Methods Homework 6

Tim Vigers

October 10, 2018

```
# Load the libraries.
library(ggplot2)
library(reshape2)
```

A) Load gvhd.txt into R, then subset the data to focus on only transplant recipients with an HLA-matched sibling donor.

```
# Read in and subset the data.
gvhd <- read.table("/Users/timvigers/Documents/School/UC Denver/Biostatistics/Biostatistical Methods 1/
# Subset to just HLA matched siblings.
hla.matched <- gvhd[gvhd$hla.matched.sibling == 1,c(1,3,4)]
```

B) Calculate the proportion of recipients that got GvHD in the Treatment A group. Repeat for Treatment B.

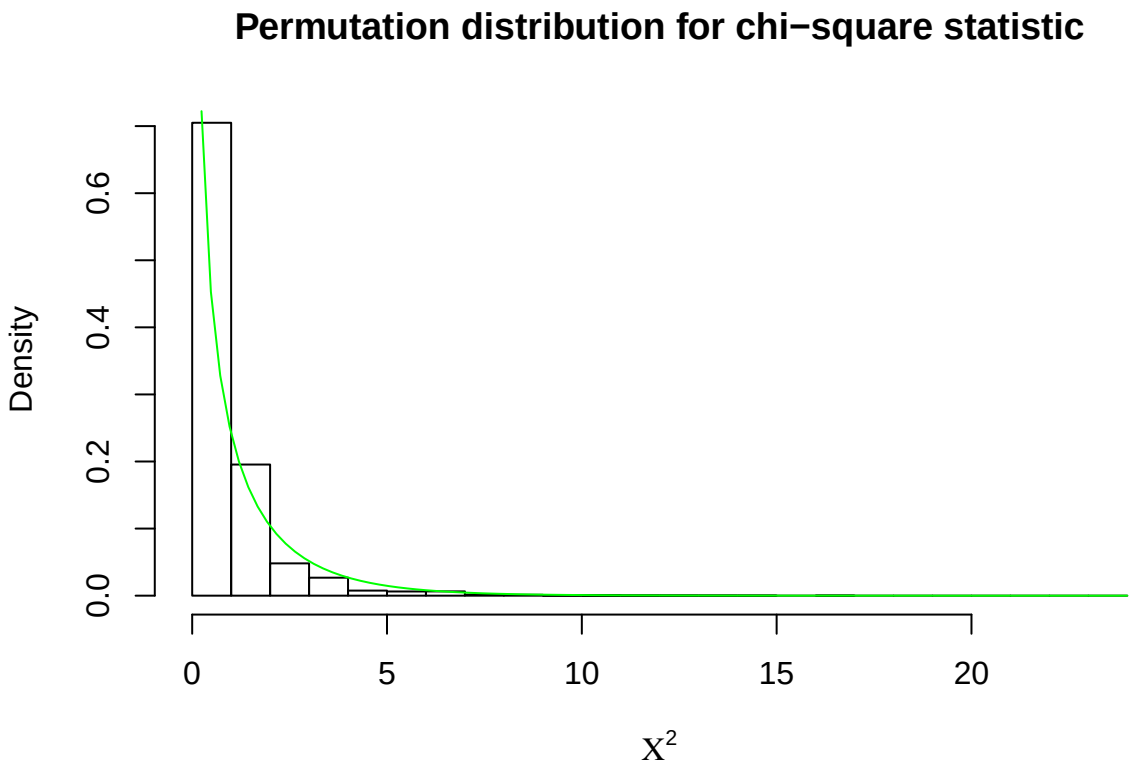
```
# Make a proportion table for both groups.
prop.table(table(hla.matched[,2],hla.matched[,3]),margin = 1)
```

```
##
##             0         1
##   A 0.7094017 0.2905983
##   B 0.7796610 0.2203390
```

