Math 152 – Fall 2015
Jo Hardin
baseball & Bayes

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

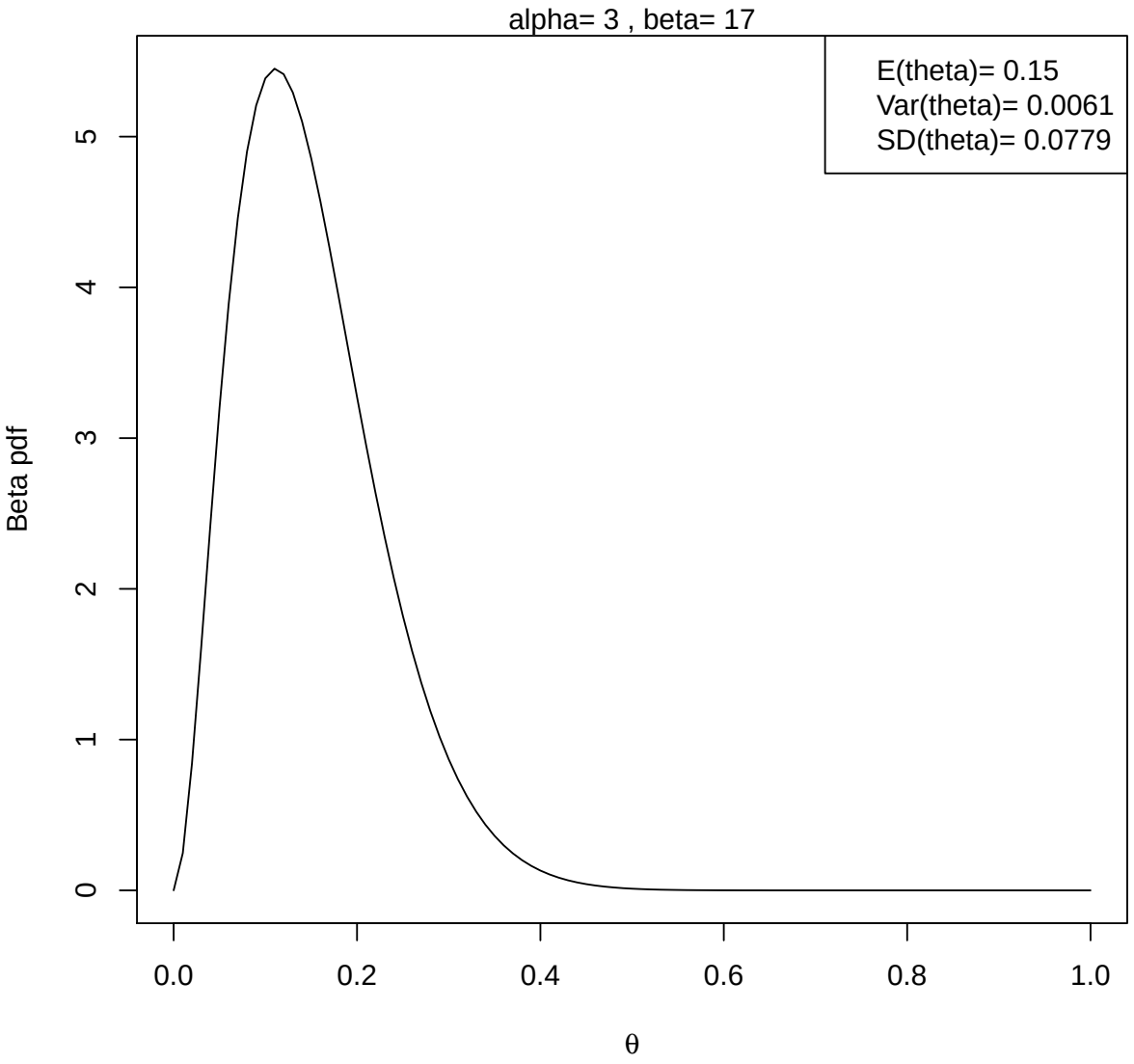
$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + bias^2(\hat{\theta}) = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

