STA 360 Lab 3: The Beta-Binomial model

STA 360: Bayesian Inference and Modern Statistical Methods

12 February, 2021

Preliminaries

Please turn in a PDF of this Rmd file on Sakai by Friday, February 12th at 11:59 PM. Exercises 2 and 4 will be graded for completion.

Repeated Binomial Trials

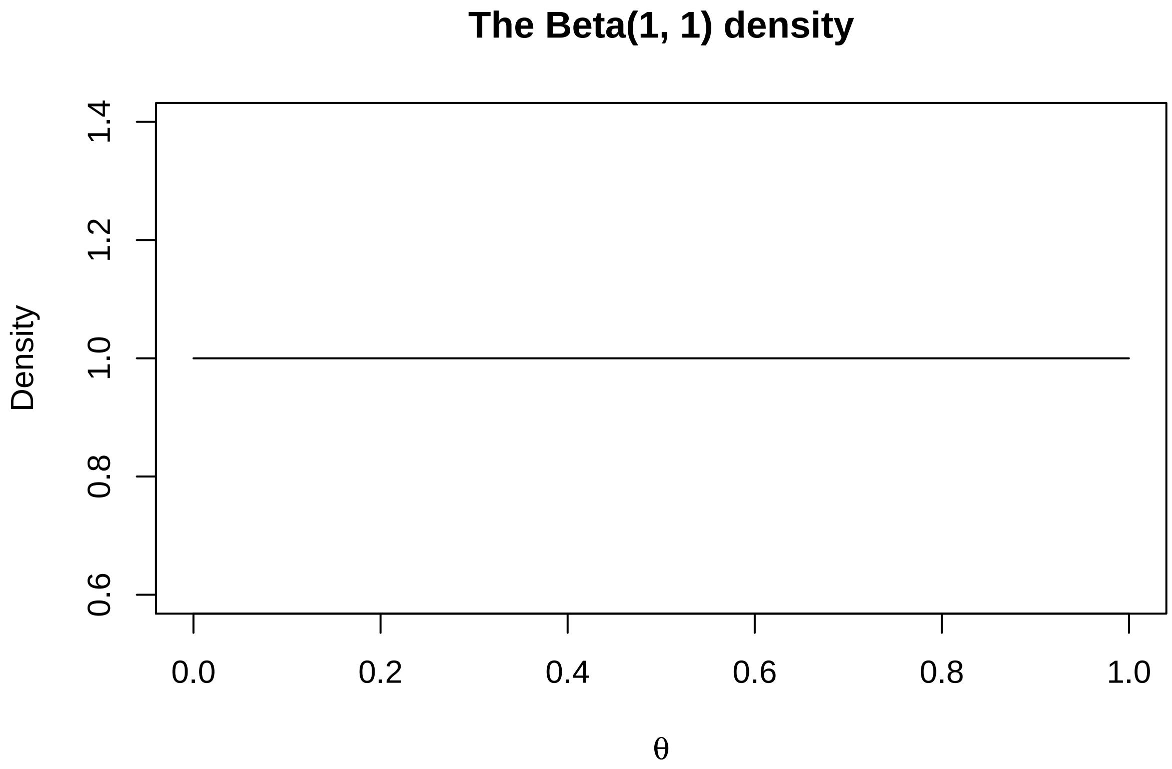
By this point, you are familiar with binomial data: If $Y \sim Binom(n, \theta)$, we assume the data are such that over n trials with success probability θ, we observe y successes. Let us consider multiple binomial realizations. We have data on rat tumor development from Tarone (1982). Specifically, we have the number of incidences of endometrial stromal polyps in 71 different groups of female lab rats of type F344. We begin by loading in the data:

```
tumors <- read.csv(file = url("http://www.stat.columbia.edu/~gelman/book/data/rats.asc"),
                     skip = 2, header = T, sep = " ")[,c(1,2)]
y <- tumors$y # number of successes
N <- tumors$N # binomial trials
n <- length(y) # sample size
```

Each row represents a group, or a draw from a binomial distribution. The y variable denotes the number of succcesses and the N variable denotes the total number of rats in that control group. For example the first group consists of 20 rats, with 0 of these 20 having developed a tumor. n is the number of groups.

