year: spent in league

Model

$$B_{ij} \sim \text{Normal}(\mu_{ij}, 1/\tau_{avg})$$

i = player index
j = year index

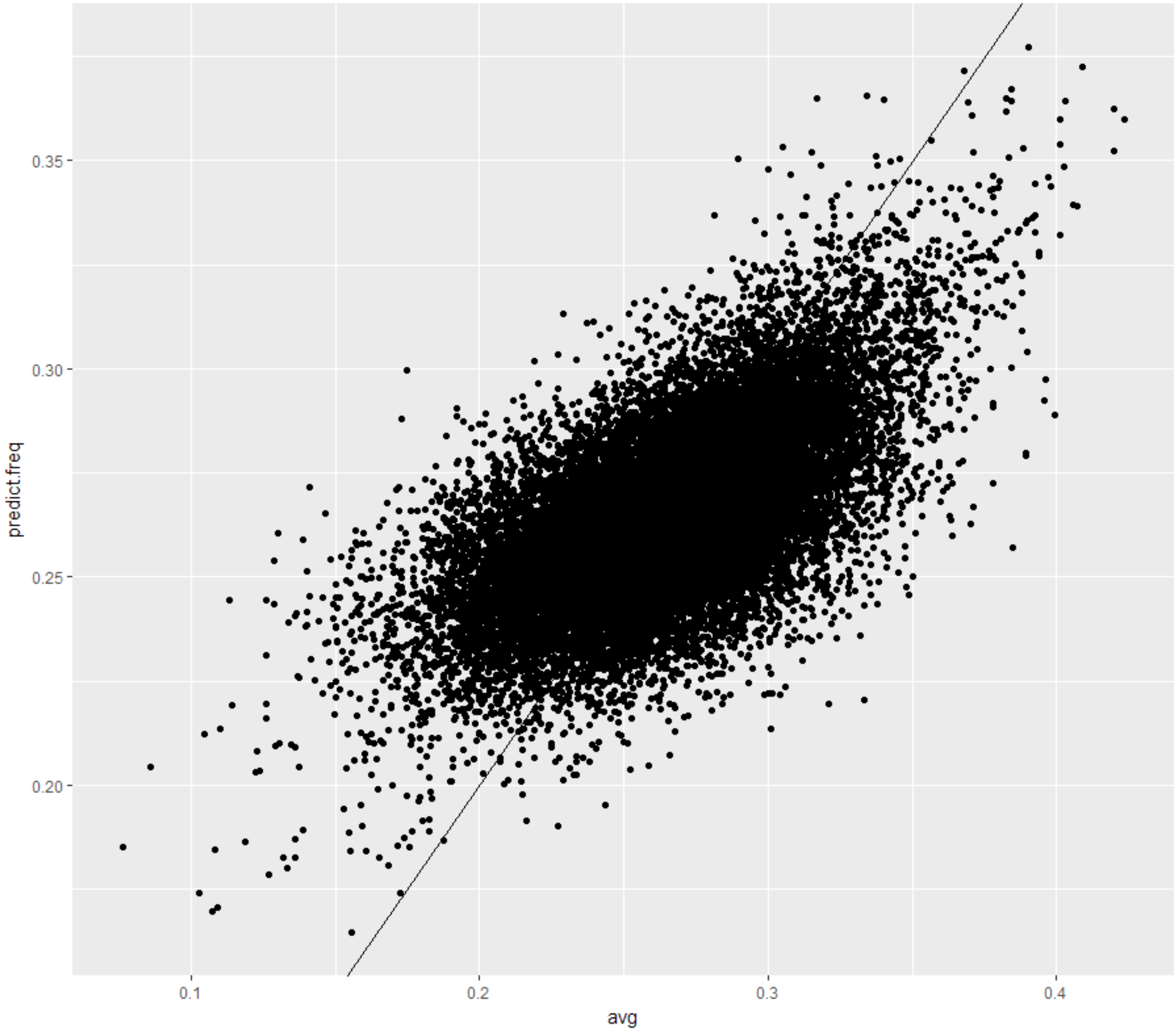
$$\mu_{ij} = \beta_{0i} + \beta_1 B_{ij-1} + \beta_2 B_{ij-2} + \beta_3 age_{ij} + \beta_4 age_{ij}^2$$

$$\beta_1, \beta_2, \beta_3, \beta_4 \sim \text{Normal}(0, 1000)$$
$$\tau_{avg} \sim \text{Gamma}(0.001, 0.001)$$