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

Creating the Beta function (to use in plotting)

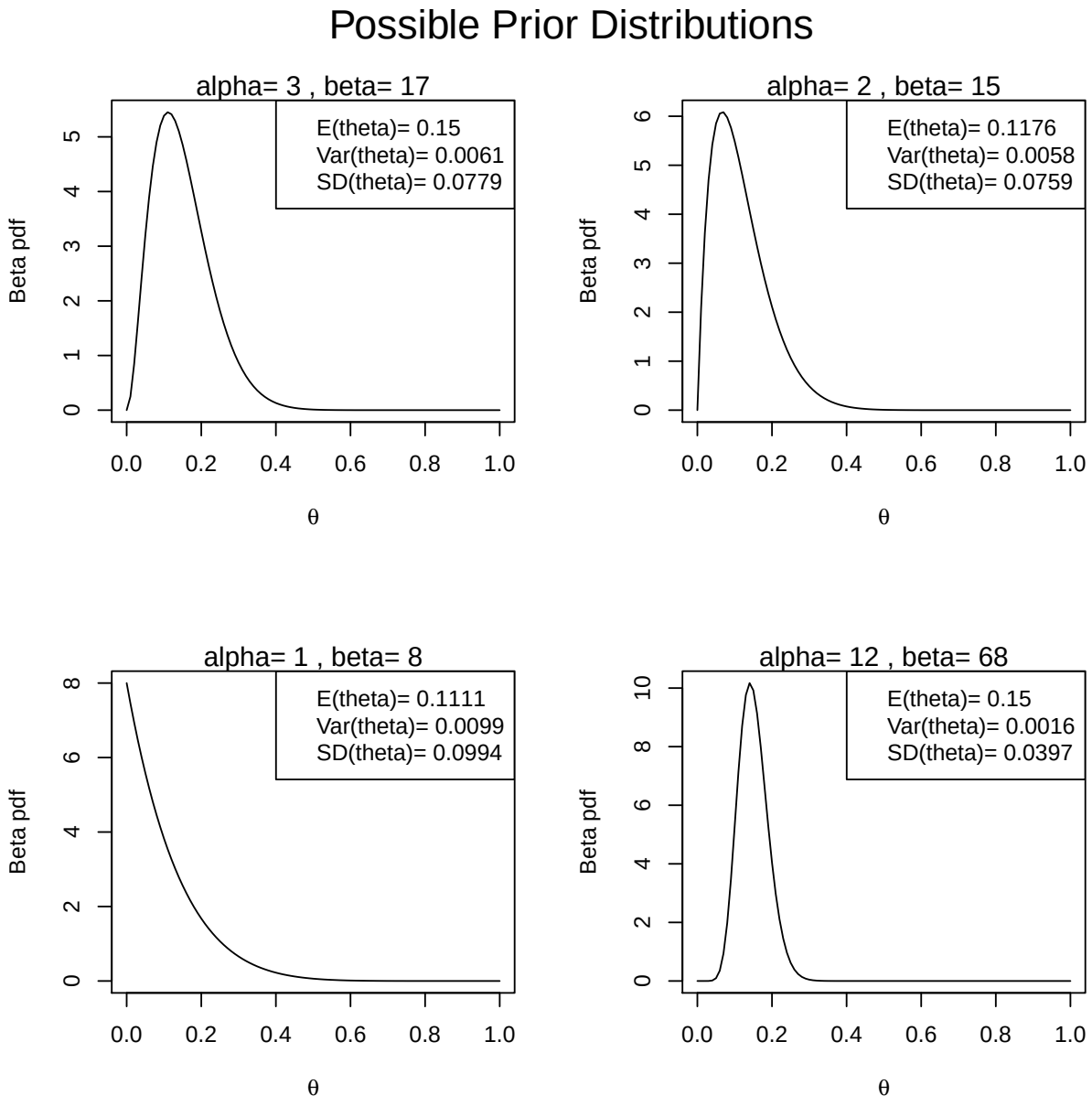
```
# note n=10 here

x<-seq(0,1,.01) #x is his true batting average

betaplot<-function(x,a,b){
plot(x,dbeta(x,a,b),type="l",xlab=expression(theta),ylab="Beta pdf")
ex<-a/(a+b)
varx<-a*b/((a+b)^2*(a+b+1))

legend(x="topright",c(paste("E(theta)=",round(ex,4)),
                      paste("Var(theta)=",round(varx,4)),
                 paste("SD(theta)=", round(sqrt(varx),4))))
mtext(paste("alpha=",a,", beta=",b))}
```

```r
hit<-seq(0,10,1)
post.mn<-function(a,b,hit=hit){(a+hit) / (10+a+b)}

betaplot(x,a=3,b=17)
```