C) Among transplant recipients with HLA-matched donors, is there a significant association between treatment and GvHD at the 5% level of significance? Carry this test out using both a permutation test, and either an exact or asymptotic method, as appropriate. Summarize your results and comment on differences, if any, between the two methods you applied.

```
# Find the test statistic for the original table.
observed <- chisq.test(table(hla.matched[,2],hla.matched[,3]))$statistic
# Make a vector to store permutation results.
B <- 10^6 - 1
result <- numeric(B)
# Repeat the chi-square test on B permutations.
for (i in 1:B) {
  permuted <- sample(hla.matched$treatment)
  table <- table(permuted,hla.matched$outcome)
  test <- chisq.test(table)
  result[i] <- test$statistic
}
# Plot.
```

```r
hist(result, freq=FALSE, xlab = expression(Chi^2), main="Permutation distribution for chi-square statis
curve(dchisq(x, 1), add=TRUE, col="green")
```




```r
# Compare the p value from the permutation distribution to the p-value from the
# chi-square distribution (the asymptotic result).
perm.p <- (sum(result >= observed)+1)/(B + 1)
true.p <- as.numeric(1 - pchisq(observed, df = 1))
perm.p
```

```
## [1] 0.233915
```

```r
true.p
```

```
## [1] 0.2777336
```

For this asymptotic test, I used a chi-square test with 1 degree of freedom (the question here is very similar to the in-class example about support for medicinal marijuana). The permutation test produced a p-value that is pretty close to the "true" p-value, and the histogram looks fairly similar to the plot of the chi-square distribution (althought maybe not quite as close as the in-class example). At the standard 5% level of significance, there does not appear to be a difference between treatment groups A and B, at least among HLA-matched siblings.

D) Using the seq() function, create a vector called p_grid that has 30 evenly spaced probabilities from 0 to 1.

```
# Create the vector.
p_grid <- seq(from=0,to=1,length.out = 30)
p_grid
```

```
##  [1] 0.00000000 0.03448276 0.06896552 0.10344828 0.13793103 0.17241379
##  [7] 0.20689655 0.24137931 0.27586207 0.31034483 0.34482759 0.37931034
## [13] 0.41379310 0.44827586 0.48275862 0.51724138 0.55172414 0.58620690
## [19] 0.62068966 0.65517241 0.68965517 0.72413793 0.75862069 0.79310345
## [25] 0.82758621 0.86206897 0.89655172 0.93103448 0.96551724 1.00000000
```

E) Assume that whether or not a patient has GvHD is a binary feature modeled by a Bernoulli distribution (see Lecture 4). Using the dbinom() function, find the likelihood of the number of GvHD cases among subjects in Treatment A at each value in p_grid. You should end up with a 30-element long vector of probabilities. Save this vector as "likelihood".

```
# Subset by treatment.
treat.a <- hla.matched[hla.matched$treatment == "A",]
treat.b <- hla.matched[hla.matched$treatment == "B",]
# Define the number of cases among treatment A, and the total number in
# treatment A.
x <- sum(treat.a$outcome)
n <- length(treat.a$outcome)
# Make the likelihood vector.
likelihood <- dbinom(x,n,prob = p_grid)
likelihood
```

