

## The Setting

You are a statistician employed by On The Ball Consulting. Veteran major-league baseball scout Rocky Chew seeks your advice regarding estimating the probability that amateur baseball player John Spurrier will get a base hit against a major-league pitcher. Rocky has arranged for Spurrier to have ten at bats against a major-league pitcher.

## The Background

The traditional batting average,  $\hat{\theta}_f = X/n$  is a frequentist estimator in that it makes use of the observed data, but ignores any prior information that might exist. (Some of you baseball enthusiasts will be a bit uncomfortable that we're going to assume that our denominator is # of times up to bat.) If we assume that the at bats are independent Bernoulli trials with a constant probability of getting a base hit, then

$$X \sim \text{Bin}(n = \text{number at bat}, \theta = \text{P}(\text{getting a base hit}))$$

$\hat{\theta}_f$ , is the maximum likelihood estimator, the method of moments estimator, and the minimum variance unbiased estimator of the unknown probability (of getting a base hit.) That makes it a good estimator, but it ignores information we might have about baseball. You have the following prior information:

- John Spurrier appears to be a good but not great player. He is one of the better batters on a somewhat above-average American Legion (high school) baseball team.
- The few major-league scouts who have watched him play do not believe that Spurrier's batting ability is at the professional level.
- A barely adequate major-league hitter has a batting average of about 0.200.
- A very good major-league batter has a batting average of about 0.300.
- Ty Cobb has the all-time best major-league batting average of 0.366.

We're going to use a Beta prior to incorporate our previous knowledge. What should that prior look like?

If we measure the goodness of an estimate  $\hat{\theta}$  using the squared error loss, then the Bayesian estimator is the expected value of the posterior distribution (i.e., the mean of the posterior distribution.) The Bayesian estimator is:

$$\hat{\theta}_b = \frac{X + \alpha}{n + \alpha + \beta}$$

## The Experiment

1. John Spurrer will have  $n=10$  at bats. The random variable,  $X$ , will be the number of base hits that he gets.
2. Determining the prior probability: As a class we will find  $\alpha$  and  $\beta$  that are consistent with our prior information.
3. Collecting data: let's calculate our estimates for all possible realizations of the random variable.

| $x$ | $\hat{\theta}_f$ | $\hat{\theta}_b$ |
|-----|------------------|------------------|
| 0   | 0.00             |                  |
| 1   | 0.10             |                  |
| 2   | 0.20             |                  |
| 3   | 0.30             |                  |
| 4   | 0.40             |                  |
| 5   | 0.50             |                  |
| 6   | 0.60             |                  |
| 7   | 0.70             |                  |
| 8   | 0.80             |                  |
| 9   | 0.90             |                  |
| 10  | 1.00             |                  |

