Bayes: Homework 2

Leslie Gains-Germain

1. The mathematical definition of probability out of Casella and Berger's *Statistical Inference*:

   Given a sample space S and an associated sigma algebra B, a probability function is a function P with domain B that satisfies

   (a) $P(A) \geq 0$ *for all* $A \in B$

   (b) $P(S) = 1$

   (c) $If A_1, A_2, ... \in B$ *are pairwise disjoint, then* $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   Two ways that we use probability to summarize the world around us:

   - We use probabilities to describe how likely an outcome is in a game of chance. For example, if we roll a fair die and we have no knowledge of the outcome, the probability of rolling a 3 is 1/6. We can also assign probabilities to the outcomes of flipping a coin, spinning a spinner, or picking a random card from a deck.
   - The second way we use probabilities is to describe our degree of certainty about something happening in the future. For example, I might say to my boyfriend, "There is a 10% chance that I will do the dishes tonight." I am trying to convey to him that it is not very likely that I will be doing the dishes, and I use the words "probability" and "chance" to convey that idea. I don't necessarily want to completely rule out the possibility of me doing the dishes, because I figure there is a 'chance' that he will become injured during the day. Even when I'm trying to remember something, like the name of someone from my childhood, I'll catch myself saying something like, "I'm 99% sure that her name was Nina." Sometimes, our use of probability can be based on something more concrete but is still used to convey uncertainty. For example, suppose I say that the chance Serena Williams will win the US open is 90%. I'm basing this off her previous performance this calendar year, and I'm using that as reasoning for my degree of certainty in the outcome of the tournament.

2. A frequentist's criticism of the likelihood principle is that likelihood inference about model parameters is the same in two studies with different designs but the same sample data. Consider the example given on the Wikipedia page for the Likelihood principle. Let X be a random variable that represents that number of heads in 12 flips of an unfair coin, where θ is the probability of heads in each trial. Then $X \sim Bin(12, \theta)$. If the experiment is conducted and we observe 3 heads, $X = 3$, then the likelihood function is as follows:

$$L(\theta|X = 3) = \binom{12}{3}\theta^3(1-\theta)^9 = 220\theta^3(1-\theta)^9$$

Now, let Y be the number of flips needed to get three heads. If we observe the same sample data as the previous example, then $Y = 12$. The likelihood function for this example is as follows:

$$L(\theta|Y = 12) = \binom{11}{2}\theta^3(1-\theta)^9 = 55\theta^3(1-\theta)^9$$

The first likelihood function is a scalar multiple of the second, so inference about the probability of getting heads, θ, is exactly the same in these two scenarios (the estimates will be the same and likelihood based confidence intervals will be the same). This is consistent with the likelihood principle because the observed data are the same in both cases (three success in 12 flips, or 12 flips to get three successes). Frequentists argue, however, that the study design - whether we count the number of flips needed to get three heads or the number of heads in 12 flips - should be considered when making inference about θ.

3. See attached handwritten sheet.

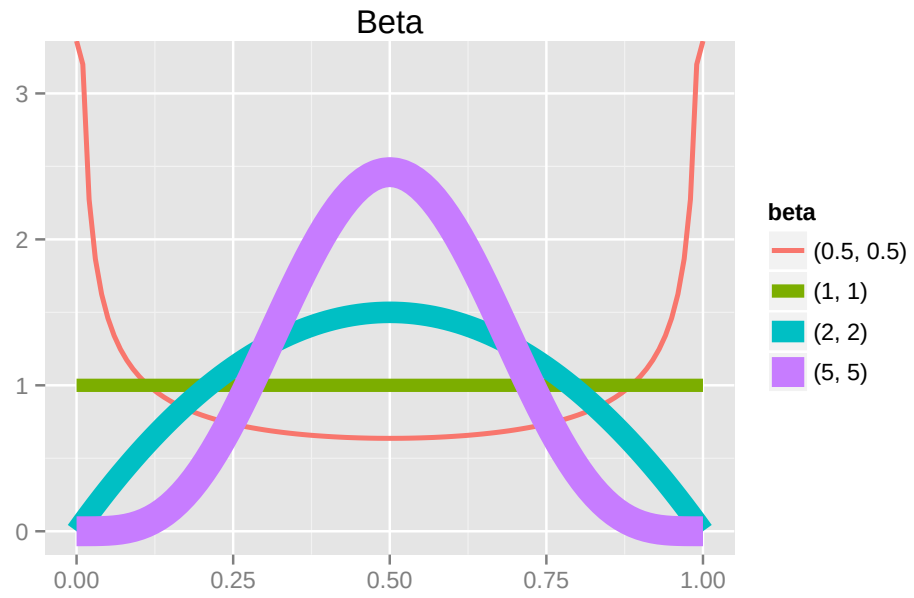
4. I am going to use a simple example to explain this concept to someone with little mathematical

background. Say you are given an unfair coin, and the probability of getting heads is θ. The natural way to learn something about θ is to pick up the coin and start flipping it. The proportion of heads you get in 10 tosses is clearly going to be your best guess for θ. This is exactly what the likelihood function does, it tells us how likely values of θ are given observed sample data. The likelihood function, $L(\theta|x)$, is a function of θ and gives us a likelihood for each possible value of θ, given that the sample data have been observed and are fixed. In the example, suppose we observe 8 heads in 10 tosses of the coin. Then the likelihood function will tell us that the most likely value for θ is 8/10 or 4/5. Values such as 7/10 and 9/10 will be likely (although slightly less likely than 4/5), and and a value such as 1/10 will not be likely at all. A probability distribution, on the other hand, describes the probability of observing sample data, given the value of θ is fixed. In the example given, suppose we knew that the probability of getting heads with the unfair coin was 3/4. Then we could calculate the probability of getting $1, 2, 3, 4, ...$, or 10 heads in 10 tosses of the coin, given that the probability is 3/4 on each flip. The difference between the likelihood function and the probability function lies in what we think about as fixed and what we think about as varying. The sample data varies in a probability function with θ fixed, and θ varies in a likelihood function, with the sample data fixed.

5. Suppose that the true probability of duckling survival is $\theta = 1/2$. Then the probability that all 24 ducklings survive in a week is $1/24^2$, a number very close to zero. Frequentists would estimate θ to be 1, the observed proportion of birds that survived. This estimate doesn't make sense because everyone knows bird survival rate is not 1. Likelihoodists would also estimate θ to be 1 - the MLE. In both frequentist and likelihoodist land, the estimates for θ are nonsense because of this one bad sample. In Bayesian land, the unrepresentative sample will still influence estimates, but at least the sample is not the *only* piece of information going into the posterior predictive distribution. For example, if we use a Beta(2,2) prior, the posterior distribution, would be a Beta(26,2). The estimator for θ under squared error loss is then $26/28 = 0.93$. This estimator is still off, but at least it makes more sense than the

likelihood and frequentist estimator of 1. Note: it turns out that if we were to construct uncertainty intervals, none of the three methods would capture the true value of θ, but the Bayesian interval would probably be closer to capturing the truth.