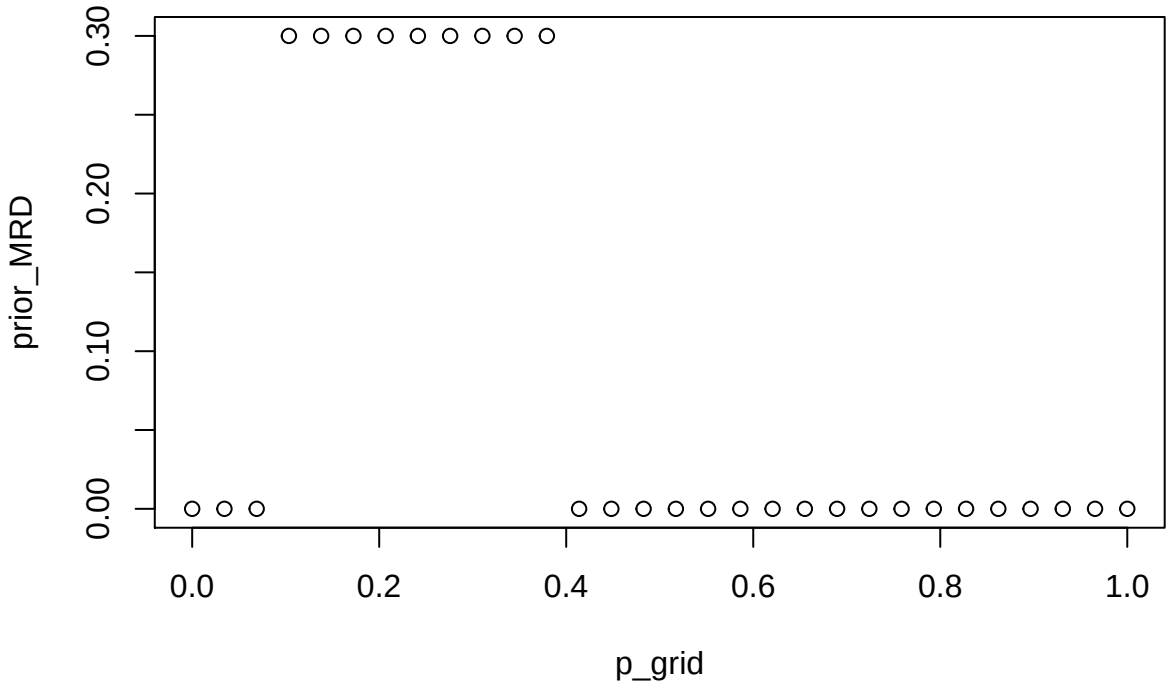
```
##  [1] 0.000000e+00 3.516132e-22 2.952303e-13 1.249854e-08 8.531178e-06
##  [6] 5.681484e-04 8.175258e-03 3.858020e-02 7.606933e-02 7.272497e-02
## [11] 3.702520e-02 1.063942e-02 1.783904e-03 1.770032e-04 1.037233e-05
## [16] 3.529217e-07 6.750349e-09 6.906457e-11 3.522140e-13 8.120280e-16
## [21] 7.397700e-19 2.207362e-22 1.649898e-26 2.076103e-31 2.363966e-37
## [26] 8.568172e-45 1.387096e-54 1.212848e-68 4.318388e-93 0.000000e+00
```

F) Use the following code to generate a possible prior distribution of the probability of GvHD for HLA-matched, related donors (which is based on the existing literature information). What prior distribution does this represent (e.g., normal, Poisson, uniform) and what parameters does this distribution have?

```
# Use the code.
prior_MRD <- ifelse(p_grid > 0.1 & p_grid < 0.4, 0.3, 0)
prior_MRD
```

```
##  [1] 0.0 0.0 0.0 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.0 0.0 0.0 0.0 0.0
## [18] 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
```

```
plot(p_grid,prior_MRD)
```