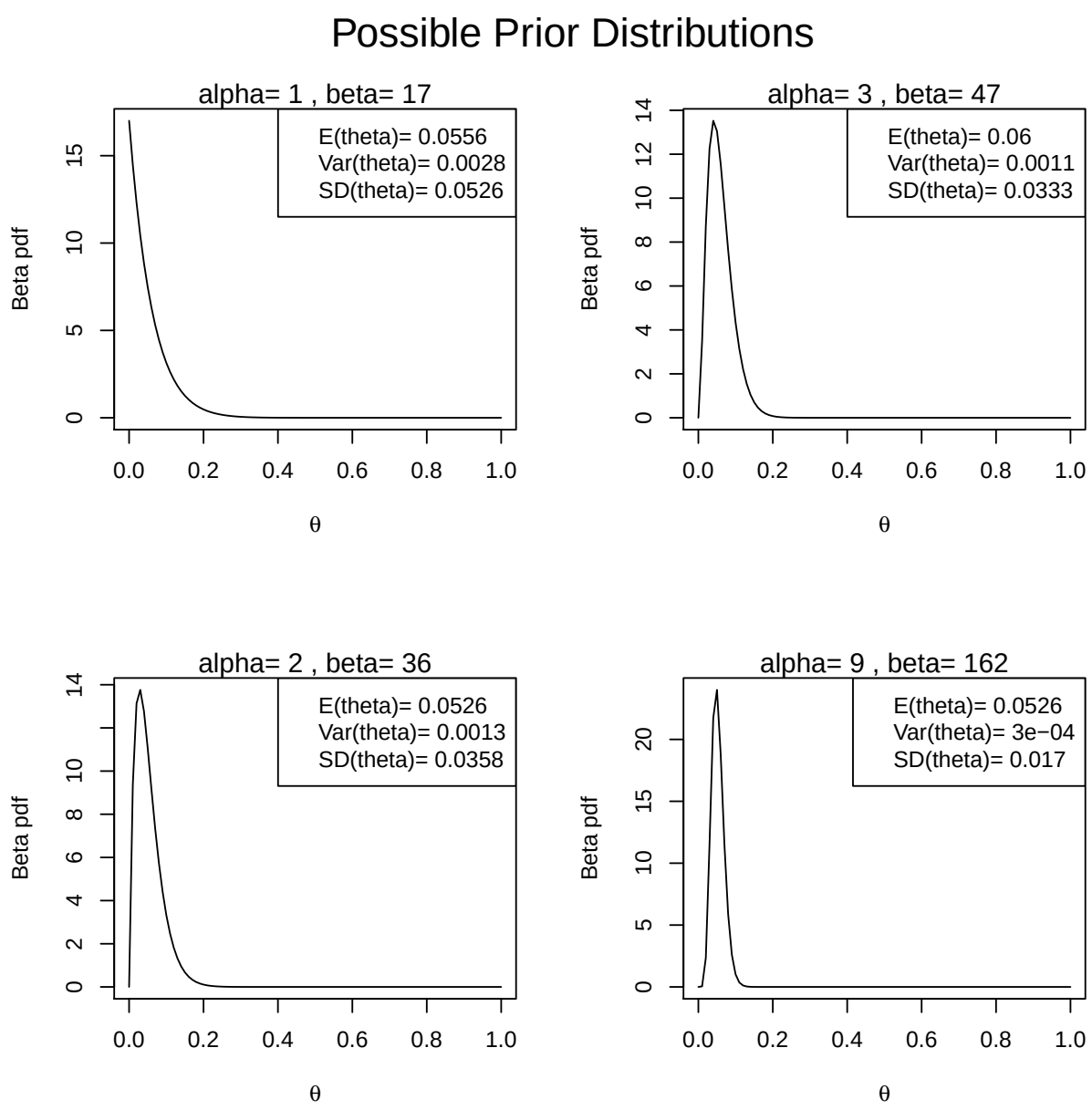


Priors

```
par(mfrow=c(2,2))
betaplot(x,a=3,b=17)
betaplot(x,a=2,b=15)
betaplot(x,a=1,b=8)
betaplot(x,a=12,b=68)
mtext("Possible Prior Distributions",line=-2,cex=1.5,outer=T)