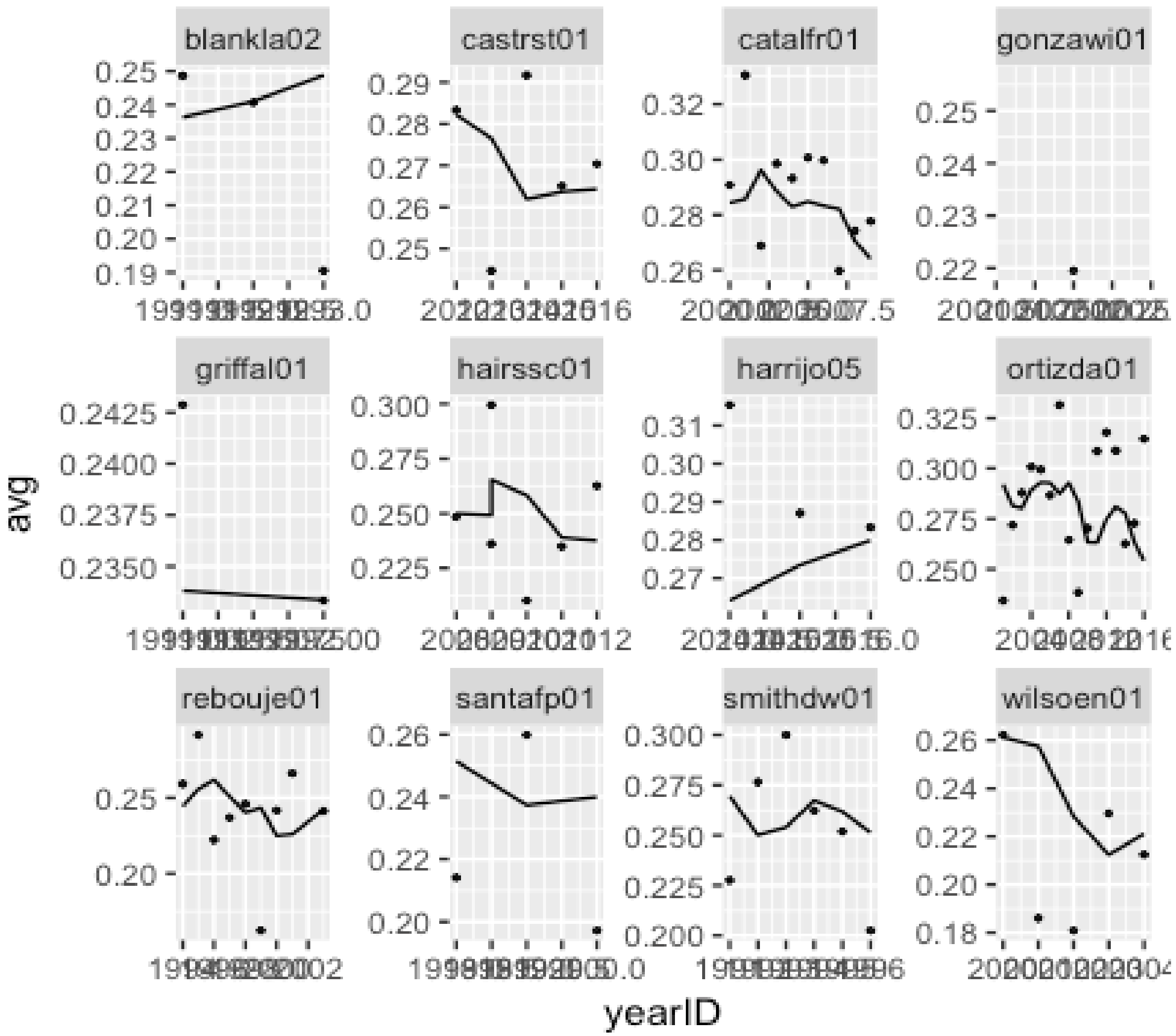
$$\beta_{0i} \sim \text{Normal}(\theta_i, 1/\tau_{\beta 0})$$
$$\theta_i = \gamma_0 + \gamma_1 height_i + \gamma_2 height_i^2 + \gamma_3 weight_i + \gamma_4 weight_i^2 + \gamma_5 year_i + \gamma_6 era_{liveball} + \gamma_7 era_{expansion} + \gamma_8 era_{freeagency} + \gamma_9 era_{steroids} + \gamma_{10} era_{modern}$$