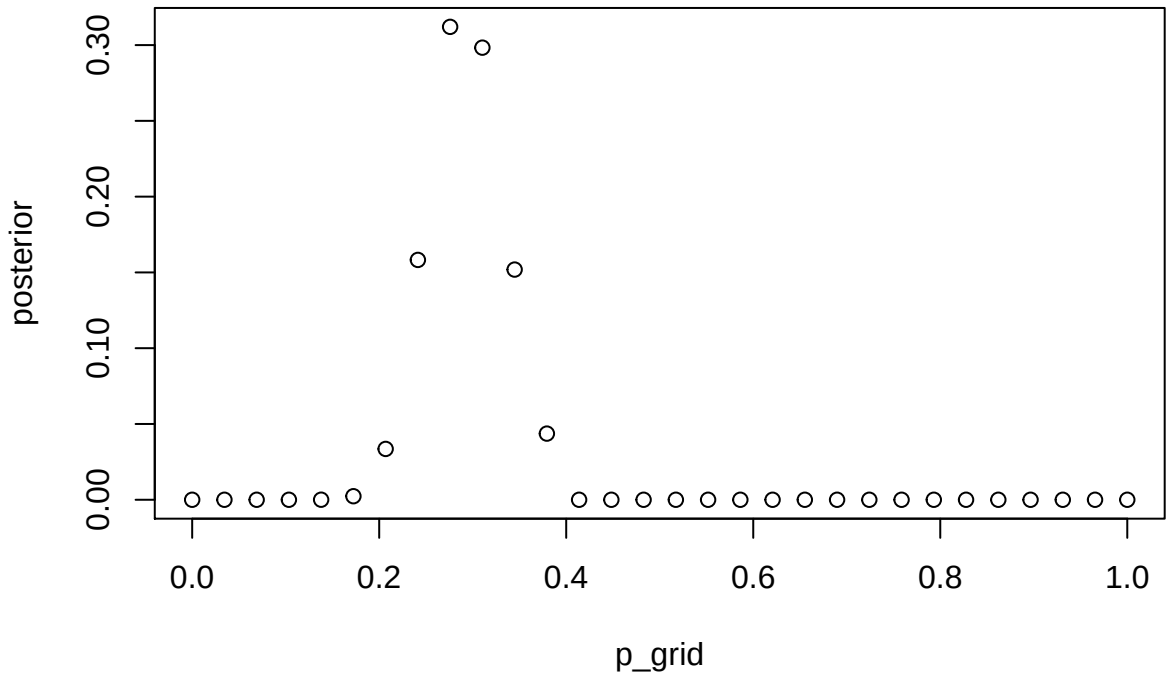
This prior distribution is a uniform(0,0.3) distribution.

G) Calculate the posterior distribution, using the following code:

```
# Use the code.
posterior <- likelihood * prior_MRD / sum(likelihood * prior_MRD)
posterior
```

```
##  [1] 0.000000e+00 0.000000e+00 0.000000e+00 5.126740e-08 3.499381e-05
##  [6] 2.330473e-03 3.353387e-02 1.582511e-01 3.120267e-01 2.983086e-01
## [11] 1.518727e-01 4.364153e-02 0.000000e+00 0.000000e+00 0.000000e+00
## [16] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## [21] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## [26] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
```

```
plot(p_grid,posterior)
```



H) Find the means of the prior distribution and the posterior distribution numerically. Hint: Recall the definition of expected value for discrete events.

```
prior.mean <- sum(p_grid * prior_MRD) / sum(prior_MRD)
prior.mean
```