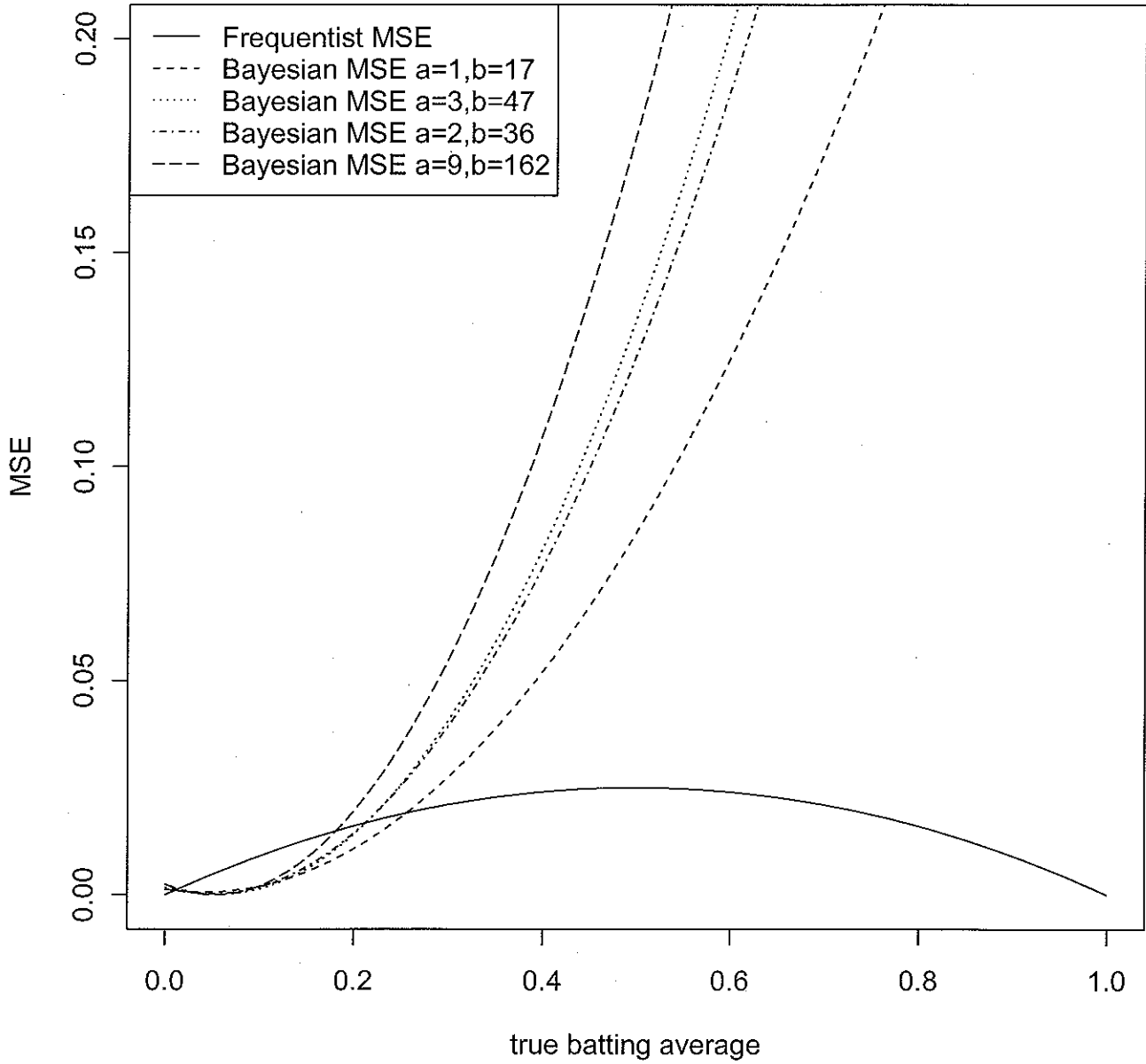
4. Comparison of the estimators:

- $\hat{\theta}_f = \frac{X}{n}$        $\hat{\theta}_b = \frac{X+\alpha}{n+\alpha+\beta}$
- We use Mean Squared Error (MSE) in the frequentist sense (that is,  $X$  is the random variable,  $\theta$  is no longer random) to compare estimators (apples to apples):

