Baseball and Bayes

Jo Hardin, Math 152

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

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

$\hat{\theta}_f$, 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 $\widehat{\theta}$ 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:

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

The Experiment

- John Spurrier will have n=10 at bats. The random variable, X, will be the number of base hits that he gets.
- Determining the prior probability: As a class we will find α and β that are consistent with our prior information.
- Comparison of the estimators:
  - $\widehat{\theta}_f = \frac{X}{n}$ $\qquad \widehat{\theta}_b = \frac{X+\alpha}{n+\alpha+\beta}$
  - We use Mean Squared Error (MSE) in the frequentist sense (that is, X is the random variable, θ is no longer random) to compare estimators (apples to apples):

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

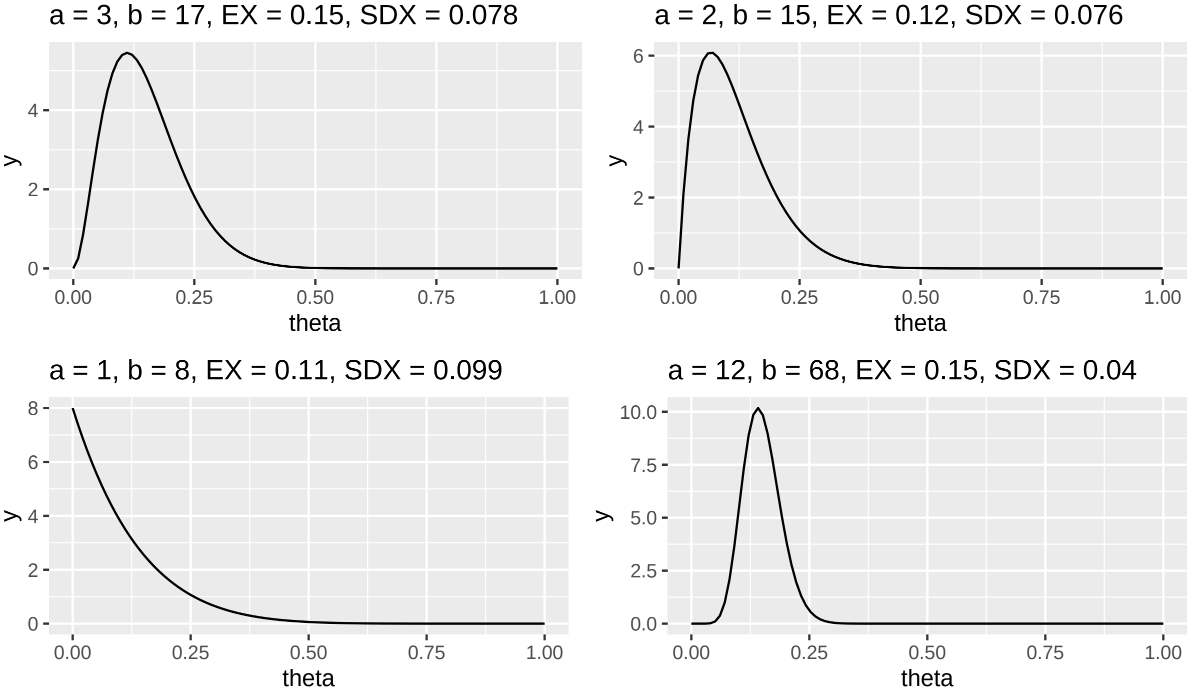
  - Under the assumption that X has a binomial distribution with parameters 10 and θ, calculate the mean and variance of X.
  - Using the mean and variance of X, what are the variance and bias of the two estimators?

Possible Prior Distributions

By trying out a variety of different values for a and b, the prior distributions can be visualized.

Which prior distribution should be used? The answer is that it depends! Of course, if we have a lot of information about the situation, we should use a steep prior that contains the known information. If our information is weak, we should use a flat prior.