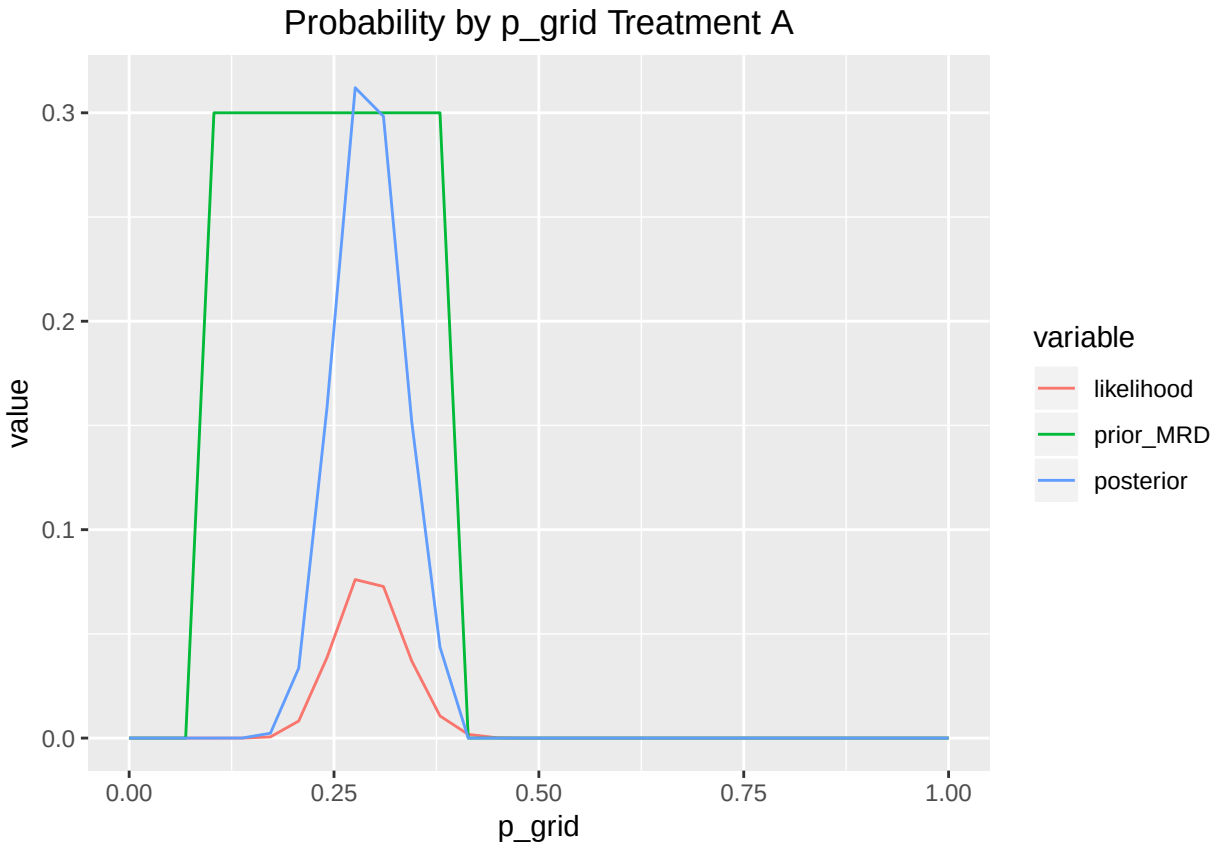
```
## [1] 0.2413793
```

```
post.mean <- sum(p_grid * posterior)
post.mean
```

```
## [1] 0.2931217
```

I) Plot the likelihood, prior, and posterior (as Y-variables) against p_grid (X-variable) for Treatment group A in the same figure. Make the line of each distribution a different color. Summarize what you observe.

```
# Make a data table for ggplot. Melt for easier plotting.
dat <- as.data.frame(cbind(p_grid,likelihood,prior_MRD,posterior))
dat <- melt(dat,id.vars = p_grid)
# Plot.
plot <- ggplot(data = dat, aes(x = p_grid,y = value,color = variable)) +
  geom_line(aes(group = variable)) +
  ggtitle("Probability by p_grid Treatment A") +
  theme(plot.title = element_text(hjust = 0.5))
plot
```



Based on this plot we can see that given just our prior distribution, we are equally certain that the probability of developing GvHD is in the range of about 0.1 - 0.4. We would say that that the most certain probability is about 0.24, but that there's an equal chance the probability is 0.15 or 0.35. After looking at the data and updating our distribution though, we are more certain that the true probability is 0.29, and we've shrunk the likely range of probabilities to something like 0.25 - 0.35.

J) EXTRA CREDIT: Repeat parts E through I for those who received Treatment B. Comment on how likely you think there is to be a difference between the two treatments.