If we assume that the probability of developing a tumor is the same across groups, then for each of the $i = 1, 2, \ldots, n$ groups, we have $y_i \sim Binom(N_i, \theta)$. We have learned that the Beta distribution is conjugate for Binomial data. For now, we place a $Beta(1, 1)$ prior on θ, which corresponds to a uniform density on the interval $[0, 1]$.

```
plot(seq(0, 1, length.out = 1000),
     dbeta(seq(0, 1, length.out = 1000), 1, 1),
     type = 'l',
     xlab = expression(theta), ylab = "Density",
     main = "The Beta(1, 1) density")
```



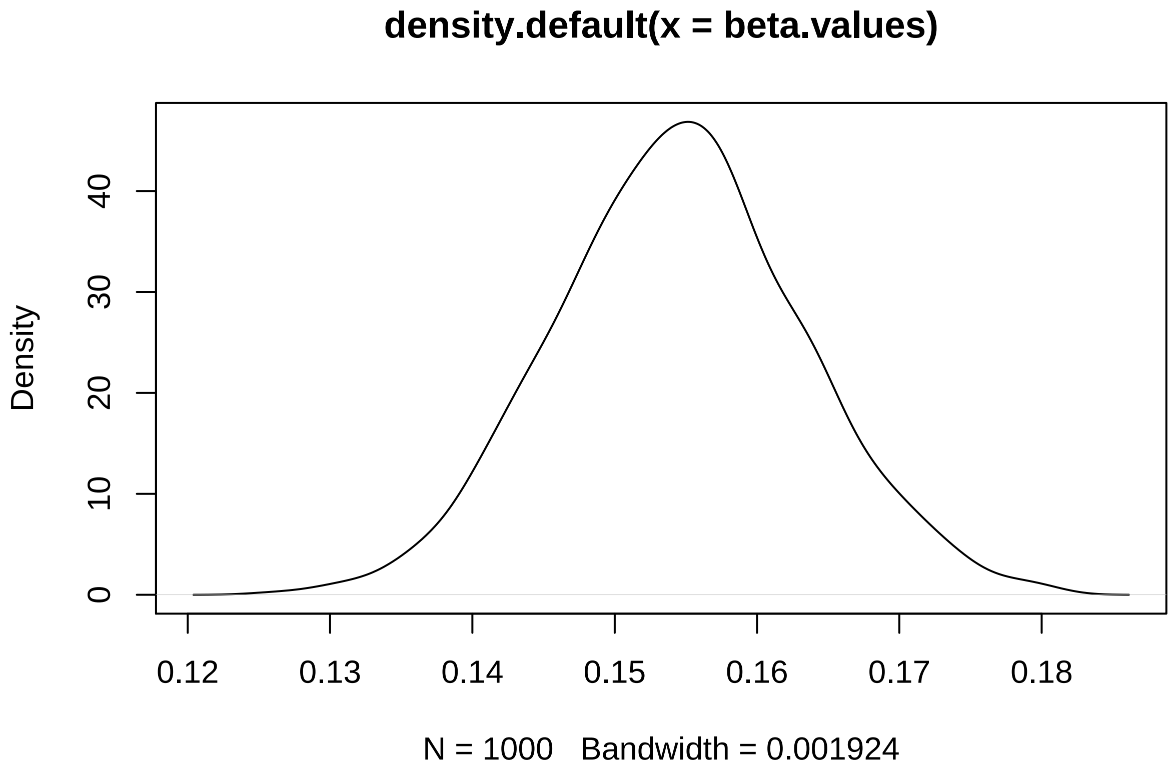
Now suppose we wanted to draw values from the posterior.

Exercise 0

Recall that if $Y \sim Binom(N, \theta)$ and $\theta \sim Beta(a, b)$, then $(\theta \mid Y) \sim Beta(a + y, b + n - y)$. Sample 1,000 observations from this posterior. Make a density plot of the observations.

```r
beta.values <- rbeta(n = 1000, shape1 = 1 + sum(y), shape2 = 1 + sum(N) - sum(y))

plot(density(beta.values), type = "l")
```



An alternative way to do this is to use `stan`. `stan` files consist of 3 parts:

- `data` that need to be input
- `parameters` that are to be estimated
- a `model` that describes the sampling model and the prior distributions

In the `Rmd` file, we supply the actual data and call file using the `stan()` function (example below). We can then extract outputs from the model that represent our posterior distribution(s) for further analysis. Let’s take a look with some examples:

```r
stan_dat <- list(n = n, N = N, y =y, a = 1, b = 1)
fit_pool <- stan('lab3_pool.stan', data = stan_dat, chains = 2, refresh = 0)
pool_output <- rstan::extract(fit_pool)
mean(pool_output$theta)
```

```
## [1] 0.1543364
```