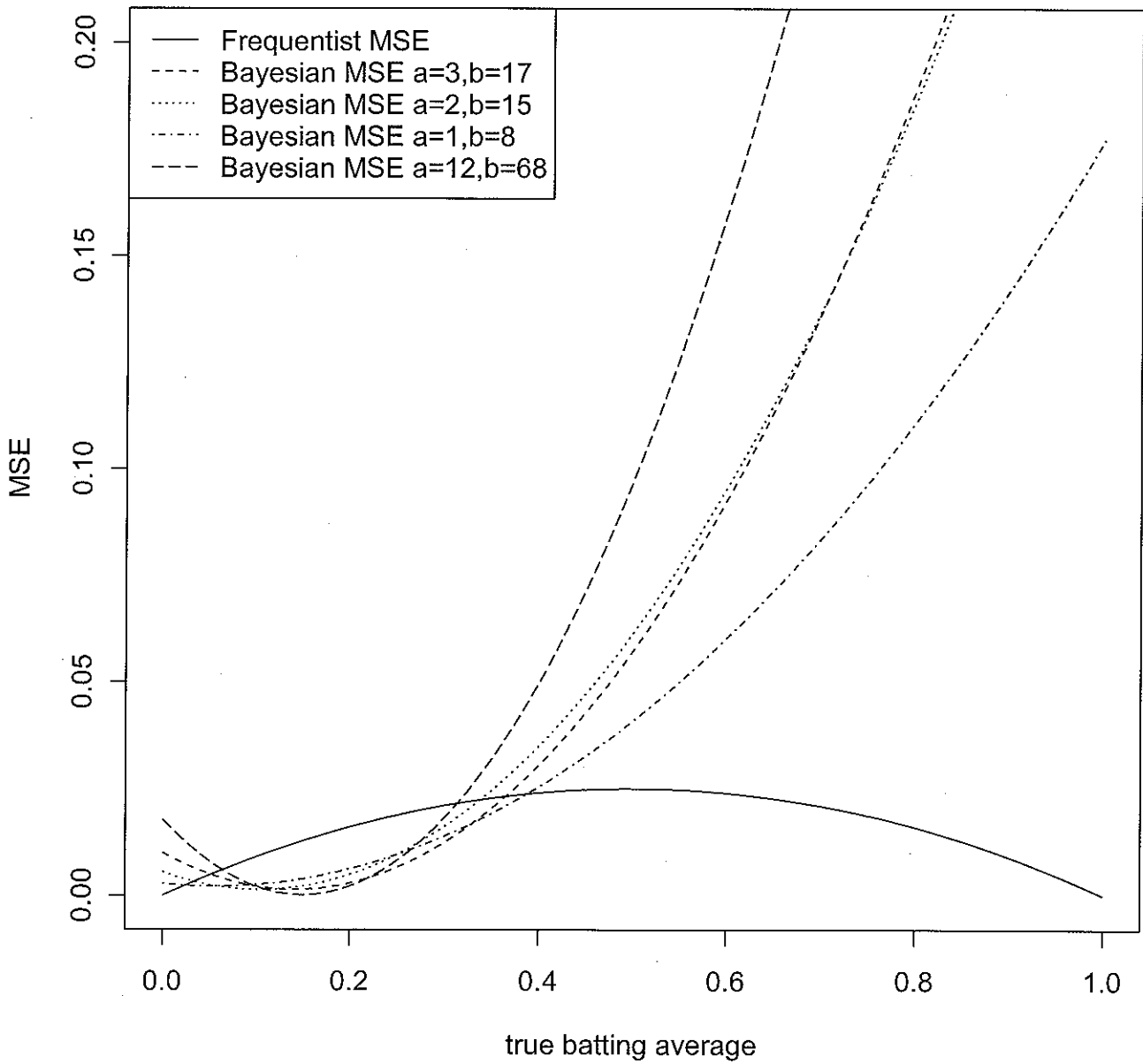
$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + bias^2(\hat{\theta}) = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

- Under the assumption that  $X$  has a binomial distribution with parameters 10 and  $\theta$ , calculate the mean and variance of  $X$ .