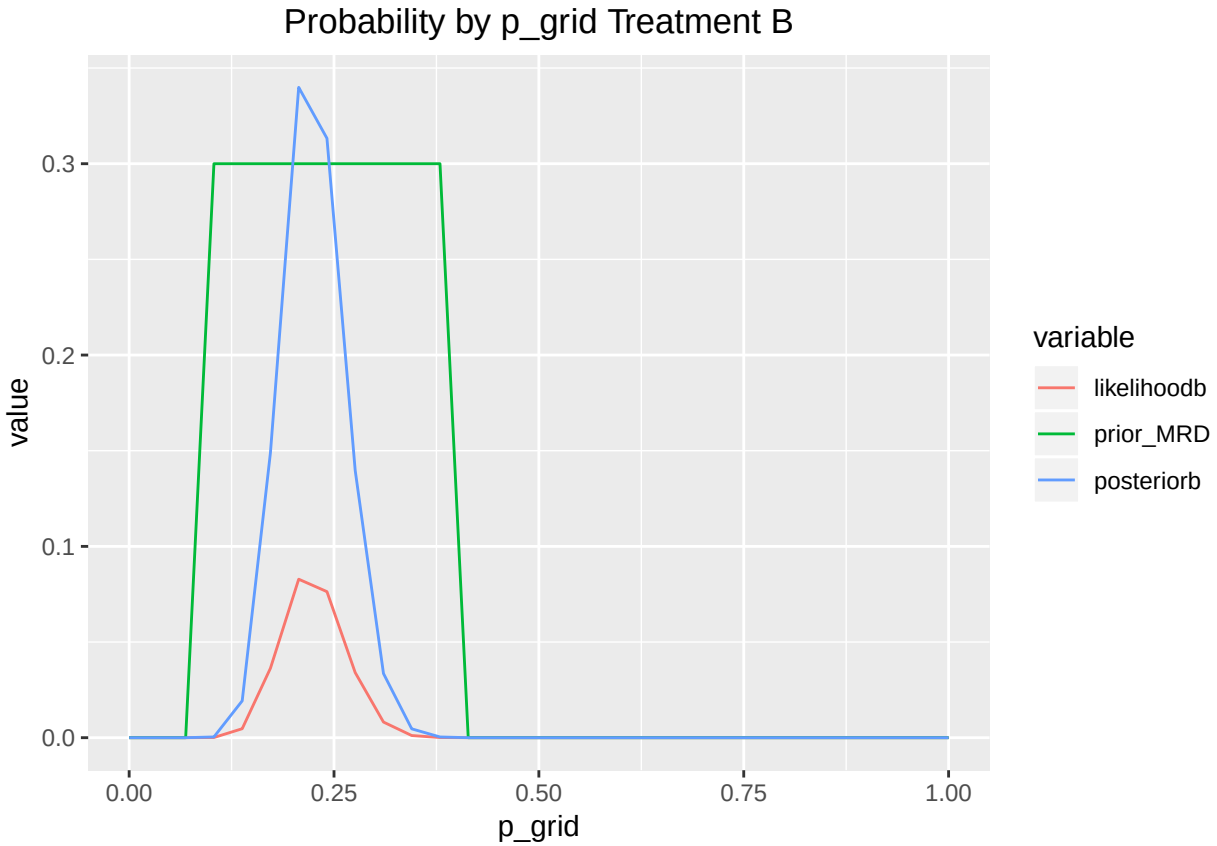
```
# Define the number of cases among treatment B, and the total number in
# treatment B.
xb <- sum(treat.b$outcome)
nb <- length(treat.b$outcome)
# Make the likelihood vector.
likelihoodb <- dbinom(xb,nb,prob = p_grid)
# Calculate the posterior distribution.
posteriorb <- likelihoodb * prior_MRD / sum(likelihoodb * prior_MRD)
# Calculate prior and posterior means.
post.mean.b <- sum(p_grid * posteriorb)
prior.mean
```

```
## [1] 0.2413793
```

```
post.mean.b
```