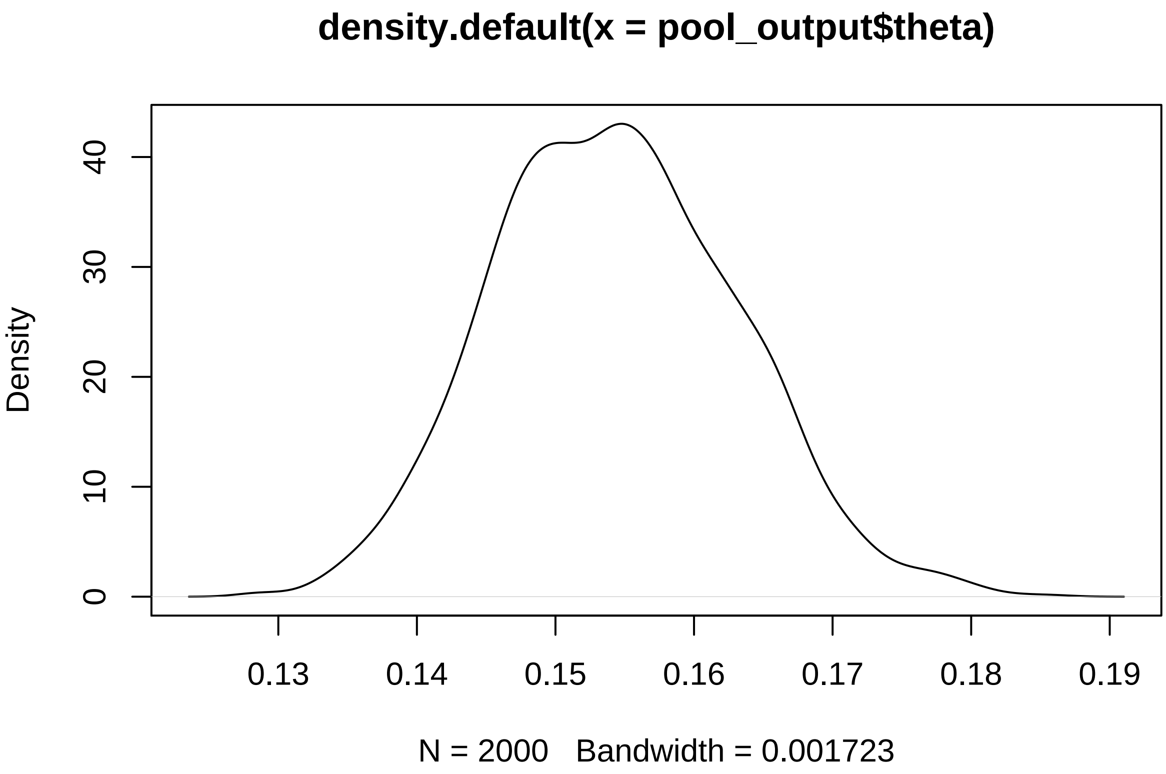
```r
mean(beta.values)
```

```
## [1] 0.1545218
```

Exercise 1

Plot a density of the prior distribution, and plot a density of θ from the `rstan` object called `pool_output` on the same graph. Since we know the posterior distribution (see, for example, Exercise 3 on HW 1), do these distributions seem reasonable? How do they compare to each other?

```
plot(density(pool_output$theta), type = 'l')
```



Alternatively, we may not have reason to believe that the probability of a rat developing a tumor should be the same across groups. Then we have the model $y_i \sim Binom(N_i, \theta_i)$ for $i = 1, 2, \ldots, n$. If we had expert knowledge about the different groups of rats, we might place different priors on each of the n θ_i's. However, for simplicity we choose to model the θ_i as i.i.d. $Beta(1, 1)$.

```
stan_dat <- list(n = n, N = N, y =y, a = 1, b = 1)
fit_nopool <- stan('lab3_nopool.stan', data = stan_dat, chains = 2, refresh = 0)
nopool_output <- rstan::extract(fit_nopool)
apply(nopool_output$theta,2,mean)
```

```
##  [1] 0.04559733 0.04660631 0.04673750 0.04533308 0.04545791 0.04550787
##  [7] 0.04526571 0.04845408 0.04806015 0.04822807 0.04781651 0.04912493
## [13] 0.04912514 0.05264819 0.08970851 0.09085872 0.09061531 0.09207577
## [19] 0.09621220 0.09657944 0.09896385 0.10133208 0.13705512 0.11027285
## [25] 0.11677478 0.12033362 0.13647024 0.13502213 0.13680292 0.13700561
## [31] 0.13715614 0.13713743 0.16639588 0.11929070 0.14427281 0.12550739
## [37] 0.15762323 0.15605798 0.16133538 0.18007730 0.18115180 0.19742842
## [43] 0.20054233 0.21139628 0.22904436 0.22474916 0.22664009 0.22948225
## [49] 0.22781343 0.22505287 0.22682769 0.22051966 0.23795275 0.23848790
## [55] 0.24037699 0.24895351 0.25142697 0.25475814 0.27496146 0.27279663
## [61] 0.27898398 0.28569229 0.29362414 0.31933265 0.31896225 0.31939555
## [67] 0.31390867 0.33054556 0.32680559 0.38421628 0.31237168
```

```
samples <- nopool_output[[1]]
dim(samples)
```

```
## [1] 2000   71
```

```
head(samples)
```

