BIOS 6611 Homework 6 *Answer Key*

Due Monday, October 15, 2018 by noon to Canvas Assignment Basket

Graft-versus-host disease (GvHD) is a secondary complication of allogeneic (from a donor) hematopoietic stem cell transplantation (the only potentially curative treatment for leukemia and many other disorders). It is characterized by the transplanted cells of the donor inappropriately attacking the tissues of the transplant recipient, or host. You're running a large cohort study of stem cell recipients in which a placebo is given to Treatment Group A and a new GvHD prophylaxis drug is given to Treatment Group B.

The risk factors for GvHD are extremely complex and still little understood. However, the characteristics of the donor are extremely important. HLA-matched siblings are considered the "gold standard" donor because they appear to be associated with substantially decreased GvHD risk (systematic reviews suggest 10-40% GvHD incidence in HLA-matched, related donors, versus 50 - 80% in other donors).

We will not perform a "complete" analysis of this data (which would include commenting on whether or not we observe a difference in GvHD risk by treatment), but focus on viewing the relationships between the prior and the posterior distributions of responses to the drugs.

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

```
gvhd <- gvhd[gvhd$hla.matched.sibling == 1, ]
> head(gvhd)
   subject.id hla.matched.sibling treatment outcome
2           2                   1         A       1
4           4                   1         A       0
7           7                   1         B       1
10         10                   1         B       1
11         11                   1         A       0
12         12                   1         A       0
```

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

```
summary <- table(gvhd$treatment, gvhd$outcome)

> summary["A", "1"] / sum(summary["A", ])
[1] 0.2905983

> summary["B", "1"] / sum(summary["B", ])
[1] 0.220339
```

29% in Treatment group A got GVHD and 22% in Treatment group B

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.

Commented [AK1]: 40 points

25 for permutation test

15 for asymptotic test

```
> chisq<-function(Obs) #uncorrected chi-square statistic
+ { #Obs is the observed contingency table
+   Expected <- outer(rowSums(Obs),colSums(Obs))/sum(Obs)
+   sum((Obs-Expected)^2/Expected)
+ }
> observed <- chisq(table(gvhd$treatment,gvhd$outcome))
> B <- 10^4-1  #set number of times to repeat this process
> result <- numeric(B) # space to save the random differences
> for(i in 1:B)
+ {
+   treat.permuted <- sample(gvhd$treatment)
+   perm.table <- table(treat.permuted, gvhd$outcome)
+   result[i] <- chisq(perm.table)
+ }

> # Compute p-value from the permutation distribution
> (sum(result >= observed)+1)/(B + 1)  #P-value
[1] 0.229
>
> # Compute p-value from chi-square distribution
> 1-pchisq(observed, df=1)
[1] 0.2168198
```

No, there is not a significant association between treatment groups and GvHD.

The p-values from the two methods are slightly different (although they give the same conclusion). Should eventually get the same number with a large number of permutations.

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

```
> 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 0.20689655
0.24137931 0.27586207 0.31034483 0.34482759
[12] 0.37931034 0.41379310 0.44827586 0.48275862 0.51724138 0.55172414 0.58620690
0.62068966 0.65517241 0.68965517 0.72413793
[23] 0.75862069 0.79310345 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`”.

```
> likelihood <- dbinom(x = sum(summary["A", "1"]),
+                      size = sum(summary["A", ]),
+                      prob = p_grid)
> likelihood
 [1] 0.000000e+00 3.516132e-22 2.952303e-13 1.249854e-08 8.531178e-06 5.681484e-04
8.175258e-03 3.858020e-02 7.606933e-02
[10] 7.272497e-02 3.702520e-02 1.063942e-02 1.783904e-03 1.770032e-04 1.037233e-05
3.529217e-07 6.750349e-09 6.906457e-11
[19] 3.522140e-13 8.120280e-16 7.397700e-19 2.207362e-22 1.649898e-26 2.076103e-31
2.363966e-37 8.568172e-45 1.387096e-54
[28] 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?

```
prior_MRD <- ifelse( p_grid > 0.1 & p_grid < 0.4, 0.3, 0)
```

Note: This is an “improper” prior, i.e. it does not integrate to 1. This is ok, but you will have to account for this when you obtain the mean proportion of GvHD based on this prior in part H below.