Possible Prior Distributions I



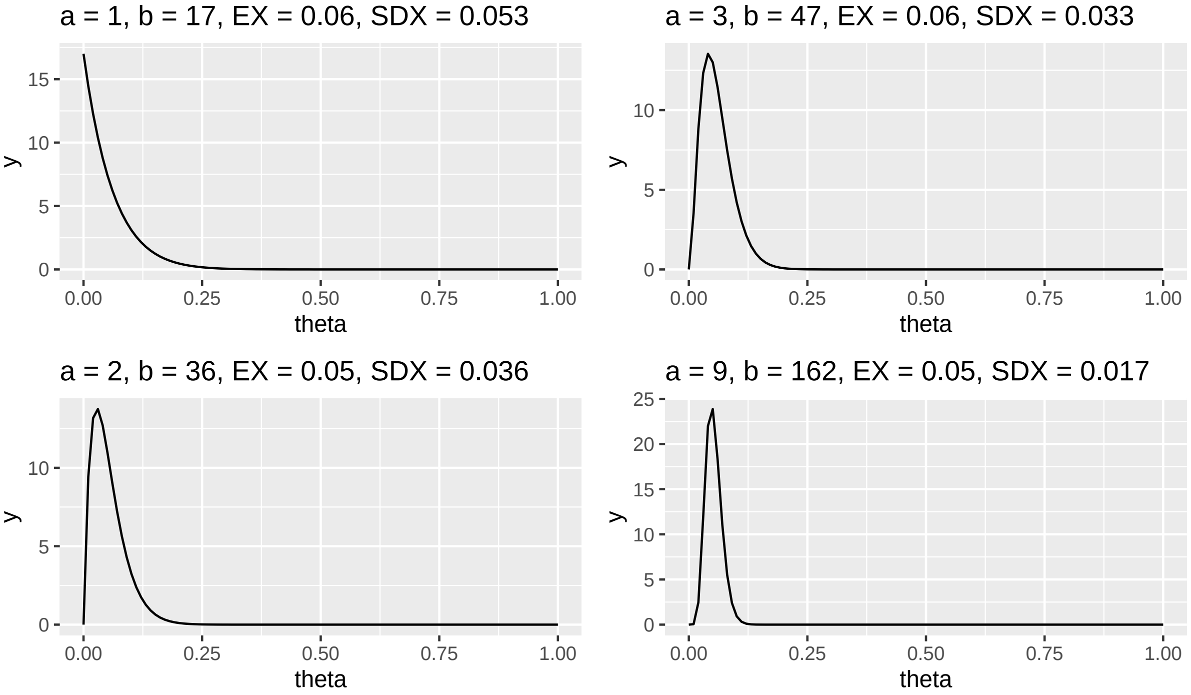
a = 3, b = 17, EX = 0.15, SDX = 0.078

a = 2, b = 15, EX = 0.12, SDX = 0.076

a = 1, b = 8, EX = 0.11, SDX = 0.099

a = 12, b = 68, EX = 0.15, SDX = 0.04

Possible Prior Distributions II



a = 1, b = 17, EX = 0.06, SDX = 0.053

a = 3, b = 47, EX = 0.06, SDX = 0.033

a = 2, b = 36, EX = 0.05, SDX = 0.036

a = 9, b = 162, EX = 0.05, SDX = 0.017

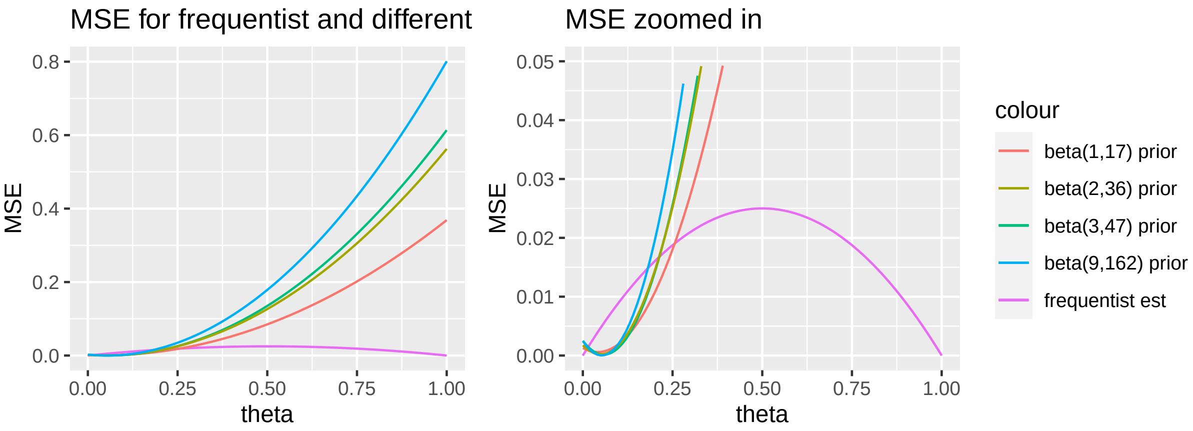
MSE

Mean Squared Error can be used to determine if the prior was the correct one to use, but only if we know the true value of θ!! In this case, we'll compare the MSE of the Bayesian estimator with the MSE of the frequentist estimator under various *truth* conditions. Because we are comparing apples to oranges (Bayesian vs. frequentist), we are forced to use the frequentist formulation of the MSE (there is no way to find an expected value or variance of the θ random variable under the frequentist paradigm).

Consider X to be the random variable with a Binomial(n=10, θ) distribution. In the Bayesian setting, $\hat{\theta} = (x+\alpha)/(n+\alpha+\beta)$. Deriving the MSE (as a function of θ) below is given as a homework problem.

$$\begin{aligned}
\text{MSE}_F(\hat{\theta}) &= \text{var}(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2 \\
&= \frac{(\alpha - \alpha\theta - \beta\theta)^2 + n\theta(1-\theta)}{(n+\alpha+\beta)^2}
\end{aligned}$$

The MSE can be used to assess the estimator (which may or may not be a function of the prior information). Note that the value on the x-axis is the **truth**, and the value on the y-axis is how good / bad the estimator is (as measured by mean squared error).


MSE for frequentist and different
MSE
theta
MSE zoomed in
MSE
theta
colour
beta(1,17) prior
beta(2,36) prior
beta(3,47) prior
beta(9,162) prior
frequentist est