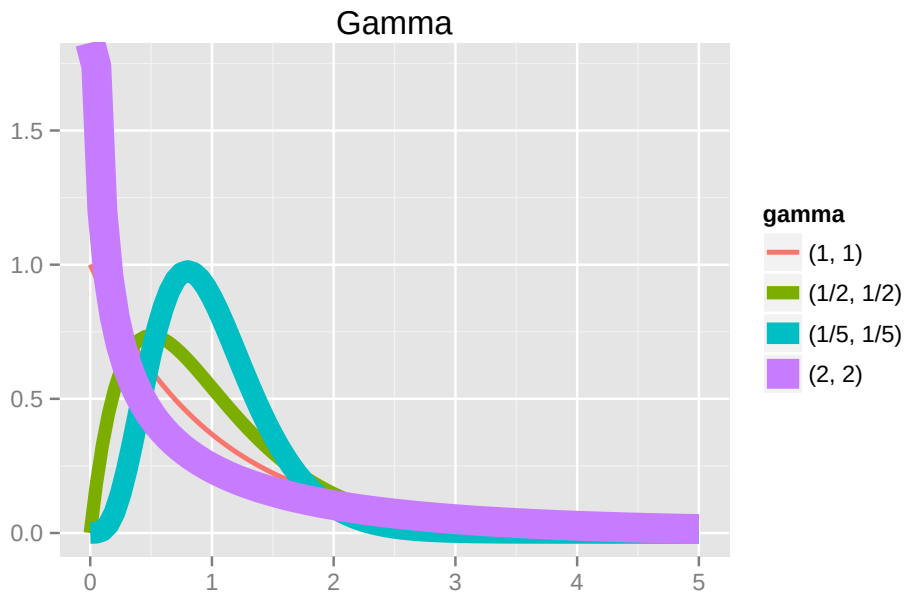
6. A plot of beta distributions with four different parameter values of $\alpha = \beta$ is shown below.



7. Note that the parameterization of the gamma distribution given by Gelman et. al is technically a gamma* distribution on Boik's equation sheet. This means that when Gelman talks about a Gam(1/2, 1/2), he is really talking about a Gam*(1/2, 1/2) which is a Gam(2, 2) distribution according to Boik, R, Wikipedia, and most everyone else. In the plot below, I am using Gelman's parameterization. A gamma prior is often used for the precision parameter (1/variance) in a normal distribution. For example, if we want large values of the variance most likely, then we choose a Gam(2, 2) prior for the precision. This makes small values of the precision and large values of the variance most likely. The distribution of the variance (1/precision) follows an inverse gamma distribution, which also has nice properties. I think the gamma distribution is a natural choice as a prior for the precision because of the shape of the distribution. If we choose a Gam(1,1) or a Gam(2,2) prior on the precision, small values

of the precision are the most likely, and larger values of the precision are less and less likely. By changing the parameters of the gamma distribution, we can change how much weight we want to put on small and large values of the precision.

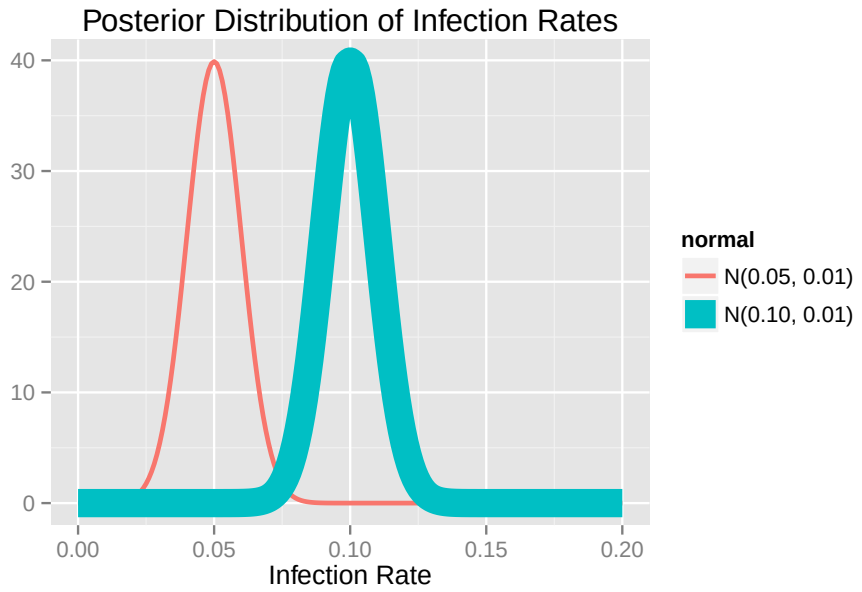
I'm guessing that Gelman uses a different parameterization for the gamma distribution than most to help the user remember what prior to choose if he/she wants to make large values of the variance most likely. The larger the α, β parameters in Gelman's parameterization of the gamma prior makes *larger* values of the variance more likely. With Boik's parameterization of the Gamma distribution, you would need to use a gamma prior with *smaller* parameter values to make larger values of the variance more likely.



8. (a) See my work below. The probability of being infected if the person tests positive is 0.326 for Doctor 1's belief and 0.505 for Doctor 2's belief about the infection rate.

$$
\begin{aligned}
P(disease|testpos) &= \frac{P(testposanddiseased)}{P(testpos)} \\
&= \frac{P(testpos|disease)P(disease)}{P(testpos|disease)P(disease) + P(testpos|nodisease)P(nodisease)} \\
Doc1 &= \frac{0.92 * .05}{0.92 * 0.05 + 0.10 * 0.95} = 0.326 \\
Doc2 &= \frac{0.92 * .10}{0.92 * 0.10 + 0.10 * 0.90} = 0.505
\end{aligned}
$$

(b) I chose to use a normal prior distribution to reflect each Doctor's belief. I chose the mean to be 0.05 and 0.10 for each doctor, respectively. I chose the standard deviation so that their beliefs would not overlap too much. It looks like Doctor 1 would think of $3 - 7\%$ as reasonable values, and Doctor 2 would think of $8 - 12\%$ as reasonable values for the infection rates.



9. Based off what I could find about this first Bayesian analysis done by Laplace, it sounds like he used a weakly informative Beta(2,2) prior for θ. Below, I use Bayes rule to find the posterior distribution of θ based on the birth data from Laplace's time. Let X be the number of female births out of the 493472 births in Paris from 1745 to 1770. Let θ represent the probability

that any birth is female. Then, by Bayes rule:

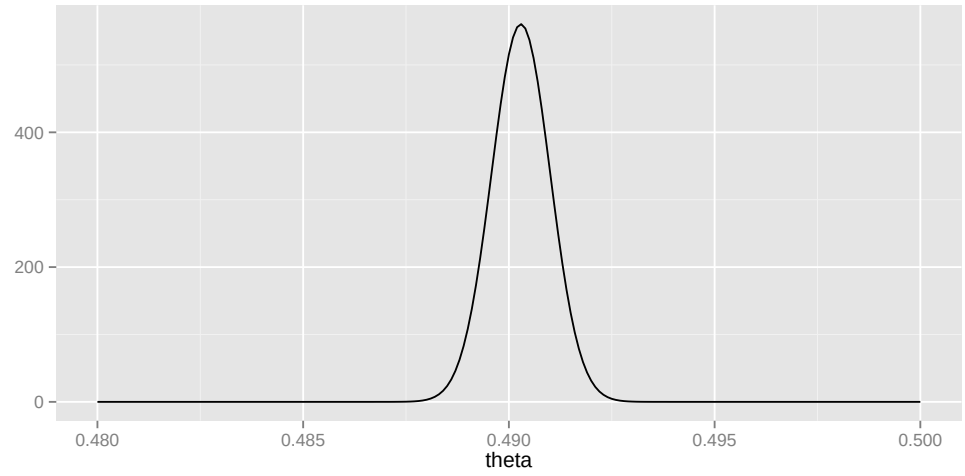
$$P(\theta|X=x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{1}$$

$$= \frac{\binom{493472}{x}\theta^x(1-\theta)^{493472-x} * 6\theta(1-\theta)}{\int_0^1 \binom{493472}{x}\theta^x(1-\theta)^{493472-x} * 6\theta(1-\theta)d\theta} \tag{2}$$

$$= \frac{6\binom{493472}{x}\theta^{x+1}(1-\theta)^{493473-x}}{6\binom{493472}{x} * \beta(x+2, 493474-x)\int_0^1 \frac{\theta^{x+1}(1-\theta)^{493473-x}}{\beta(x+2,493474-x)}} \tag{3}$$

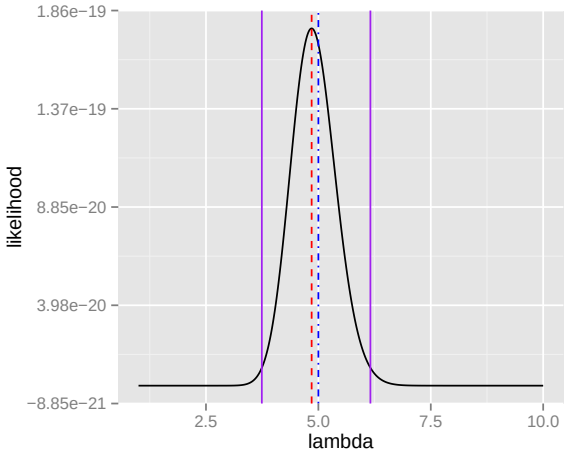
$$= \frac{\theta^{x+1}(1-\theta)^{493473-x}}{\beta(x+2, 493474-x)} \tag{4}$$

Note that the integral is equal to 1 in line 3, because we integrate the pdf for a Beta(x+2, 493474-x) distribution. We then find that the posterior distribution for θ, given the observed $X = 241945$, is a $Beta(241947, 251529)$ distribution. This distribution is plotted below, the mean is $241947/493476 = 0.4902913$, and the standard deviation is approximately 0.0007. The mean is only very slightly larger than the observed proportion of girls in the sample, $241945/493472 = 0.4902912$. The probability that θ is less than 0.5 is very very close to 1 because almost all of the density of the posterior probability distribution for θ lies below 0.5. There is very strong evidence that θ is less than 0.5.

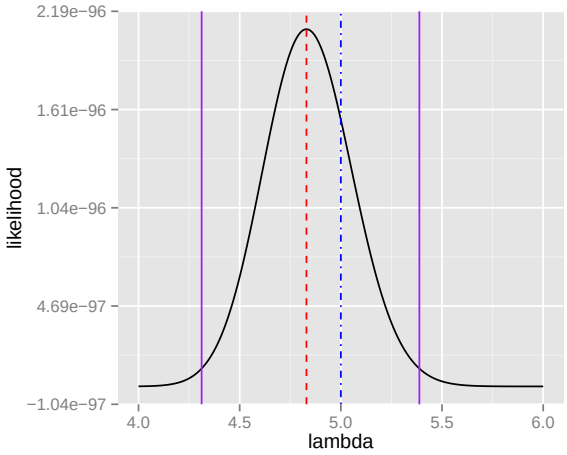


10. (a) In the plot below, the dashed line is the mle, which is also the mean of the likelihood

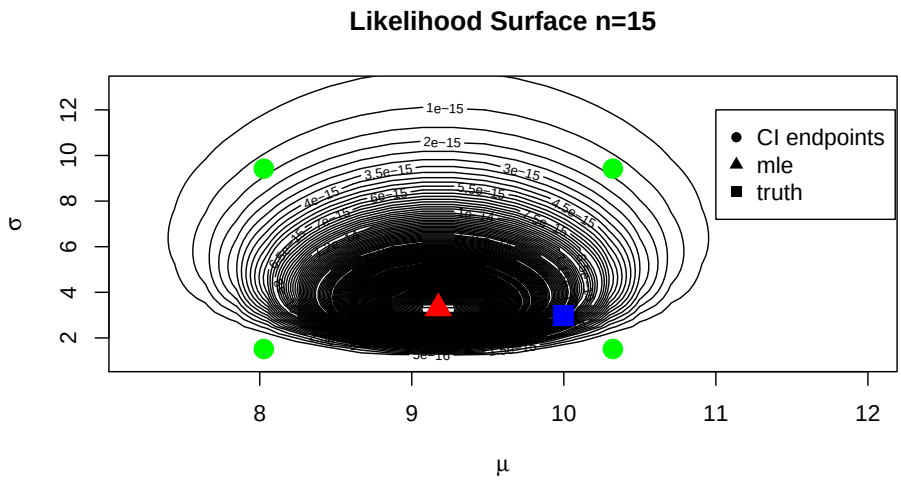
function, 4.85. The endpoints of the 95% confidence interval are 3.74 and 6.16. These endpoints were found by find the values of λ that have a likelihood that is 5% of the likelihood at the mle.



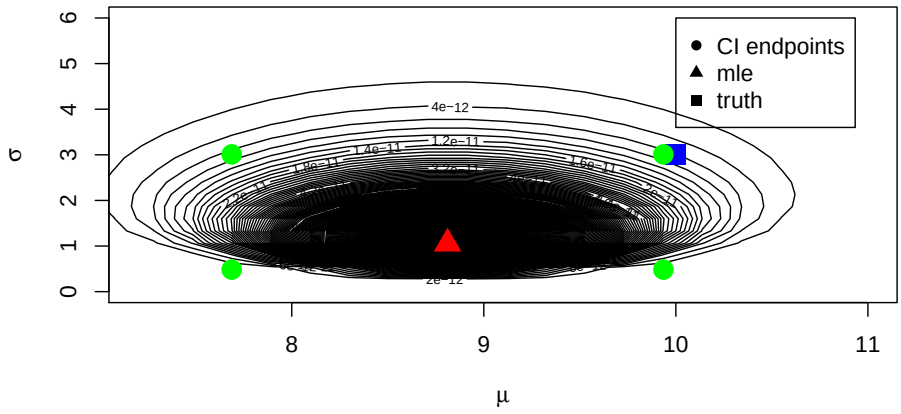
(b) With a sample size of 100 observations, the mle is 4.83, and the endpoints of the 95% likelihood interval are 4.31 and 5.39. Although the accuracy of the mle has not improved, the width of the interval has decreased.