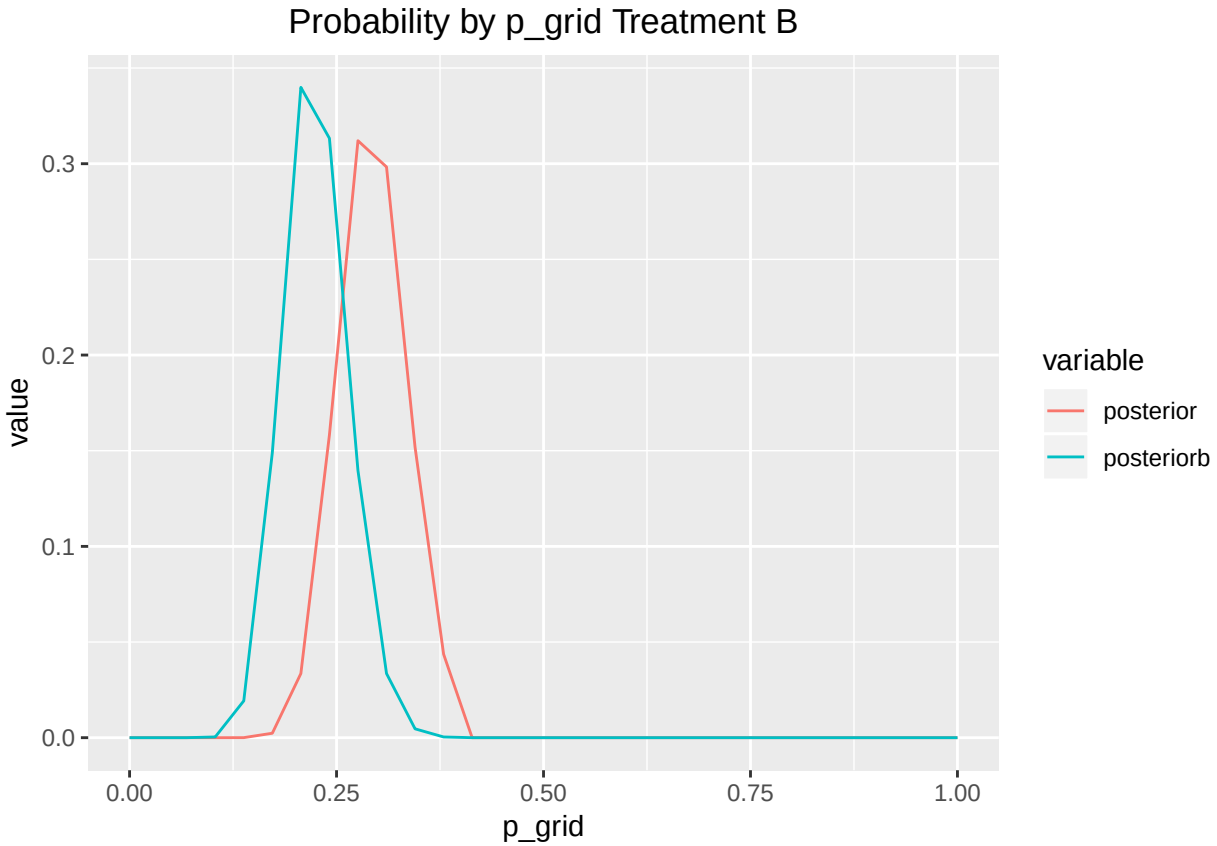
```
## [1] 0.2249963
```

```
# Make a data table for ggplot. Melt for easier plotting.
dat <- as.data.frame(cbind(p_grid,likelihoodb,prior_MRD,posteriorb))
dat <- melt(dat,id.vars = p_grid)
# Plot.
plot <- ggplot(data = dat, aes(x = p_grid,y = value,color = variable)) +
  geom_line(aes(group = variable)) +
  ggtitle("Probability by p_grid Treatment B") +
  theme(plot.title = element_text(hjust = 0.5))
plot
```



The posterior distribution for treatment B looks pretty similar to treatment A, in the sense that we can be a little more certain about the probability of developing the disease. However the most likely probability for treatment B is a little lower than treatment A (0.22 instead of 0.29). This suggests that treatment B probably does reduce the risk of developing the disease in this group, but it's useful to compare the posterior distributions side by side:

```
# Make a data table of both treaments for ggplot. Melt for easier plotting.
dat <- as.data.frame(cbind(p_grid,posterior,posteriorb))
dat <- melt(dat,id.vars = p_grid)
# Plot.
plot <- ggplot(data = dat, aes(x = p_grid,y = value,color = variable)) +
  geom_line(aes(group = variable)) +
  ggtitle("Probability by p_grid Treatment B") +
  theme(plot.title = element_text(hjust = 0.5))
plot
```



I would still say that there isn't a huge difference between the treatments, as the posterior distributions overlap quite a lot. However, this interpretation depends somewhat on what is clinically significant. This probably indicates that the difference in probability of disease is at the very least worth further investigation, since the tips of the distriubutions are pretty well separated.