```

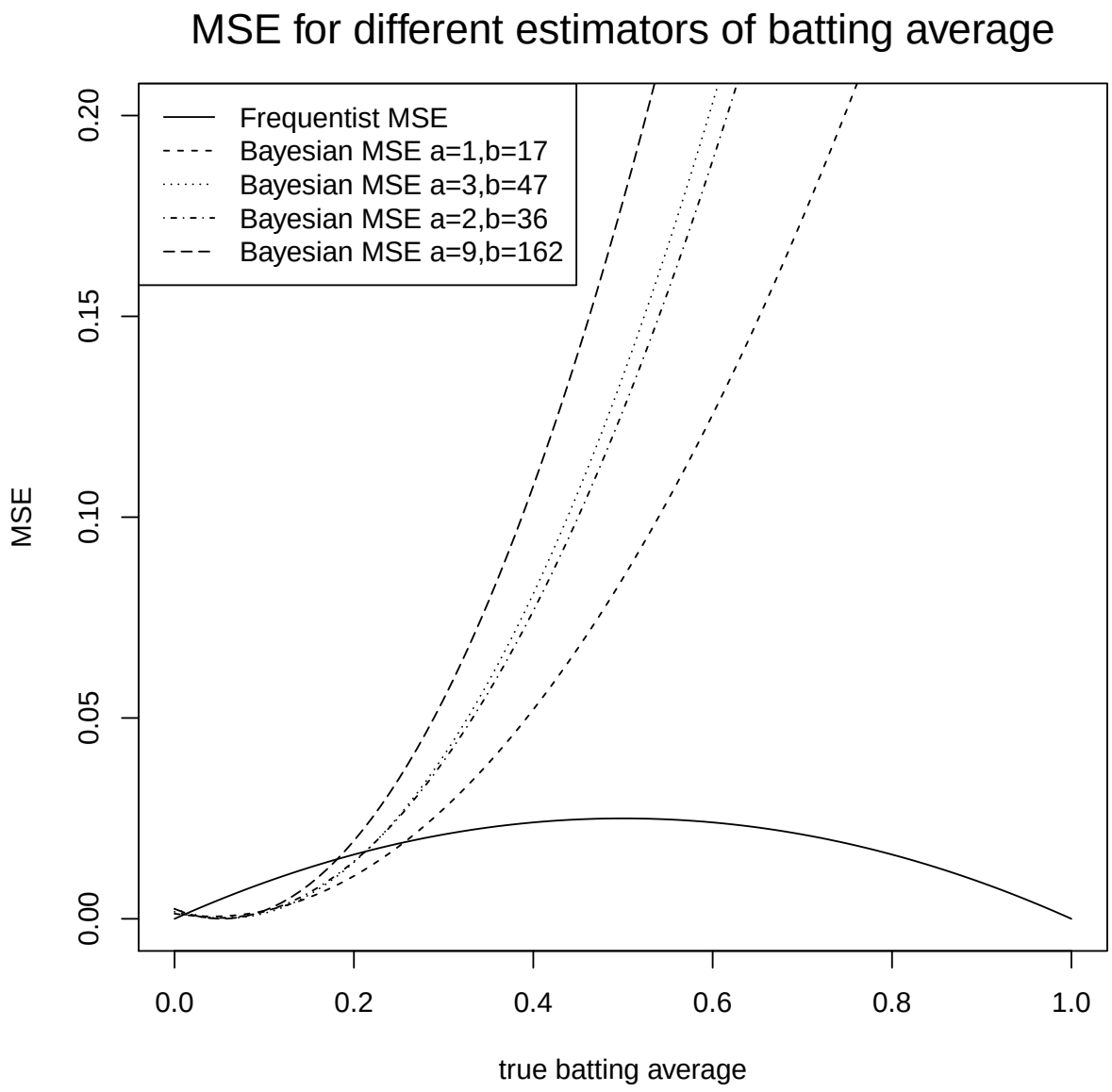


```r
par(mfrow=c(2,2))
betaplot(x,a=1,b=17)
betaplot(x,a=3,b=47)
betaplot(x,a=2,b=36)
betaplot(x,a=9,b=162)
mtext("Possible Prior Distributions",line=-2,cex=1.5,outer=T)
```