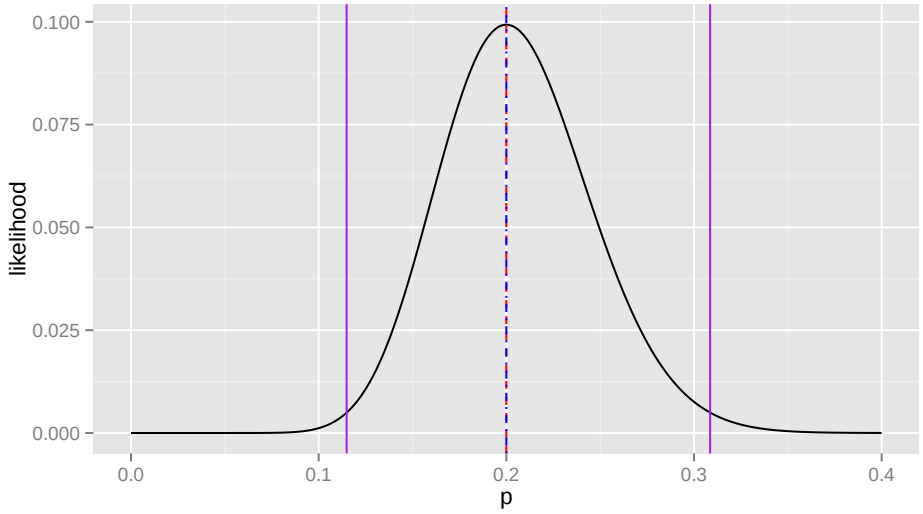
(c) The contour plot is shown below. The mles were found by maximizing the three dimensional likelihood surface with respect to μ and σ^2. With a sample size of 15 observations, the mle for μ is 9.17, and the mle for σ^2 is 3.30. The *profile* 95% confidence intervals are $(8.03, 10.32)$ and $(1.51, 9.42)$ for μ and σ^2 respectively.



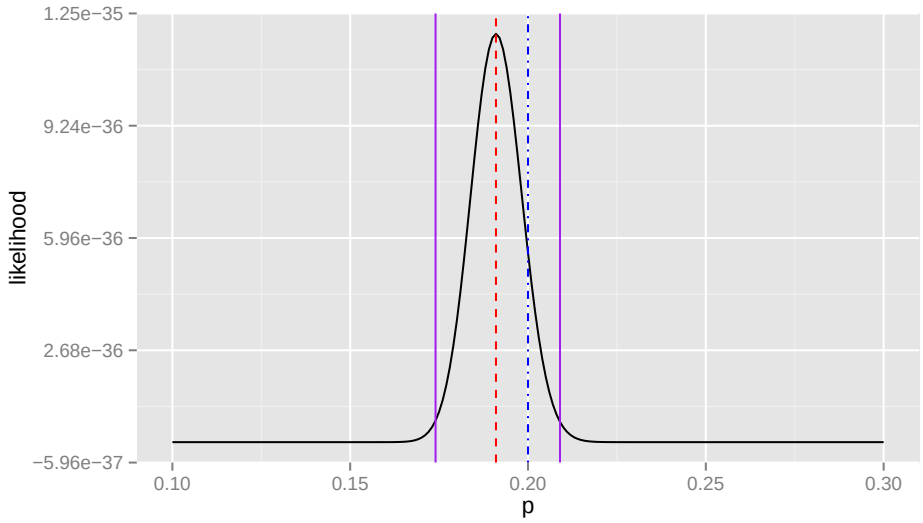
(d) With a sample size of 5, the mles are farther from the truth, and neither of the profile confidence intervals contain their respective true value. The mle for μ is 8.81 and the mle for σ^2 is 1.05. The 95% profile confidence interval for μ is $(7.69, 9.94)$, and the 95% profile confidence interval for σ^2 is $(0.48, 3.01)$.



(e) For the binomial likelihood, I assumed that m was fixed at 100 and found the MLE for p. The MLE for p was 0.2 (the same as the truth) because it just so happened that in the sample drawn, there were 20 successes. The 95% confidence interval for p is 0.11 to 0.31.

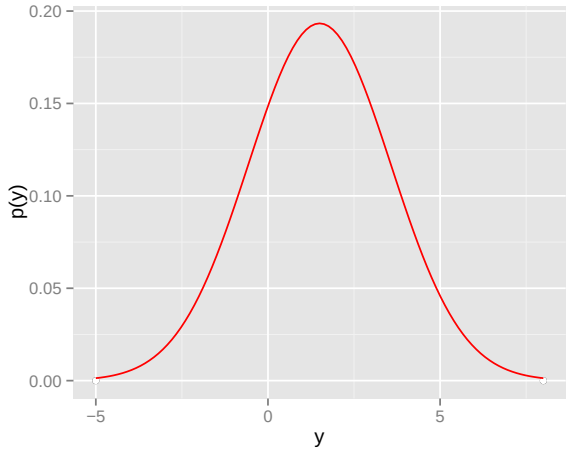


(f) Again, I assumed that m was fixed at 100 and found the MLE for p. The MLE for p was 0.19, which is actually farther from the truth than the sample of 1. But, the 95% confidence interval for p is much narrower (0.17, 0.21).



11. (a) The formula for the marginal probability density for y is shown below, as well as the density curve.

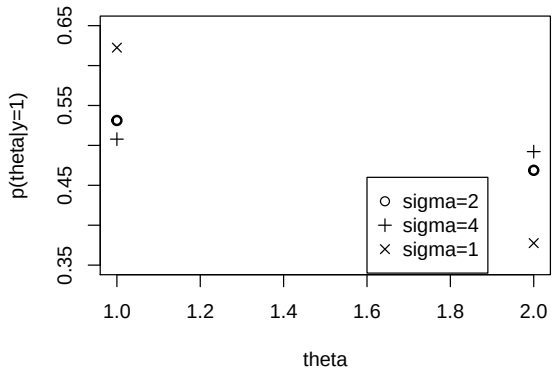
$$\begin{aligned} p(y) &= \sum_{\theta \in (1,2)} p(\theta, y) = \sum_{\theta \in (1,2)} p(y|\theta)p(\theta) \\ &= \frac{1}{2\sqrt{2\pi}} e^{-(y-1)^2/8} * 0.5 + \frac{1}{2\sqrt{2\pi}} e^{-(y-2)^2/8} * 0.5 \\ &= \frac{1}{4\sqrt{2\pi}} (e^{-(y-1)^2/8} + e^{-(y-2)^2/8}) \end{aligned}$$



(b) See work below.

$$\begin{aligned} P(\theta = 1|Y = 1) &= \frac{P(Y = 1|\theta = 1)P(\theta = 1)}{P(Y = 1)} \\ &= \frac{\frac{1}{2\sqrt{2\pi}} * 0.5}{\frac{1}{4\sqrt{2\pi}}(1 + e^{-1/8})} \\ &= \frac{1}{1 + e^{-1/8}} = 0.54 \end{aligned}$$

Note that the posterior distribution of theta is a probability distribution with all its mass at $\theta = 1$ and $\theta = 2$. When $\sigma = 2$, the distribution of θ given $y = 1$ looks like:



If we increase σ to 4, then $P(\theta = 1|Y = 1) = \frac{1}{(1+e^{-1/32})} = 0.51$ and $P(\theta = 2|Y = 1) = \frac{e^{-1/32}}{(1+e^{-1/32})} = 0.49$. If we decrease σ to 1, then $P(\theta = 1|Y = 1) = \frac{1}{(1+e^{-1/2})} = 0.62$ and $P(\theta = 2|Y = 1) = \frac{e^{-1/2}}{(1+e^{-1/2})} = 0.38$. This is shown on the above plot. What we see is that as σ decreases, the posterior density becomes more concentrated at $\theta = 1|Y = 1$ and less concentrated at $\theta = 2|Y = 1$. This means that the posterior probability of $\theta = 1$ is larger for smaller σ. The opposite is true as σ increases. For larger σ, the posterior probability is greater for $\theta = 2|Y = 1$ and less for $\theta = 1|Y = 1$. It looks like the $P(\theta = 1|Y = 1)$ and $P(\theta = 2|Y = 1)$ is approaching 0.5 as σ increases.

12. The probability that Elvis is an identical twin, given that he has a twin brother, is $\frac{5}{11}$.

$$\begin{aligned} P(identicaltwins|twinbrother) &= \frac{P(identicaltwins, twinbrother)}{P(twinbrother)} \\ &= \frac{P(twinbrother|identicaltwins)P(identicaltwins)}{P(twinbrother)} \\ &= \frac{\frac{1}{2}\frac{1}{300}}{\frac{1}{2}\frac{1}{300} + \frac{1}{4}\frac{1}{225}} = \frac{5}{11} \end{aligned}$$

13. (a) Let E=event that a 6 is rolled. Let I_A=knowledge of the outcome of the roll and I_E=no knowledge. Then $P(E|I_E)$ would be 1/6 because person E has no knowledge, so her best guess is to assume equal probability for all the possible outcomes of the roll. $P(E|I_A)$, however, would be either a 0 or a 1 depending on whether a 6 was rolled or not. So, by

the definition of subjective probability given here, the probability of rolling a 6 would be considered subjective. Persons A and B are both rational, and they have assigned unequal probabilities to event E, conditional upon the knowledge they each have of the event.

(b) Let event E be the event that Brazil wins the next World Cup. If A is ignorant of soccer knowledge, she would probably give all the teams equal probability of winning the World Cup, so that $P(E|I_A) = \frac{1}{no.teamsplayingWorldCup}$. If B is knowledgeable about soccer, B knows that Brazil usually has a very good soccer team and has won the World Cup *five* times. Person B might weight each teams probability of winning the World Cup by their number of previous wins. In this case, person B would give Brazil a much higher probability of winning than a country less well known for soccer such as South Africa. Again, persons A and person B are both rational, and they are both assigning reasonable probabilities to even E based on their current knowledge. By the definition of subjective given here, the probability of Brazil winning the World Cup would be considered subjective.

14. (b) We showed in class that if we have a binomial pmf and a Beta(1,1) prior, then the posterior distribution is a Beta(y+1, n-y+1). In problem 9, I showed the steps I would go through to prove that the posterior distribution for θ wth α and β unknown is $Beta(y+\alpha, n-y+\beta)$. This means that our posterior mean of θ is $\frac{y+\alpha}{n+\alpha+\beta}$. Now we must show that this posterior mean will always be between the prior mean, $\frac{\alpha}{\alpha+\beta}$ and the observed proportion of heads, $\frac{y}{n}$. This proof is shown on the attached handwritten sheet.

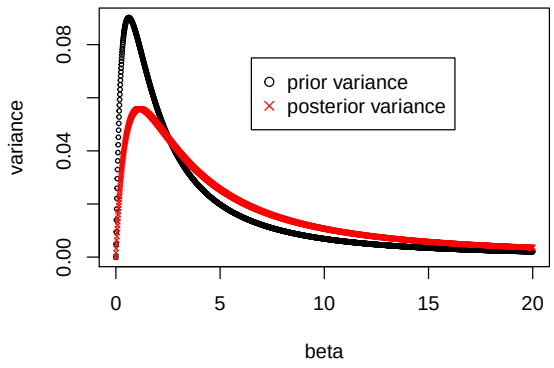
(c) Suppose the prior on θ is Uniform(0,1) (equivalent to a Beta(1,1)). Then the posterior distribution is $Beta(y+1, n-y+1)$, as shown in class, and the posterior variance is $\frac{(y+1)(n-y+1)}{(n+2)^2(3+n)}$. The prior variance is $\frac{1}{12}$. To show that the posterior variance is less than

1/12,

$$\frac{(y+1)(n-y+1)}{(n+2)^2(3+n)} \leq \frac{(n+1)}{(n+2)^2(3+n)} < \frac{(n+3)}{(n+2)^2(3+n)} = \frac{1}{(n+2)^2}$$

which is less than 1/12 for $n \geq 2$. For $n = 1$, $\frac{(n+1)}{(n+2)^2(3+n)} = \frac{1}{18}$, and $\frac{(y+1)(n-y+1)}{(n+2)^2(3+n)} \leq \frac{1}{18} < \frac{1}{12}$. The posterior variance is less than the prior variance of 1/12 for all values of y and n.

(d) Let's try a $Beta(1, \beta)$ prior, with $y = 1$ observations out of $n = 1$ trials. The prior variance is then $\frac{\beta}{(\beta+1)^2(\beta+2)}$. The posterior distribution is $Beta(2, \beta)$, and the posterior variance is then $\frac{2\beta}{(\beta+2)^2(\beta+3)}$. A plot showing the prior and posterior variance for different values of β is shown below. It looks like the posterior variance is larger for values of β greater than 3. When $\beta = 3$, the posterior variance is 0.04 and the prior variance is 0.0375.



15. (a) The details of the transformation are shown below. The prior on $log(\theta)$ does not appear to be non-informative. An exponential "prior" (I put prior in quotes because it is an improper prior) gives larger values of $log(\theta)$ higher probability. Small and negative values

of $log(\theta)$ are given lower probability.

$$
\begin{aligned}
p_\theta(\theta) &= 1I(\theta)_{(0,\infty)} \\
Y = log(\theta) \rightarrow e^y = \theta \rightarrow \frac{d\theta}{dy} &= e^y \\
f_Y(y) = p_\theta(e^y) * e^y &= 1I(e^y)_{(0,\text{inf})} e^y \\
f_Y(y) &= e^y I(y)_{(-\infty,\infty)}
\end{aligned}
$$