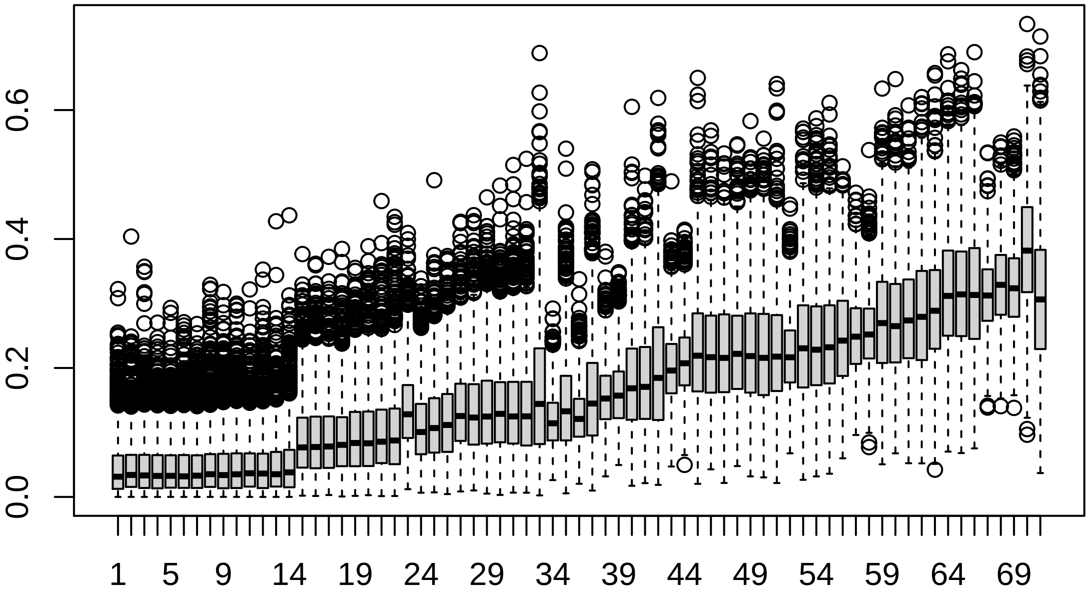
```
##
## iterations        [,1]       [,2]       [,3]       [,4]        [,5]       [,6]
##       [1,] 0.011158940 0.01892612 0.02421017 0.01035539 0.004755946 0.06854434
##       [2,] 0.048996798 0.02300957 0.01613694 0.01329809 0.053441740 0.04711003
##       [3,] 0.099810295 0.05778326 0.03263492 0.06201163 0.001755189 0.07766285
##       [4,] 0.009710874 0.02391560 0.07305299 0.05096916 0.023712241 0.01052782
##       [5,] 0.030995053 0.06783511 0.04977657 0.05484250 0.024177707 0.06537038
##       [6,] 0.030176556 0.07148629 0.02556720 0.06957754 0.045256284 0.11384425
##
## iterations         [,7]       [,8]        [,9]        [,10]      [,11]
##       [1,] 0.0148380310 0.02486241 0.032619291 0.0018825329 0.01915603
##       [2,] 0.0914888328 0.15854277 0.025296436 0.0003180447 0.03653842
##       [3,] 0.0095109290 0.11291892 0.034489239 0.0116247919 0.07600647
##       [4,] 0.0253191265 0.01814312 0.003535319 0.0341148199 0.05716536
##       [5,] 0.0009009329 0.04707350 0.047246828 0.0234132480 0.18840673
##       [6,] 0.0278338298 0.02689257 0.050503622 0.0080448748 0.05622780
##
## iterations      [,12]       [,13]      [,14]      [,15]      [,16]      [,17]
##       [1,] 0.01069719 0.005713984 0.06366839 0.13857025 0.07142231 0.06216607
##       [2,] 0.01496026 0.036204561 0.07680428 0.10339937 0.06006729 0.13102151
##       [3,] 0.04118438 0.018340178 0.13851919 0.08246072 0.07859123 0.05928887
##       [4,] 0.07374952 0.003775776 0.07491328 0.12387529 0.06978088 0.12581576
##       [5,] 0.05599183 0.007513349 0.08659478 0.02820316 0.09997584 0.01313482
##       [6,] 0.03718416 0.022787728 0.04209071 0.02131038 0.02967377 0.01803662
##
## iterations      [,18]      [,19]      [,20]      [,21]      [,22]      [,23]
##       [1,] 0.03128270 0.22109047 0.03632147 0.18408797 0.20560419 0.06424227
##       [2,] 0.02818412 0.03075150 0.09900180 0.06585701 0.03695191 0.12962679
##       [3,] 0.07727479 0.07990973 0.15728225 0.07232928 0.12894388 0.16582432
##       [4,] 0.03140952 0.11223402 0.04341504 0.04026992 0.07042168 0.16261205
##       [5,] 0.05718868 0.10477416 0.05654217 0.13585431 0.05305995 0.17959697
##       [6,] 0.01310822 0.13126050 0.07920422 0.16445166 0.04665120 0.12416028
##
## iterations      [,24]      [,25]      [,26]      [,27]      [,28]      [,29]
##       [1,] 0.08821438 0.23097311 0.02469202 0.23847592 0.07699124 0.15456161
##       [2,] 0.08977188 0.14446675 0.20586725 0.18041014 0.11443402 0.16774171
##       [3,] 0.09488880 0.08330983 0.21509480 0.07183051 0.20103316 0.02829577
##       [4,] 0.07366198 0.20602342 0.06376357 0.11152869 0.19425454 0.12513427
##       [5,] 0.15484587 0.29484448 0.06644171 0.08399352 0.09563659 0.13375009
##       [6,] 0.24061327 0.12097062 0.13050574 0.12875645 0.12678468 0.13554997
##
## iterations     [,30]      [,31]      [,32]      [,33]      [,34]     [,35]
##       [1,] 0.1191871 0.07367754 0.12411295 0.37846379 0.07778463 0.2041699
##       [2,] 0.2420157 0.19706588 0.10446315 0.10339847 0.09287382 0.2670115
##       [3,] 0.1487886 0.05965777 0.17237082 0.06037806 0.10235659 0.1600393
##       [4,] 0.1649426 0.36496535 0.15343373 0.10682398 0.07613307 0.1660878
##       [5,] 0.1667525 0.19156313 0.24822464 0.05914671 