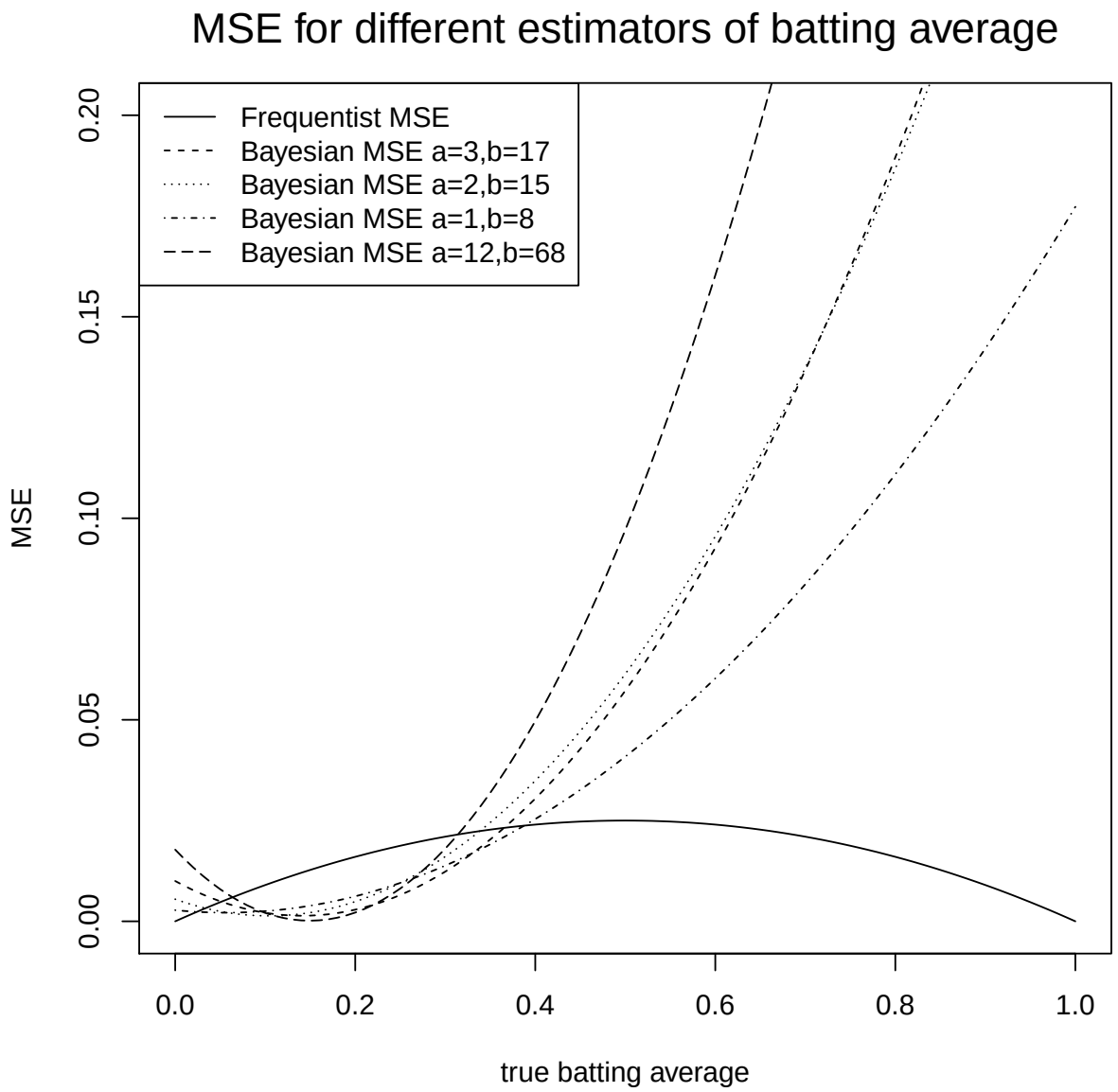


MSE

```
plot(x,x*(1-x)/10,type="l",lty=1,xlab="true batting average",
     ylab="MSE",ylim=c(0,.2)) #mse.f
lines(x,x*(1-x)/78.4 + ((10*x+1)/28 - x)^2,lty=2) #mse.b (a=1,b=17)
lines(x,x*(1-x)/360 + ((10*x+3)/60 - x)^2,lty=3) #mse.b (a=3,b=47)
lines(x,x*(1-x)/230.4 + ((10*x+2)/48 - x)^2,lty=4) #mse.b (a=2, b=36)
lines(x,x*(1-x)/3276.1 + ((10*x+9)/181 - x)^2,lty=5) #mse.b (a=9,b=162)
legend(x="topleft",c("Frequentist MSE", "Bayesian MSE a=1,b=17",
                "Bayesian MSE a=3,b=47", "Bayesian MSE a=2,b=36",
                "Bayesian MSE a=9,b=162"),lty=c(1:5))
mtext("MSE for different estimators of batting average",line=1,cex=1.5)
```



```r
plot(x,x*(1-x)/10,type="l",lty=1,xlab="true batting average",
     ylab="MSE",ylim=c(0,.2)) #mse.f
lines(x,x*(1-x)/90 + ((10*x+3)/30 - x)^2,lty=2) #mse.b (a=3,b=17)
lines(x,x*(1-x)/72.9 + ((10*x+2)/27 - x)^2,lty=3) #mse.b (a=2,b=15)
lines(x,x*(1-x)/36.1 + ((10*x+1)/19 - x)^2,lty=4) #mse.b (a=1,b=8)
lines(x,x*(1-x)/810 + ((10*x+12)/90 - x)^2,lty=5) #mse.b (a=12,b=68)
legend(x="topleft",c("Frequentist MSE", "Bayesian MSE a=3,b=17",
                "Bayesian MSE a=2,b=15", "Bayesian MSE a=1,b=8",
                "Bayesian MSE a=12,b=68"),lty=c(1:5))
mtext("MSE for different estimators of batting average",line=1,cex=1.5)
```



caveat Note, in order to compare MSE for Bayesian and Frequentist settings, we need to compare apples to appples. In the Bayesian setting, $\hat{\theta} = (x + \alpha)/(n + \alpha + \beta)$. So to find the "Bayesian MSE", use the Frequentist idea of variability, and find the var of X (and bias), and treat α, β, n as constants.