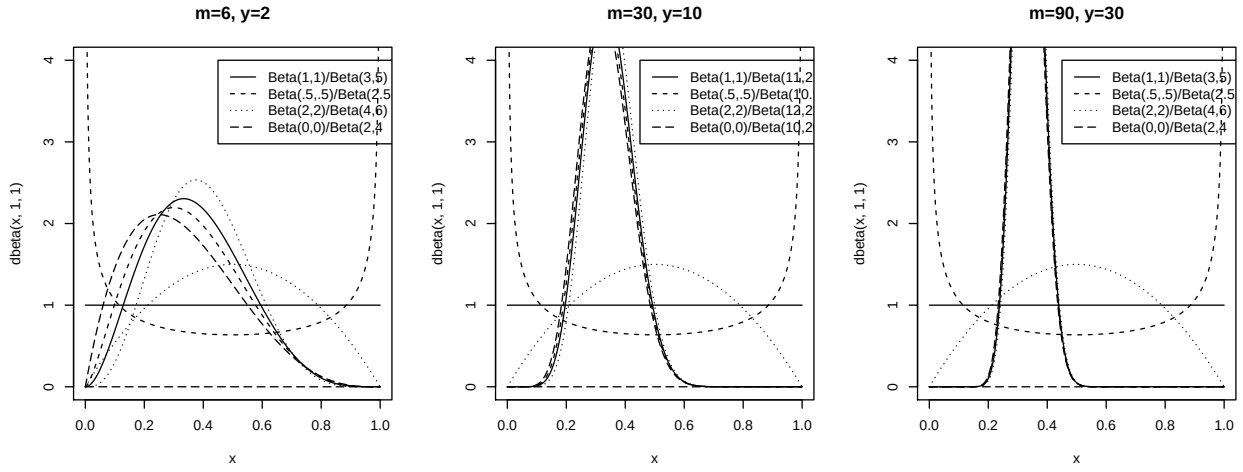
(b) To get Jeffrey's non-informative prior density, we find the Fisher Information:

$$
\begin{aligned}
-E[\frac{d^2}{d\theta^2} logp(y|\theta)] &= -E[\frac{d^2}{d\theta^2} log(\frac{\theta^y e^{-\theta}}{y!})] \\
&= -E[\frac{d^2}{d\theta^2} ylog\theta - \theta - log(y!)] \\
&= -E[-\frac{y}{\theta^2}] = \frac{1}{\theta} I(\theta)_{(0,\infty)}
\end{aligned}
$$

This is the kernel of a $Gam(1/2, 0)$ distribution, but it is an improper prior because it does not have the correct constant to ensure that the density sums to 1 over the support.

16. I am using the previously discussed result that for a binomial pmf with y success in m trials and a $Beta(\alpha, \beta)$ prior, the posterior is $Beta(y + \alpha, m - y + \beta)$.

Prior	Posterior m=6, y=2	Posterior m=30, y=10	Posterior m=90, y=30
Beta(1, 1)	Beta(3, 5)	Beta(11,21)	Beta(31,61)
Beta(0.5, 0.5)	Beta(2.5, 4.5)	Beta(10.5,20.5)	Beta(30.5, 60.5)
Beta(2, 2)	Beta(4, 6)	Beta(12, 22)	Beta(32, 62)
Beta(0, 0)	Beta(2, 4)	Beta(10, 20)	Beta(30, 60)



The plots show the four prior/posterior pairs for each sample size. I think the thing that stands out most from the plots is that the posterior is a lot more sensitive to the prior for smaller m and y. As far comparing sensitivity across *priors*, it's kind of hard to tell. None of the posterior distributions look too much like the corresponding prior. The $Beta(2, 2)$ is closest in shape to the corresponding posterior distribution.

17. I've always thought that a statement or decision is subjective if someone's personal beliefs are involved. I think there is a gray area, however, when personal beliefs are combined with more objective statements. Consider the example I gave in question 1 about Serena Williams. If I say that there is a 90% chance that Serena will win the US Open, this reflects both the objective fact that she has won 13 Grand Slam titles, and it reflects my personal belief that she is an incredible tennis player. It seems like there is a similar gray area in choosing prior distributions. The choice of prior may be based on some objective facts, but also incorporates personal beliefs. I think of subjectivity on a spectrum with some statements or decisions

incorporating more subjectivity than others. Practically, if I'm trying to decide whether something is more subjective or objective, I think of whether two different reasonable people could arrive at different conclusions and both justify their thought processes.

Steps of Statistical Inference:

(a) Define a research question!! (subjective, way question is defined can be different for different people)

(b) Define and think about your explanatory and response variables (objective, all of my 217 students have the same answer for this one and if they don't I mark it wrong)

(c) Look at and plot the raw data (subjective. many different reasonable ways to plot data)

(d) Decide what to do with missing data, or decide if any observations should not be included (subjective. there could be many different reasonable ideas about what to do with the observation that was taken when the bird had a wet beak)

(e) think about potential statistical procedures that could be appropriate (subjective, many different procedures could be justified)

(f) check assumptions (subjective, some may have differing ideas about when a violation is severe enough to interfere with the statistical procedures)

(g) select a final model, graph, or other to use for inference (subjective, same reason as (e))

(h) make a conclusion to address the question of interest (subjective wording. use a p-value cutoff or a strength of evidence statement?)

(i) provide estimates for effects of interest (I would call the actual reporting of the estimate objective)

(j) describe the scope of inference (subjective - there are general rules about when the design buys inference to the population or causality, but often, even when we don't have a random sample or random assignment we can justify inference to a larger population (or not) based off other knowledge)