0.07701541 0.2200851
##       [6,] 0.1160058 0.33243828 0.09715994 0.17051774 0.03667823 0.1907366
##
## iterations      [,36]      [,37]     [,38]      [,39]     [,40]      [,41]
##       [1,] 0.14024417 0.05488039 0.1613948 0.08549236 0.2155844 0.10485941
##       [2,] 0.21827457 0.20031982 0.2250113 0.16379863 0.1438020 0.05666776
##       [3,] 0.13838233 0.03978048 0.1085668 0.10817034 0.2498188 0.16173665
##       [4,] 0.09527648 0.27295018 0.1617028 0.11870061 0.1752736 0.13426450
##       [5,] 0.16711770 0.10811660 0.2166199 0.13516597 0.1571485 0.10739161
##       [6,] 0.13836933 0.26599391 0.1031543 0.10979633 0.2908770 0.16640698
##
## iterations      [,42]     [,43]     [,44]     [,45]     [,46]      [,47]
##       [1,] 0.35329074 0.2258400 0.2471876 0.2517813 0.1307310 0.30604284
##       [2,] 0.28185004 0.2524044 0.2686088 0.1991106 0.1523768 0.13533005
##       [3,] 0.30115957 0.1757689 0.3089883 0.3819626 0.2576025 0.12778992
##       [4,] 0.09088177 0.3322115 0.1374561 0.1994119 0.2136749 0.19627975
##       [5,] 0.32055095 0.2302104 0.1590615 0.1328216 0.3403530 0.19118059
##       [6,] 0.07529766 0.2613730 0.2919094 0.1558526 0.1102827 0.08372028
##
## iterations      [,48]     [,49]     [,50]     [,51]     [,52]      [,53]
##       [1,] 0.08282846 0.2709810 0.2560916 0.1675690 0.2101210 0.17716168
##       [2,] 0.40648229 0.1274611 0.4272847 0.1843034 0.2667005 0.08319503
##       [3,] 0.31664833 0.3023644 0.1228365 0.4036240 0.2454858 0.34243679
##       [4,] 0.18945505 0.2318058 0.1669375 0.1586808 0.1366918 0.15359136
##       [5,] 0.15495612 0.3229150 0.1710484 0.2574455 0.1890728 0.22785222
##       [6,] 0.45924172 0.1491732 0.1690950 0.2499994 0.2278267 0.21147973
##
## iterations     [,54]     [,55]     [,56]     [,57]     [,58]     [,59]
##       [1,] 0.3927432 0.1839623 0.2146384 0.2361586 0.2067198 0.2152110
##       [2,] 0.1804415 0.2999603 0.3376993 0.2060214 0.2202515 0.2188381
##       [3,] 0.2588068 0.2845174 0.1638930 0.1171935 0.2592863 0.2402267
##       [4,] 0.1944204 0.2934312 0.2576884 0.2857053 0.3226638 0.2926132
##       [5,] 0.1739235 0.2740395 0.1809210 0.3313998 0.1657787 0.1782450
##       [6,] 0.1594663 0.2946342 0.3395171 0.2395693 0.2581460 0.2221947
##
## iterations     [,60]     [,61]     [,62]     [,63]     [,64]     [,65]
##       [1,] 0.2396771 0.3781553 0.2597736 0.2505873 0.3300317 0.3873389
##       [2,] 0.1757614 0.2918782 0.2154699 0.3134780 0.2240534 0.2921039
##       [3,] 0.2620416 0.2656038 0.3867211 0.3664373 0.4259723 0.1394835
##       [4,] 0.2269559 0.2513918 0.3000461 0.2546757 0.3251789 0.2103585
##       [5,] 0.2193776 0.3055292 0.3056128 0.4085649 0.4898659 0.3477149
##       [6,] 0.2809192 0.2210642 0.5571987 0.3967850 0.2955531 0.4014651
##
## iterations     [,66]     [,67]     [,68]     [,69]     [,70]     [,71]
##       [1,] 0.4049913 0.3037468 0.4087239 0.4468145 0.3819561 0.2461760
##       [2,] 0.2020348 0.3053393 0.3931928 0.2692189 0.5372708 0.1909767
##       [3,] 0.3961408 0.3273122 0.3255622 0.3235053 0.2777061 0.2533299
##       [4,] 0.4487204 0.3399551 0.3080260 0.3762501 0.4596049 0.3646242
##       [5,] 0.2752795 0.3508006 0.2386204 0.3844777 0.2509161 0.2534361
##       [6,] 0.2489006 0.3431155 0.3484342 0.4702904 0.4834402 0.2055262
```

