Applied Bayesian Analysis : NCSU ST 540

Homework 1

Bruce Campbell

Bayesian analysis of Binomial Model

In this assignment we are performing a Bayesian analysis of count data. The response $Y \sim Binomial(N, \theta)$ measures the number of successes of N independent Bernoulli trials. The parameter θ is unknown and modeled with a Beta prior $\theta \sim Beta(\alpha, \beta)$. This is a conjugate prior for the Binomial and makes for a Beta distributed posterior $\theta|Y \sim Beta(Y + \alpha, N - Y + \beta)$ For clarity, we work this last statement out below. Bayes theorem $P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$ is our starting point. The likelihood of our model is $P(Y|\theta) = \binom{N}{Y}\theta^Y(1-\theta)^{N-Y}$ so putting this together with our prior and dropping terms without θ we see that

$$P(\theta|Y) \propto \theta^{\alpha-1+Y}(1-\theta)^{\beta-1+N-Y}$$

which we recognize as the kernel of a $Beta(Y + \alpha, N - Y + \beta)$ distributed random variable.

(1) R function for plotting the prior and posterior

Write an R function that takes Y , N, a, and b as inputs. The function should produce a plot (clearly labeled!) that overlays the prior and posterior density functions (both using the dbeta function), and it should return a list with the posterior mean and posterior standard deviation.

```
makePlots <- function(Y, N, a = 1, b = 1) {
    titleString <- paste("Prior and Posterior for Binomial Analysis \n(N,Y,a,b) =(",
        N, ",", Y, ",", a, ",", b, ")", sep = " ")

    # Calculate prior mean and variance
    priorMean <- a/(a + b)
    priorVariance <- a * b/((a + b)^2 * (a +
        b + 1))
    priorSD <- sqrt(priorVariance)

    # Parameters for the posterior
    # distribution
    posteriorA <- Y + a
    posteriorB <- N - Y + b

    # Calculate posterior mean and variance
    posteriorMean <- posteriorA/(posteriorA +
        posteriorB)
```

```r
    posteriorVariance <- posteriorA * posteriorB/((posteriorA +
        posteriorB)^2 * (posteriorA + posteriorB +
        1))
    posteriorSD <- sqrt(posteriorVariance)

    # We use ggplot2's stat_function instead
    # of curve
    p1 <- ggplot(data.frame(x = c(0, 1)),
        aes(x)) + stat_function(fun = function(x) dbeta(x,
        shape1 = a, shape2 = b), aes(colour = "prior")) +
        ylab("density") + xlab(TeX("$\\theta$"))
    p2 <- p1 + stat_function(fun = function(x) dbeta(x,
        shape1 = posteriorA, shape2 = posteriorB),
        aes(colour = "posterior")) + ggtitle(titleString) +
        scale_colour_manual("Density", values = c("red",
            "blue"))
    print(p2)

    # Place the posterior mean and standard
    # deviation in a list to return to the
    # user.
    results <- list(posteriorMean = posteriorMean,
        posteriorSD = posteriorSD)

    return(results)
}
```

(2) Non informative Prior

What values of a and b would make good default values to represent a prior that carries little information about θ? Make these the default values in your function.

An uninformative prior is provided by setting the parameters to $a = 1$ $b = 1$ since setting these values in the density

$f_\theta(x) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^0(1-\theta)^0 = 1$ shows us that the prior is uniform $U[0,1]$.

(3) What values of a and b give prior mean 0.7 and prior standard deviation 0.2?

To solve this we use the expressions for the mean and variance of a beta distributed random variable $Y \sim Beta(\alpha, \beta)$

$E[Y] = \frac{\alpha}{\alpha+\beta}$ and $Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. working out the algebra for the case $E(Y) = 0.7$ and $Var(Y) = 0.04$ yields the results provided by the R code below.

```r
# Informed prior parameters
bp <- sqrt(1/(3 + 1/3) * ((2 + 1/3)/(0.04 *
    (3 + 1/3)^2) - 1))
ap <- 0.7/0.3 * bp

pander(data.frame(ap = ap, bp = bp),
    caption = "paramters when prior mean=0.7 and prior standard deviation=.2")
```