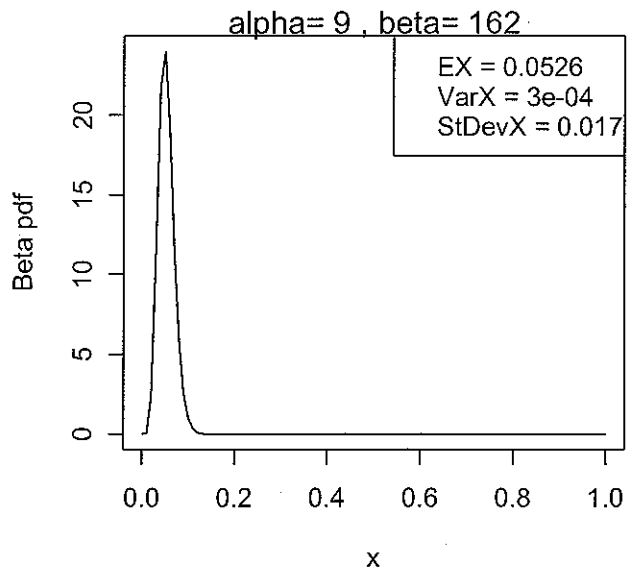
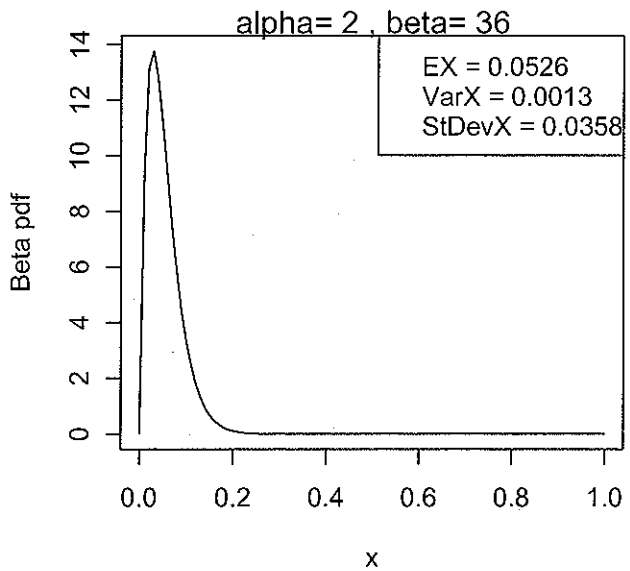
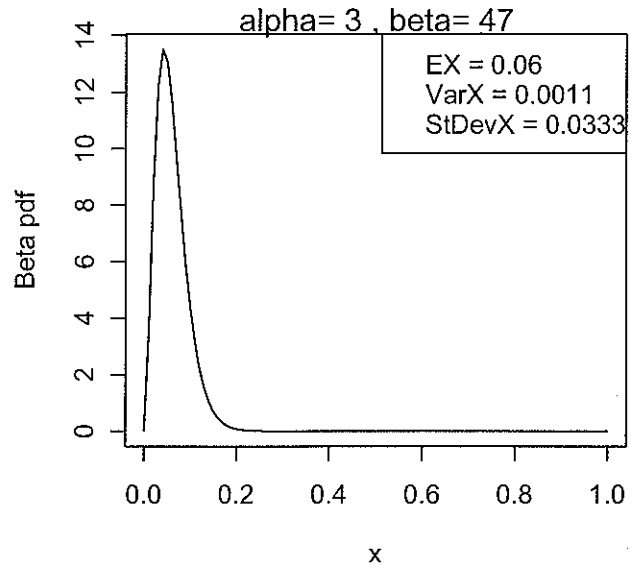
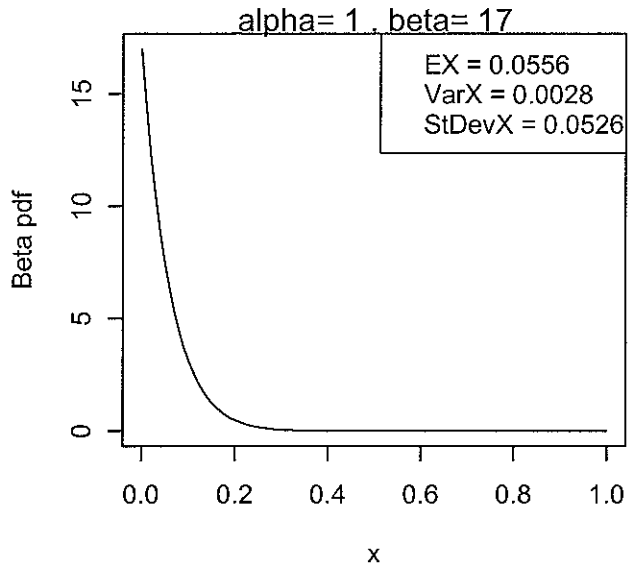
# MSE for different estimators of batting average



# MSE for different estimators of batting average



## Possible Prior Distributions



## Possible Prior Distributions