Exercise 2

Visualize the posterior distributions of the θ_i with boxplots. In the plot, there should be one box and whiskers object for each θ_i.

What is actually being plotted here (i.e., you can describe this with a mathematical expression and/or in words)? What does each point represent?

```
boxplot(nopool_output$theta)
```



Each boxplot is the distribution of thetas in each set of trials. ***

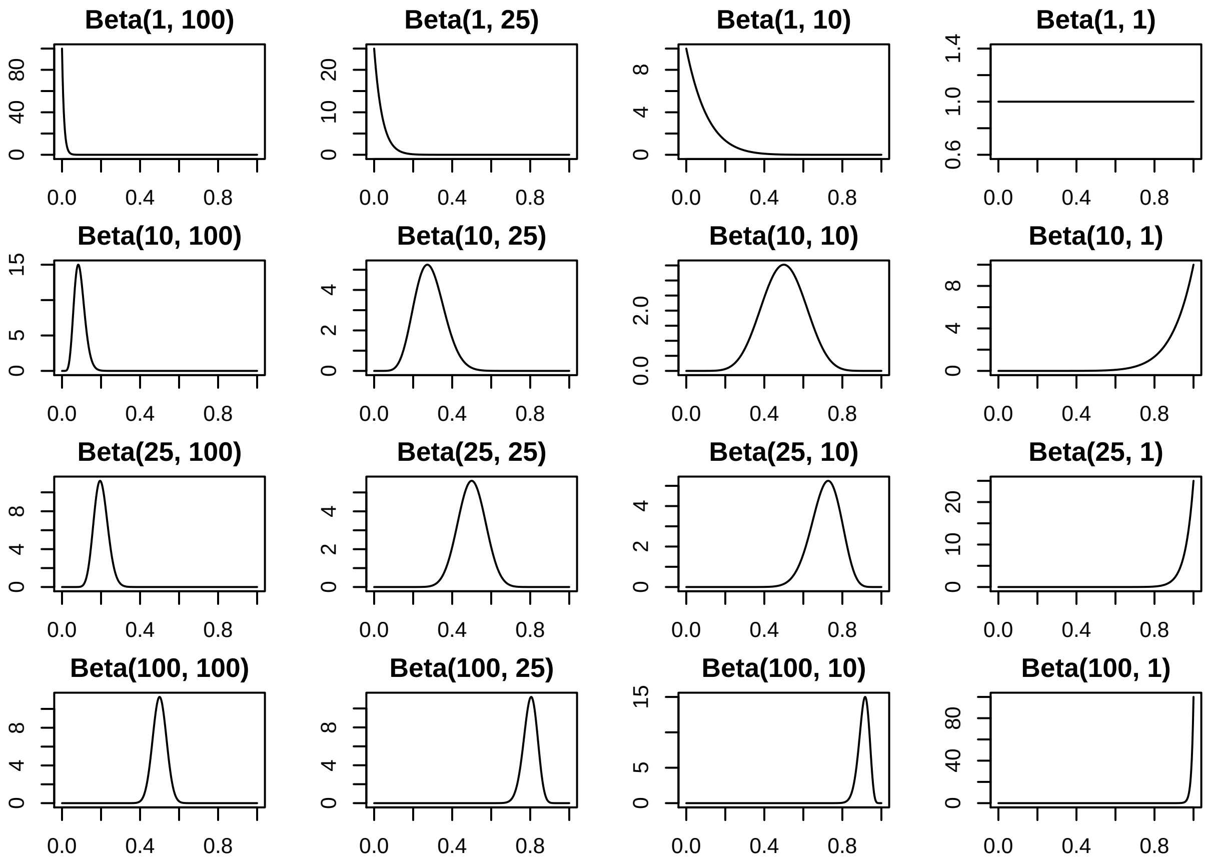
Sensitivity analysis

With the Beta-Binomial model, we know that the posterior is $\theta|Y \sim Beta(a + \sum y_i, b + \sum N_i - \sum y_i)$. Therefore, the posterior mean is

$$E[\theta|Y] = \frac{a + \sum y_i}{a + b + \sum N_i}$$

We fit the above models with $a = 1, b = 1$, but it is good practice to perform an analysis to determine how sensitive the posterior is to the choice of prior. Considering the first model where we assumed the same success probability θ across groups, let us sample from the posterior distribution of θ over a range of a and b values. These parameter settings produce very different pictures of our prior beliefs about θ:

```r
par(mfrow = c(4, 4))
par(mar=c(2,2,2,2))
for(a_val in c(1, 10, 25, 100)){
  for(b_val in rev(c(1, 10, 25, 100))){
    plot(seq(0, 1, length.out = 1000),
     dbeta(seq(0, 1, length.out = 1000), a_val, b_val),
     type = ’l’,
     xlab = expression(theta), ylab = "Density",
     main = paste0("Beta(", a_val, ", ", b_val, ")"))
  }
}
```