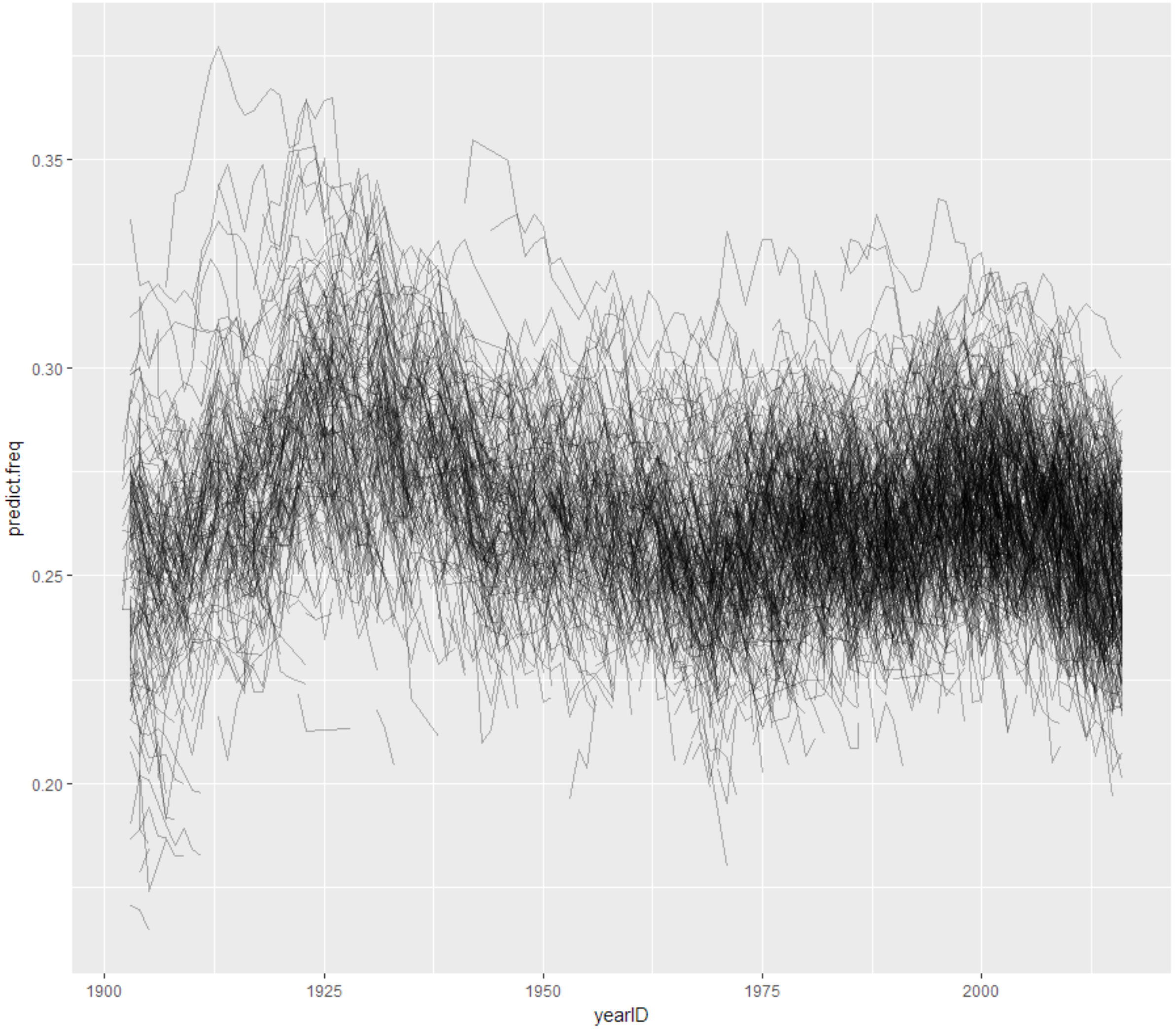
$$\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_8, \gamma_9, \gamma_{10} \sim \text{Normal}(0, 1000)$$
$$\tau_{\beta 0} \sim \text{Gamma}(0.001, 0.001)$$