Table 1: paramters when prior mean=0.7 and prior standard deviation=.2

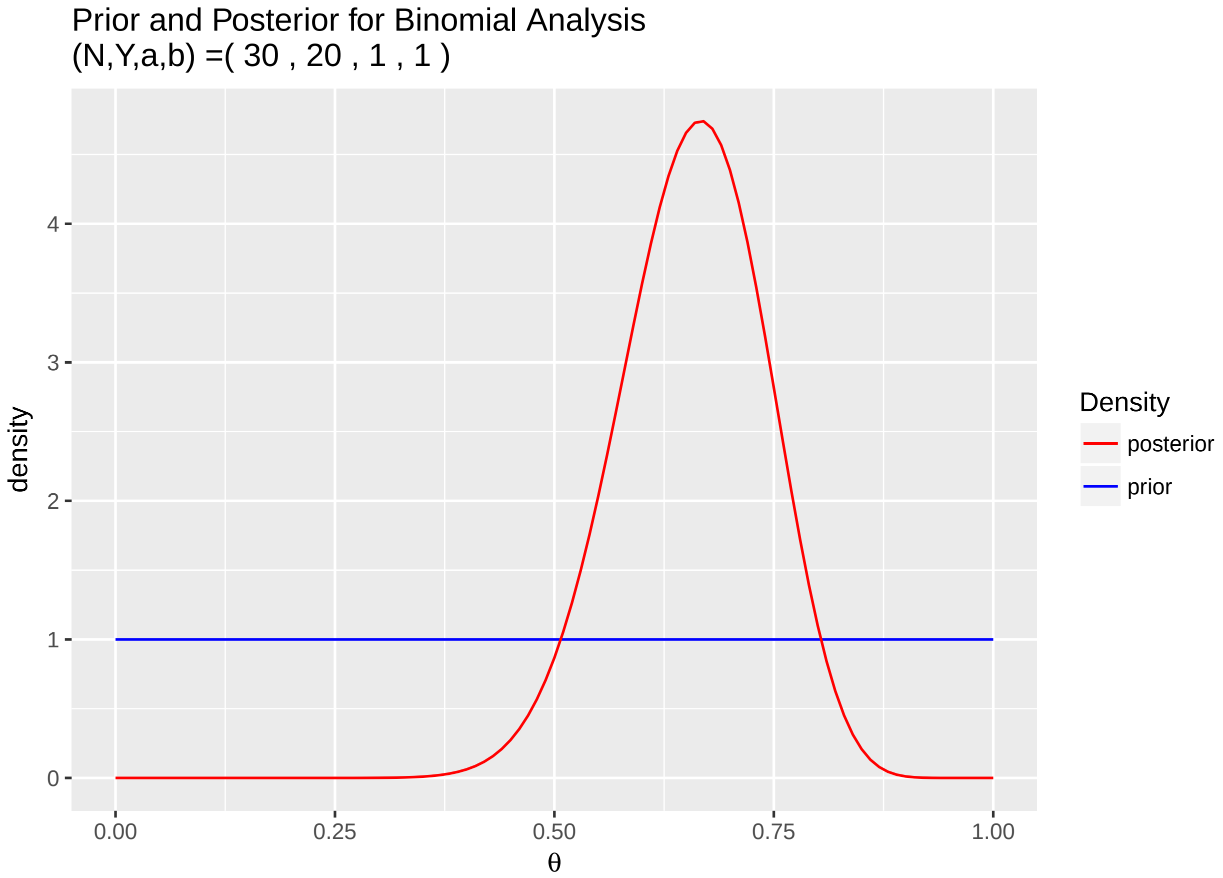
ap	bp
2.635	1.129

(4) Analysis

Now we observe Y = 20 events in N = 30 trials. Use your code from (1) to conduct a Bayesian analysis of these data. Perform the analysis twice, once with the uninformative prior from (2) and once with the informative prior in (3).

```
Y <- 20
N <- 30

resultsUninformative <- makePlots(Y,
    N, 1, 1)
```