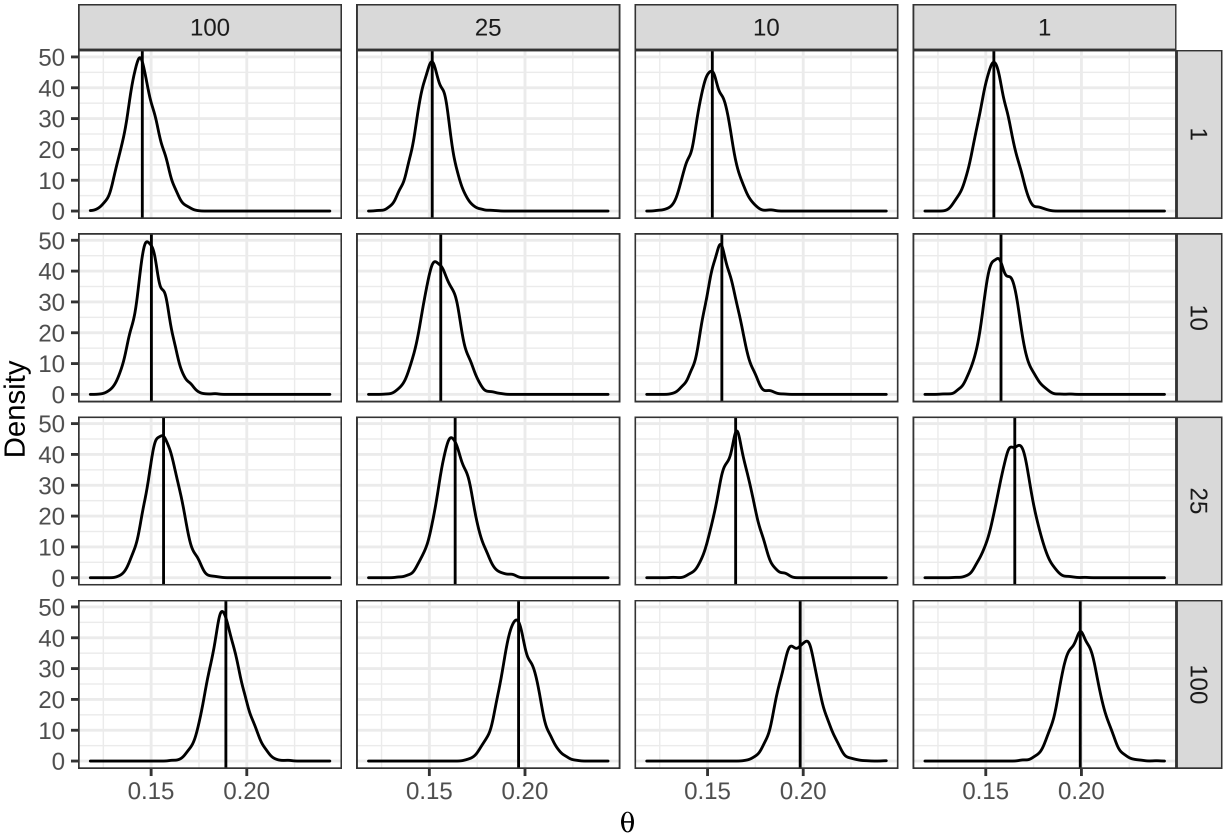


To get samples from the posterior distribution of θ for each one of the prior distributions above, we run:

```r
output_list <- list()
for(a_val in c(1, 10, 25, 100)){
  for(b_val in c(1, 10, 25, 100)){
    stan_dat <- list(n = n, N = N, y = y, a = a_val, b = b_val)
    fit_pool <- stan(’lab3_pool.stan’, data = stan_dat, chains = 2, refresh = 0)
    output_list[[paste0("a_", a_val, ":b_", b_val)]] <- rstan::extract(fit_pool)[["theta"]]
  }
}
```