Predictive accuracy of fitted frequentist model

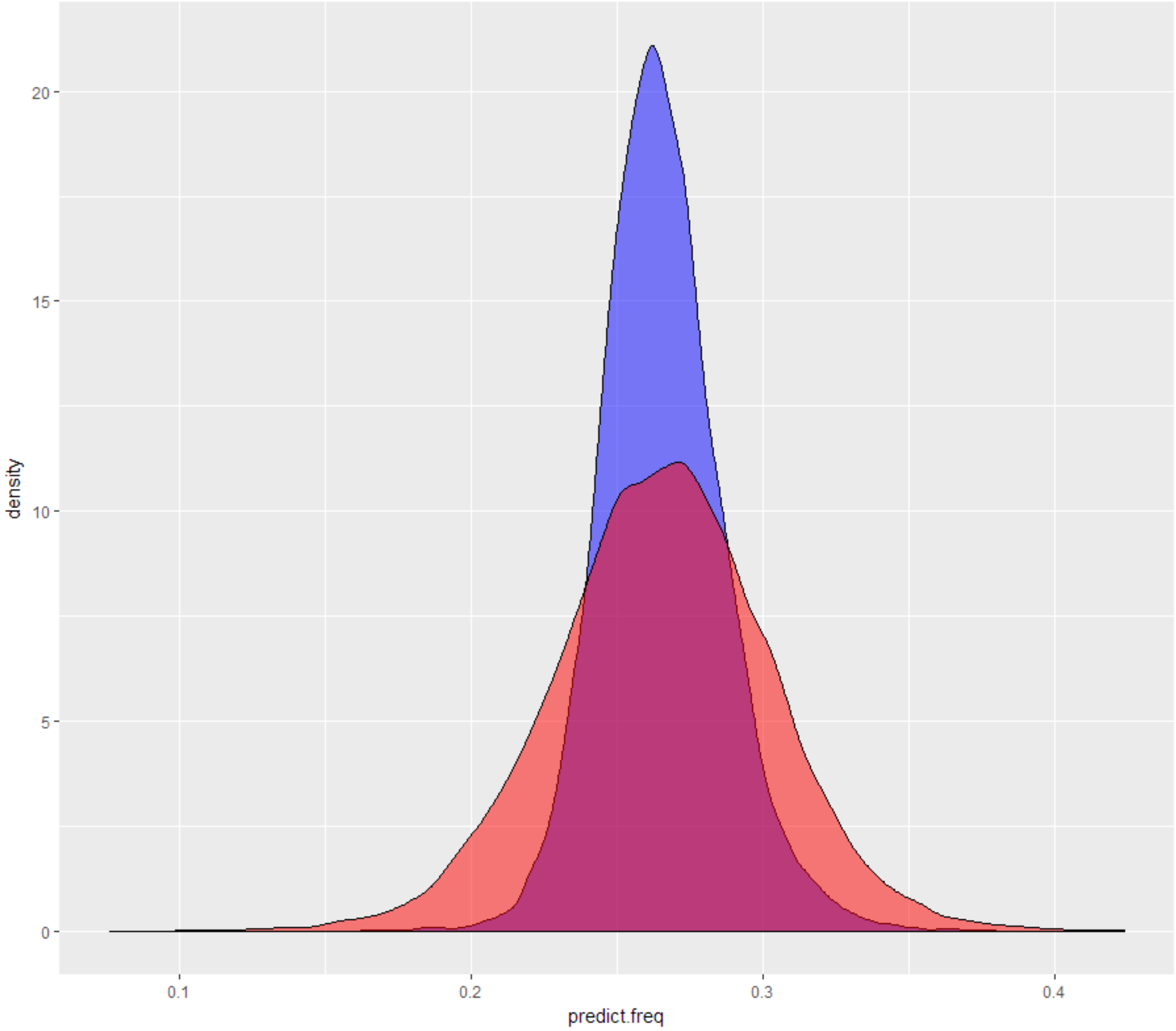


Fitted frequentist model predicted trajectories





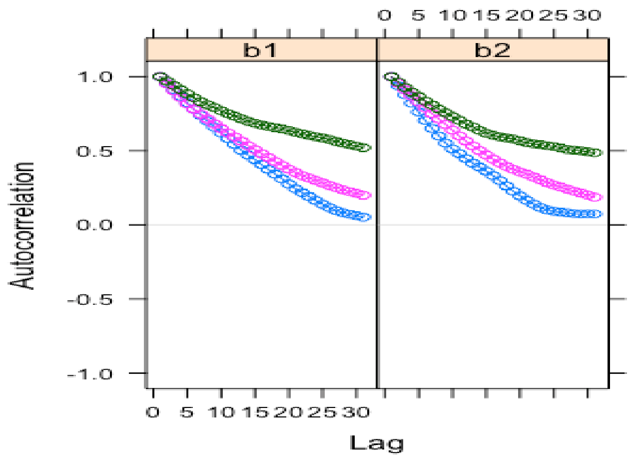
Frequentist – Accurate Mean, no Model for Variability