R Code Appendix

```
require(ggplot2)
x <- seq(0,1, by=0.01)
y <- c(dbeta(x,0.5,0.5), dbeta(x,1,1), dbeta(x,2,2), dbeta(x,5,5))
beta <- c(rep("(0.5, 0.5)", 101), rep("(1, 1)", 101), rep("(2, 2)", 101), rep("(5, 5)", 101))
z <- cbind.data.frame(x,y, beta)
ggplot(aes(x, y, group=beta), data=z) +
  geom_line(aes(color=beta, size=beta))+
  xlab("")+
  ylab("")+
  ggtitle("Beta")
```

```
require(ggplot2)
x <- seq(0,5, by=0.05)
y <- c(dgamma(x,0.5,0.5), dgamma(x,1,1), dgamma(x,2,2), dgamma(x,5,5))
gamma <- c(rep("(2, 2)", 101), rep("(1, 1)", 101), rep("(1/2, 1/2)", 101), rep("(1/5, 1/5)", 101))
z <- cbind.data.frame(x,y, gamma)
ggplot(aes(x, y, group=gamma), data=z) +
  geom_line(aes(color=gamma, size=gamma))+
  xlab("")+
  ylab("")+
  ggtitle("Gamma")
```

```
require(ggplot2)
x <- seq(0,0.2, by=0.001)
y <- c(dnorm(x,0.05,0.01), dnorm(x, 0.10, 0.01))
normal <- c(rep("N(0.05, 0.01)", 201), rep("N(0.10, 0.01)", 201))
z <- cbind.data.frame(x,y, normal)
ggplot(aes(x, y, group=normal), data=z) +
  geom_line(aes(color=normal, size=normal))+
  xlab("Infection Rate")+
  ylab("")+
  ggtitle("Posterior Distribution of Infection Rates")
```

```
require(ggplot2)
x <- seq(0.48,0.50, by=0.0001)
y <- dbeta(x, 241947, 251529)
z <- cbind.data.frame(x,y)
ggplot(aes(x, y), data=z) +
  geom_line()+
  xlab("theta")+
  ylab("")
#pbeta(0.5, 241947, 251529)
```

```
require(ggplot2)
set.seed(99)
x <- rpois(20, 5)
```

```r
#write a likelihood function
poislikelihood <- function(lambda) {exp(-20*lambda)*lambda^sum(x)/prod(factorial(x))}
lambda <- seq(1,10, by=0.01)

#find mle and endpoints of confidence interval
require(stats)
mle <- optimize(poislikelihood, c(2.5, 7.5), maximum=TRUE)$maximum
values <- optimize(poislikelihood, c(2.5, 7.5), maximum=TRUE)$objective*0.05
require(mosaic)
ci.l <- findZeros(poislikelihood(lambda)-values ~ lambda, near=0)[1,]
ci.u <- findZeros(poislikelihood(lambda)-values ~ lambda, near=0)[2,]

#organize into a dataframe and plot
parta <- cbind.data.frame(lambda, likelihood=poislikelihood(lambda))
ggplot(aes(lambda, likelihood), data=parta)+
  geom_line()+
  geom_vline(xintercept=mle, color="red", linetype="dashed")+
  geom_vline(xintercept=5, color="blue", linetype="dotdash")+
  geom_vline(xintercept=ci.l, color="purple")+
  geom_vline(xintercept=ci.u, color="purple")
```

```r
require(ggplot2)
set.seed(99)
x <- rpois(100, 5)

#write a likelihood function
poisloglikelihood <- function(lambda) {((-100*lambda)+sum(x)*log(lambda))-(sum(log(factorial(x))))}
poislikelihood100 <- function(lambda) {exp(poisloglikelihood(lambda))}
lambda <- seq(4,6, by=0.01)

#find mle and endpoints of confidence interval
require(stats)
mle <- optimize(poislikelihood100, c(2.5, 7.5), maximum=TRUE)$maximum
values <- optimize(poislikelihood100, c(2.5, 7.5), maximum=TRUE)$objective*0.05
require(mosaic)
ci.l <- findZeros(poislikelihood100(lambda)-values ~ lambda, near=0)[1,]
ci.u <- findZeros(poislikelihood100(lambda)-values ~ lambda, near=0)[2,]

#organize into a dataframe and plot
partb <- cbind.data.frame(lambda, likelihood=poislikelihood100(lambda))
ggplot(aes(lambda, likelihood), data=partb)+
  geom_line()+
  geom_vline(xintercept=mle, color="red", linetype="dashed")+
  geom_vline(xintercept=5, color="blue", linetype="dotdash")+
  geom_vline(xintercept=ci.l, color="purple")+
  geom_vline(xintercept=ci.u, color="purple")
```

```r
require(ggplot2)
set.seed(55)
x <- rnorm(15, 10, sqrt(5))
```

```
#write a likelihood function
normloglikelihood <- function(vec) {
  -15/2*log(2*pi)-15/2*log(vec[2])-1/(2*vec[2])*sum((x-vec[1])^2)
}
normlikelihood <- function(vec) {exp(normloglikelihood(vec))}
vec <- c(5,10)

#find mle and endpoints of confidence interval
require(stats)
mles <- optim(c(5,3), normloglikelihood, control = list(fnscale = -1))$par

values <- exp(optim(c(5,3), normloglikelihood, control = list(fnscale = -1))$value)*0.05

normloglikelihood2 <- function(mu, sigmasq) {
  -15/2*log(2*pi)-15/2*log(sigmasq)-1/(2*sigmasq)*sum((x-mu)^2)
}
normlikelihood2 <- function(mu, sigmasq) {exp(normloglikelihood2(mu, sigmasq))}

require(mosaic)
muci.l <- findZeros(normlikelihood2(mu, sigmasq)-values ~ mu, near=0, sigmasq=3.299682)[1,]
muci.u <- findZeros(normlikelihood2(mu, sigmasq)-values ~ mu, near=0, sigmasq=3.299682)[2,]

sigmasqci.l <- findZeros(normlikelihood2(mu, sigmasq)-values ~ sigmasq, near=0, mu=9.1744)[1,]
sigmasqci.u <- findZeros(normlikelihood2(mu, sigmasq)-values ~ sigmasq, near=0, mu=9.1744)[2,]


#contour plot
ndim <- 50
mu.vals <- mles[1] + seq(-5,5, length=ndim)
sig.vals <- seq(0, (mles[2] + 12), length=ndim)
grid.vals <- expand.grid(mu.vals,sig.vals)
loglik.surf <- matrix(apply(grid.vals,1,normlikelihood), nrow=ndim, ncol=ndim)

contour(mu.vals, sig.vals, loglik.surf, nlevels = 200, xlab = expression(mu),
        ylab = expression(sigma), main="Likelihood Surface n=15", xlim=c(7.2,12),
        ylim=c(1,13))
  points(mles[1], mles[2], pch=17, col="red", cex=2)
  points(10, 3, pch=15, col="blue", cex=2)
  points(muci.l, sigmasqci.l, pch=16, col="green", cex=2)
  points(muci.l, sigmasqci.u, pch=16, col="green", cex=2)
  points(muci.u, sigmasqci.l, pch=16, col="green", cex=2)
  points(muci.u, sigmasqci.u, pch=16, col="green", cex=2)
  legend(11,12, c("CI endpoints", "mle", "truth"), pch=c(16, 17, 15))
```

```
require(ggplot2)
set.seed(55)
x <- rnorm(5, 10, sqrt(5))

#write a likelihood function
normloglikelihood <- function(vec) {
  -15/2*log(2*pi)-15/2*log(vec[2])-1/(2*vec[2])*sum((x-vec[1])^2)
}
```

```r
normlikelihood <- function(vec) {exp(normloglikelihood(vec))}
vec <- c(5,10)

#find mle and endpoints of confidence interval
require(stats)
mles <- optim(c(5,3), normloglikelihood, control = list(fnscale = -1))$par