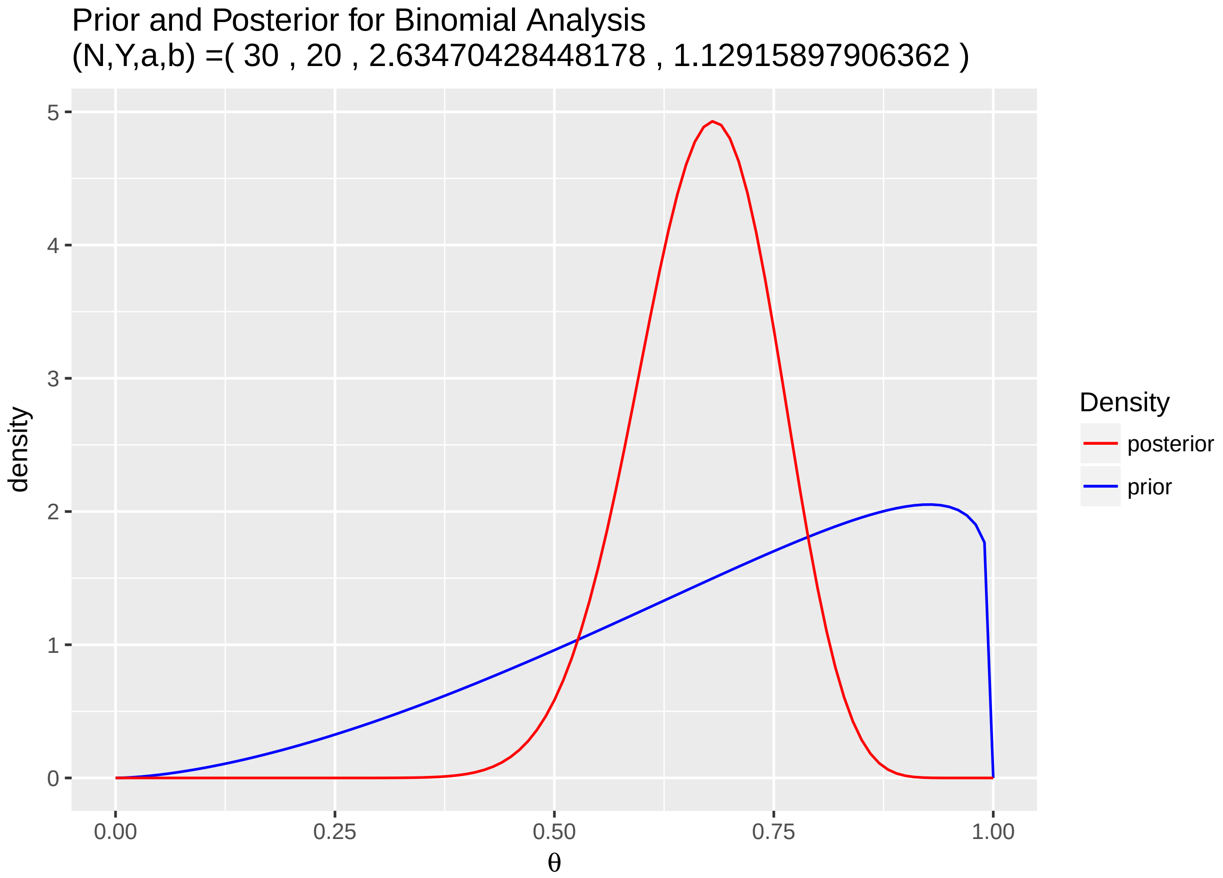


```
pander(data.frame(resultsUninformative),
    caption = "Posterior Mean and SD using Uninformative Prior ")
```

Table 2: Posterior Mean and SD using Uninformative Prior

posteriorMean	posteriorSD
0.6562	0.08268

```
resultsInformative <- makePlots(Y, N,
    ap, bp)
```



```r
pander(data.frame(resultsInformative),
    caption = "Posterior Mean and SD using Informative Prior ")
```

Table 3: Posterior Mean and SD using Informative Prior

posteriorMean	posteriorSD
0.6704	0.07973

(5) Summary

Summarize the results. In particular, how does this analysis compare to a frequentist analysis.

We note that the Bayesian analysis with both priors provide similar results. The frequentist approach is to use the maximum likelihood estimator of θ assuming a $Binomial(N, \theta)$ model for Y. This is the same likelihood in Bayes theorem above. We're only given one data element so the likelihoods agree in both cases but it's straightforward to extend to more samples. The MLE for the single sample case is $\frac{y}{N}$ which is close to the results obtained in the Bayesian analysis.