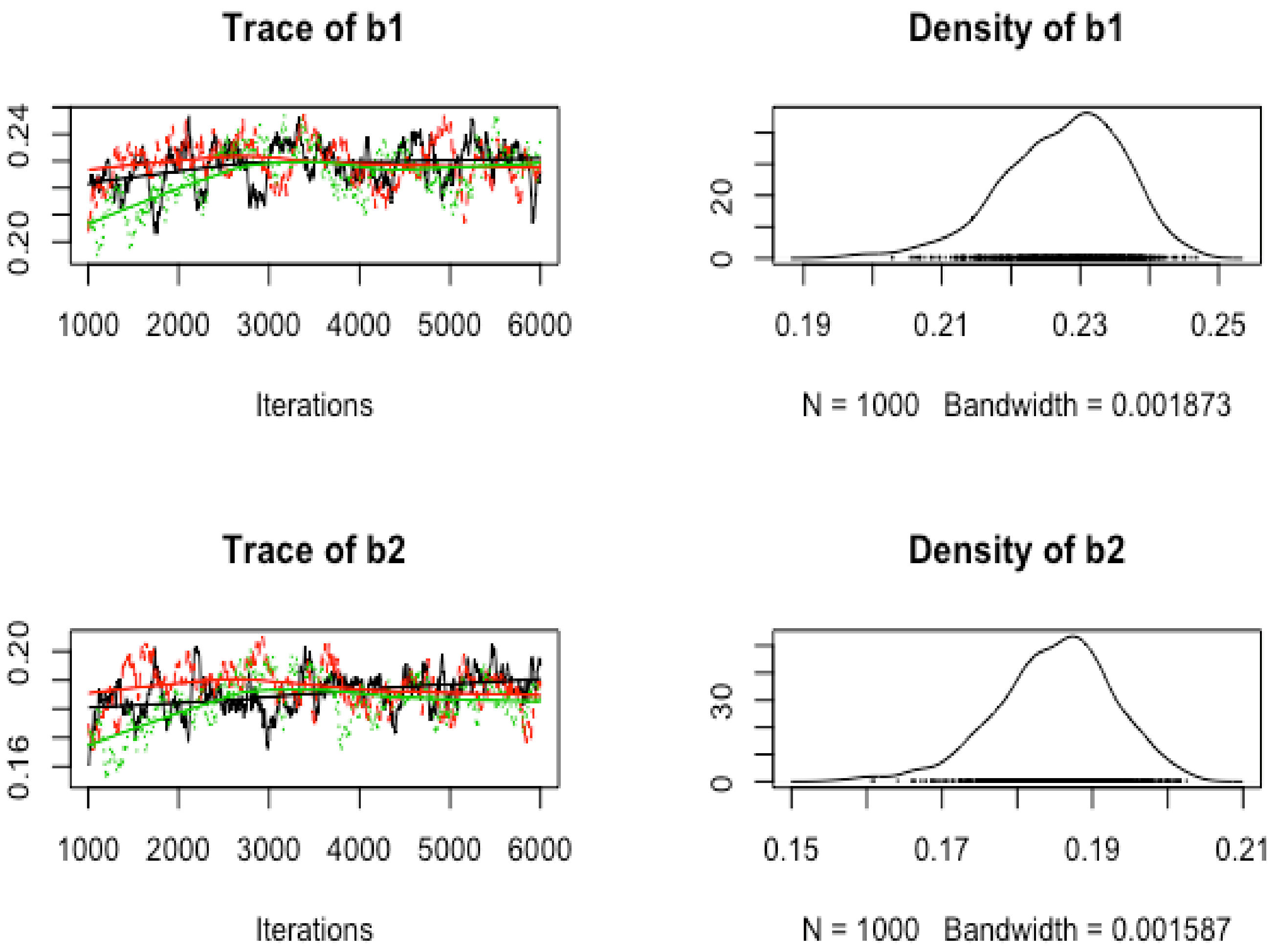


Bayesian model parameters demonstrate convergence

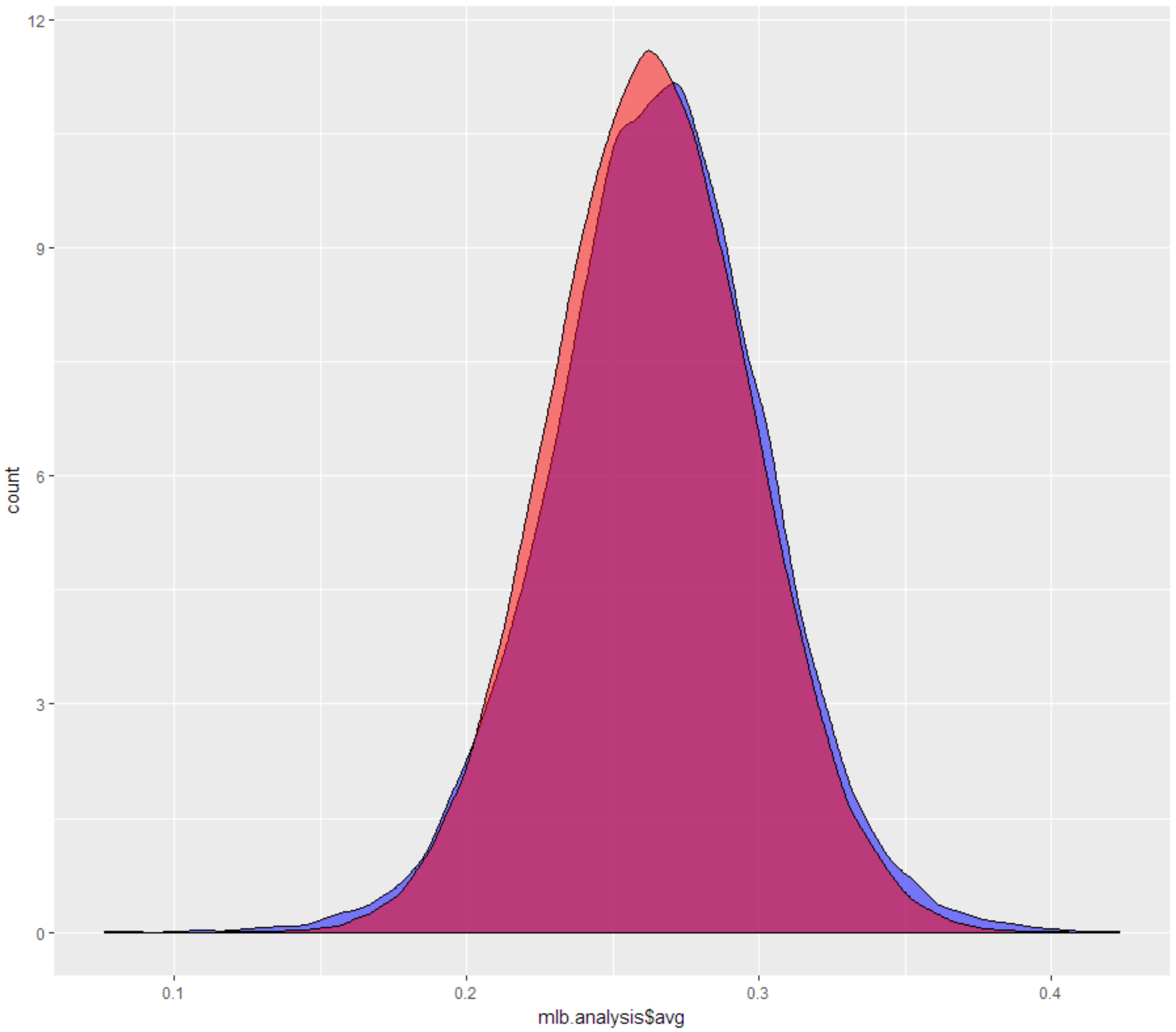
(MCMC sampling implemented R package JAGS)

```
> gelman.diag(beta.samples)
Potential scale reduction factors:

   Point est. Upper C.I.
b1       1.03       1.09
b2       1.05       1.17

Multivariate psrf

1.08
>
```





Bayesian posterior predictive reflects full variability in data



Developed a multi-level Bayesian model in predicting MLB player BA for a given season

1. Predictors included: BA in previous and prior years, age, height, weight, era, and years in league
2. Frequentist and Bayesian models are similar
3. Frequentist model provides less uncertainty in prediction
4. Bayesian posterior predictive distribution fits true BA well

Questions?