We then compile the samples from the different prior specifications into a data.frame, which will help us visualize the results.

```r
output_list %>%
  plyr::ldply(function(theta){
    reshape2::melt(theta) %>%
      dplyr::mutate(post_mean = mean(theta))
  }, .id = "prior") %>%
  tidyr::separate("prior", into = c("a", "b"), sep = ":") %>%
  dplyr::mutate(a = as.numeric(gsub("._", "", a)),
                b = as.numeric(gsub("._", "", b))) %>%
  ggplot2::ggplot() +
  geom_density(aes(x = value)) +
  geom_vline(aes(xintercept = post_mean)) +
  facet_grid(a~factor(b, levels = rev(c(1, 10, 25, 100)))) +
  scale_colour_brewer(palette = "Set1") +
  labs(x = expression(theta), y = "Density")
```



In the plot above, increasing values of the parameter a are displayed moving from top to bottom along the vertical direction. Decreasing values of the parameter b are displayed moving from left to right along the horizontal direction. We can see that all of the posterior distributions look roughly normal with roughly equal variance. They are all fairly concentrated on values of θ within the range $[0.1, 0.2]$.

Here is a further observation that is specific to the concept of sensitivity analysis: for fixed a, as b increases the posterior mean shifts slightly towards lower values of θ. But for fixed b, as a increases the posterior mean shifts more dramatically towards higher values of θ. We might say that the posterior mean of θ is more sensitive to our prior beliefs about a than it is to our prior beliefs about b. Why might this be the case?

We can look at the formula for the posterior mean above to find an explicit answer. What might be a more intuitive explanation for the posterior's high sensitivity to the parameter a?

Exercise 3

Here are some questions to consider:

1. What observable quantity does the parameter a represent about our prior beliefs with respect to these data? What does b represent?
2. What do we actually observe in the rat tumor data with respect to these quantities?
3. How well do our different prior beliefs – the ones represented by the different parameter settings above – match up with the data?

Exercise 4

Returning to the initial exploration where we considered a single θ versus allowing θ_i to vary across groups: You should have noticed that if we allow the groups to have different success probabilities, then our estimates $\hat{\theta}_i$ vary from 0.05 to 0.30. However when we assumed a single probability of success, we obtained $\hat{\theta} \approx 0.15$. In this first approach and assuming the 71 groups are independent, we essentially have one large binomial trial: $Y^* = \sum y_i \sim Binom(\sum N_i, \theta)$. Applying ideas from the sensitivity analysis. Why might we have observed such a difference between the two approaches when using the prior $Beta(1, 1)$ in both cases?

(a) Derive or state the maximum likelihood estimate (MLE) for the binomial model $y \mid \theta \sim \text{Binom}(n, \theta)$.

Approach 1:

$$\widehat{\theta}^{ML} = sum(yi)/sum(Ni)$$

Approach 2:

$$\widehat{\theta}_i^{ML} = yi/ni$$

(b) Compute the MLE estimates for both problems (i.e., when we have the same θ for each group and when we have a different θ for each group).

```
# approach 1
sum(y) / sum(N)
```

```
## [1] 0.1535365
```

```
# approach 2
y/N
```

```
##  [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [7] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## [13] 0.00000000 0.00000000 0.05000000 0.05000000 0.05000000 0.05000000
## [19] 0.05263158 0.05263158 0.05555556 0.05555556 0.11111111 0.08000000
## [25] 0.08333333 0.08695652 0.10000000 0.10000000 0.10000000 0.10000000
## [31] 0.10000000 0.10000000 0.10000000 0.10204082 0.10526316 0.10869565
## [37] 0.11764706 0.14285714 0.14893617 0.15000000 0.15000000 0.15384615
## [43] 0.18750000 0.20000000 0.20000000 0.20000000 0.20000000 0.20000000
## [49] 0.20000000 0.20000000 0.20000000 0.20833333 0.21052632 0.21052632
## [55] 0.21052632 0.22727273 0.23913043 0.24489796 0.25000000 0.25000000
## [61] 0.26086957 0.26315789 0.27272727 0.30000000 0.30000000 0.30000000
## [67] 0.30769231 0.32608696 0.31914894 0.37500000 0.28571429
```

(c) State what the expectation is for the prior beta distribution $\theta \sim \text{Beta}(a, b)$:

$$\mathsf{E}[\theta] = a/(a + b)$$

(d) Recall that if $Y \sim Binom(N, \theta)$ and $\theta \sim Beta(a, b)$, then $(\theta \mid Y) \sim Beta(a + y, b + n - y)$. State what the expectation is for this posterior distribution, and show that you can write it as a weighted average of the MLE $\widehat{\theta}^{ML}$ and the prior expectation $\mathsf{E}[\theta]$.

$$\mathsf{E}[\theta \mid Y = y] = \frac{a+y}{b+a+n} = \frac{(a+b)}{(b+a+n)} * \frac{a}{(a+b)} + \frac{n}{b+a+n} * \frac{y}{n}$$

(e) How does this help explain the difference between our estimates of θ in our two approaches?

Looking at the weighted average, the more samples there are, the greater the weight given to the MLE is. Thus, our pooled estimate will place a greater weight on the observations and less weight on the prior. Our nonpooled theta estimates will give a greater weight to the prior.