values <- exp(optim(c(5,3), normloglikelihood, control = list(fnscale = -1))$value)*0.05

normloglikelihood2 <- function(mu, sigmasq) {
  -15/2*log(2*pi)-15/2*log(sigmasq)-1/(2*sigmasq)*sum((x-mu)^2)
}
normlikelihood2 <- function(mu, sigmasq) {exp(normloglikelihood2(mu, sigmasq))}


require(mosaic)
muci.l <- findZeros(normlikelihood2(mu, sigmasq)-values ~ mu, near=0, sigmasq=1.054140)[1,]
muci.u <- findZeros(normlikelihood2(mu, sigmasq)-values ~ mu, near=0, sigmasq=1.054140)[2,]

sigmasqci.l <- findZeros(normlikelihood2(mu, sigmasq)-values ~ sigmasq, near=0, mu=8.811188)[1,]
sigmasqci.u <- findZeros(normlikelihood2(mu, sigmasq)-values ~ sigmasq, near=0, mu=8.811188)[2,]


#contour plot
ndim <- 50
mu.vals <- mles[1] + seq(-5,5, length=ndim)
sig.vals <- seq(0, (mles[2] + 12), length=ndim)
grid.vals <- expand.grid(mu.vals,sig.vals)
loglik.surf <- matrix(apply(grid.vals,1,normlikelihood), nrow=ndim, ncol=ndim)

contour(mu.vals, sig.vals, loglik.surf, nlevels = 200, xlab = expression(mu),
        ylab = expression(sigma), main="Likelihood Surface n=5", xlim=c(7.2,11),
        ylim=c(0,6))
  points(mles[1], mles[2], pch=17, col="red", cex=2)
  points(10, 3, pch=15, col="blue", cex=2)
  points(muci.l, sigmasqci.l, pch=16, col="green", cex=2)
  points(muci.l, sigmasqci.u, pch=16, col="green", cex=2)
  points(muci.u, sigmasqci.l, pch=16, col="green", cex=2)
  points(muci.u, sigmasqci.u, pch=16, col="green", cex=2)
  legend(10,6, c("CI endpoints", "mle", "truth"), pch=c(16, 17, 15))
```

```r
require(ggplot2)
set.seed(55)
x <- rbinom(1, 100, 0.2)

#write a likelihood function
binomlikelihood <- function(p) {
  dbinom(x, 100, p)
}

#find mle and endpoints of confidence interval
require(stats)
```

```
mle <- optimize(binomlikelihood, c(.1, .3), maximum=TRUE)$maximum

values <- optimize(binomlikelihood, c(.1, .3), maximum=TRUE)$objective*0.05

require(mosaic)
pci.l <- findZeros(binomlikelihood(p)-values ~ p, near=0)[1,]
pci.u <- findZeros(binomlikelihood(p)-values ~ p, near=0)[2,]

#organize into a dataframe and plot
p <- seq(0, 0.4, by=0.001)
partb <- cbind.data.frame(p, likelihood=binomlikelihood(p))
ggplot(aes(p, likelihood), data=partb)+
  geom_line()+
  geom_vline(xintercept=mle, color="red", linetype="dashed")+
  geom_vline(xintercept=.2, color="blue", linetype="dotdash")+
  geom_vline(xintercept=pci.l, color="purple")+
  geom_vline(xintercept=pci.u, color="purple")
```

```
require(ggplot2)
set.seed(55)
x <- rbinom(30, 100, 0.2)

#write a likelihood function
binomloglikelihood <- function(p) {
  sum(log(choose(100,x)))+sum(x)*log(p)+sum(100-x)*log(1-p)
}
binomlikelihood <- function(p){exp(binomloglikelihood(p))}

#find mle and endpoints of confidence interval
require(stats)
mle <- optimize(binomlikelihood, c(.1, .3), maximum=TRUE)$maximum

values <- optimize(binomlikelihood, c(.1, .3), maximum=TRUE)$objective*0.05

require(mosaic)
pci.l <- findZeros(binomlikelihood(p)-values ~ p, near=0)[1,]
pci.u <- findZeros(binomlikelihood(p)-values ~ p, near=0)[2,]

#organize into a dataframe and plot
p <- seq(0.1, 0.3, by=0.001)
partb <- cbind.data.frame(p, likelihood=binomlikelihood(p))
ggplot(aes(p, likelihood), data=partb)+
  geom_line()+
  geom_vline(xintercept=mle, color="red", linetype="dashed")+
  geom_vline(xintercept=.2, color="blue", linetype="dotdash")+
  geom_vline(xintercept=pci.l, color="purple")+
  geom_vline(xintercept=pci.u, color="purple")
```

```
require(ggplot2)
x <- c(-5, 8)
plot.fun <- function(x){
```

```
  1/(4*sqrt(2*pi))*(exp(-(x-1)^2/8)+exp(-(x-2)^2/8))
}

y <- rep(0,2)
qplot(x,y)+geom_point(color="white")+
  stat_function(fun = plot.fun, colour = "red")+
  xlab("y")+
  ylab("p(y)")
```

```
y <- c(1/(1+exp(-1/8)), exp(-1/8)/(1+exp(-1/8)))
z <- c(1/((1+exp(-1/32))), exp(-1/32)/((1+exp(-1/32))))
f <- c(1/((1+exp(-1/2))), exp(-1/2)/((1+exp(-1/2))))
theta <- c(1,2)
plot(theta, y, lwd=2, ylim=c(0.35,0.65), ylab="p(theta|y=1)")
points(theta, z, pch=3)
points(theta, f, pch=4)
legend(1.6,0.46, c("sigma=2", "sigma=4", "sigma=1"), pch=c(1,3,4))
```

```
beta <- seq(0,20, by=0.01)
priorvar <- beta/((beta+1)^2*(beta+2))
postvar <- 2*beta/((beta+2)^2*(beta+3))
plot(beta, priorvar, ylab="variance", cex=0.5)
points(beta, postvar, col="red", pch=4, cex=0.5)
legend(6.5,0.075, c("prior variance", "posterior variance"), pch=c(1,4), col=c("black", "red"))
```

```
par(mfrow=c(1,3))
x <- seq(0,1,by=0.001)
plot(x, dbeta(x, 1,1), ylim=c(0,4), type="l", main="m=6, y=2")
lines(x, dbeta(x, 3,5), lty=1)
lines(x, dbeta(x, .5,.5), lty=2)
lines(x, dbeta(x, 2.5, 4.5), lty=2)
lines(x, dbeta(x, 2,2), lty=3)
lines(x, dbeta(x, 4, 6), lty=3)
lines(x, dbeta(x, 0,0), lty=5)
lines(x, dbeta(x, 2, 4), lty=5)
legend(0.45,4, c("Beta(1,1)/Beta(3,5) Prior/Post", "Beta(.5,.5)/Beta(2.5,4.5)", "Beta(2,2)/Beta(4,6)", "Beta(0

x <- seq(0,1,by=0.001)
plot(x, dbeta(x, 1,1), ylim=c(0,4), type="l", main="m=30, y=10")
lines(x, dbeta(x, 11,21), lty=1)
lines(x, dbeta(x, .5,.5), lty=2)
lines(x, dbeta(x, 10.5, 20.5), lty=2)
lines(x, dbeta(x, 2,2), lty=3)
lines(x, dbeta(x, 12, 22), lty=3)
lines(x, dbeta(x, 0,0), lty=5)
lines(x, dbeta(x, 10, 20), lty=5)
legend(0.45,4, c("Beta(1,1)/Beta(11,21) Prior/Post", "Beta(.5,.5)/Beta(10.5,20.5)", "Beta(2,2)/Beta(12,22)", "

x <- seq(0,1,by=0.001)
```

```
plot(x, dbeta(x, 1,1), ylim=c(0,4), type="l", main="m=90, y=30")
lines(x, dbeta(x, 31,61), lty=1)
lines(x, dbeta(x, .5,.5), lty=2)
lines(x, dbeta(x, 30.5, 60.5), lty=2)
lines(x, dbeta(x, 2,2), lty=3)
lines(x, dbeta(x, 32, 62), lty=3)
lines(x, dbeta(x, 0,0), lty=5)
lines(x, dbeta(x, 30, 60), lty=5)
legend(0.45,4, c("Beta(1,1)/Beta(3,5) Prior/Post", "Beta(.5,.5)/Beta(2.5,4.5)", "Beta(2,2)/Beta(4,6)", "Beta(0
```