This is a discrete prior distribution and it has uniform probability mass of 0.3 at each of the (discrete) points in the range from 0.1 to 0.4. These points represent the individual Bernoulli (prior) probabilities (p) of developing GvHD.

G) Calculate the posterior distribution, using the following 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
2.330473e-03 3.353387e-02 1.582511e-01 3.120267e-01
[10] 2.983086e-01 1.518727e-01 4.364153e-02 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
[19] 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
[28] 0.000000e+00 0.000000e+00 0.000000e+00
```

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

Mean of Prior:

```
> sum(p_grid  * prior_MRD/sum(prior_MRD) )
[1] 0.2413793
```

Note, we had to divide by the sum of our prior because we assigned it a mass of 0.3 between probabilities of 0.1 and 0.4 and a mass of 0 everywhere else, and need to account for this decision so that the prior probabilities sum to 1.

Mean of Posterior (proper because of how it is calculated):

```
> sum(p_grid  * posterior)
[1] 0.2931217
```

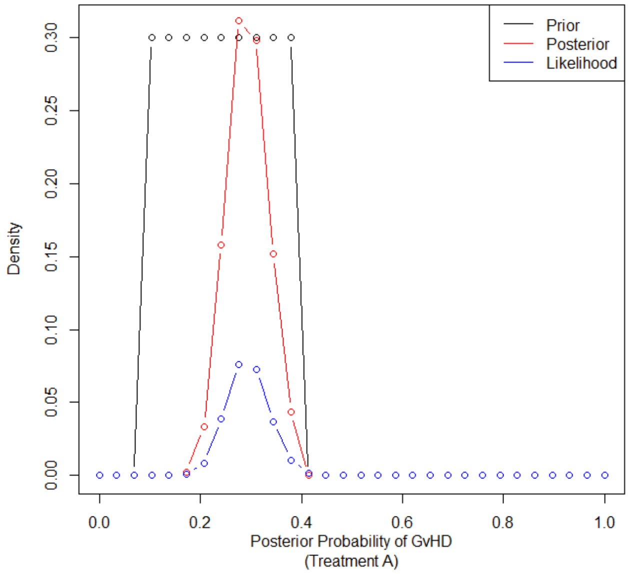
I) Plot the likelihood, prior, and posterior (as Y-variables) against `p_grid` (X-variable) for Treatment group A in the same figure. Make the line of each distribution a different color. Summarize what you observe.

Commented [AK2]: 60 points

10 points for each curve

30 points for summary which should touch on the general relationships between these three lines

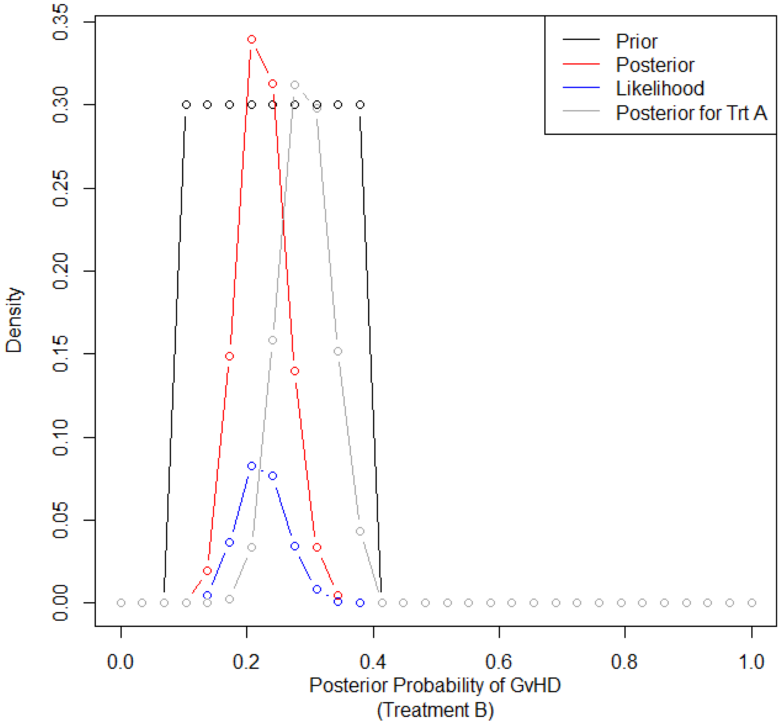
```
plot(p_grid, prior_MRD,
     xlab = "Posterior Probability of GvHD\n(Treatment A)",
     ylab = "Density", type = 'b', col = "black",ylim=c(0,0.31))
lines(p_grid , posterior, type = 'b', col = "red")
lines(p_grid , likelihood, type = 'b', col = "blue")
```



The black line represents the prior and had the shape of a uniform distribution, and is either 0 or 0.3, based on the p_grid we set. The blue line represents the binomial likelihood based off the parameters of treatment A. The red line represents the posterior and takes into account the binomial likelihood and our prior likelihood. When the prior has a value of 0, we see that the posterior also has values of 0. Likewise, when the likelihood is approximately 0, even if the prior is 0.3, we see that the posterior is 0. The posterior distribution has a more similar shape to our likelihood, but is much higher because it takes into account our prior value.

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.

Commented [AK3]: 2 extra credit points, note you don't need to overlay posterior for A on the figure, it was done here to help compare A and B



The posterior mean for group B is lower than it is for group A. There is substantial overlap in the posterior distributions indicating that it's unlikely the groups differ in terms of incidence of GvHD.

Part E

```
> likelihood <- dbinom(x = sum(summary["B", "1"]),
+                      size = sum(summary["B", ]),
+                      prob = p_grid)
```

```
>
```

Part G

```
> posterior <- likelihood * prior_MRD / sum(likelihood * prior_MRD)
```

Part H

Mean of Prior:

```
> sum(p_grid  * prior_MRD/sum(prior_MRD) )
[1] 0.2413793
```

Posterior mean for group B:

```
> sum(p_grid  * posterior)

[1] 0.2249963
```

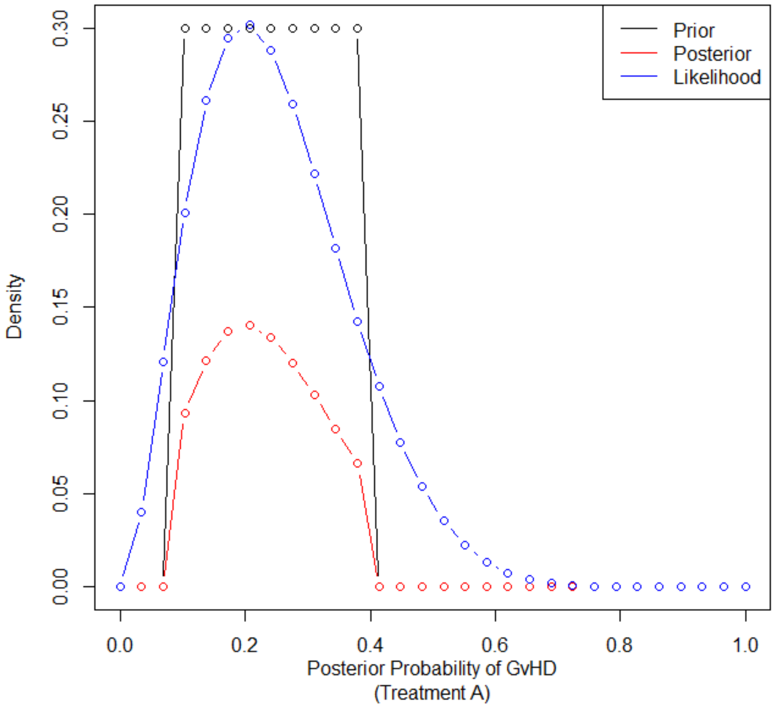
Part I

```
plot(p_grid, prior_MRD,
     xlab = "Posterior Probability of GvHD\n(Treatment B)",
     ylab = "Density", type = 'b', col = "black",ylim=c(0,0.34))
lines(p_grid , posteriorB, type = 'b', col = "red")
lines(p_grid , likelihoodB, type = 'b', col = "blue")
lines(p_grid , posterior, type = 'b', col = "gray65")
legend('topright', col=c('black','red','blue','gray65'),
      legend=c('Prior','Posterior','Likelihood','Posterior for Trt A'), lty=1 )
```

K) **EXTRA CREDIT:** To see what happens with smaller sample sizes, randomly (but reproducibly!) subsample your data so that there are only 5-10 subjects for Treatment A. Remake the plot in part I (i.e. the plot with the prior, likelihood and posterior for Treatment A) with this subsample. Obtain the posterior mean based on the smaller sample. What do you *qualitatively* notice about the posterior distribution and mean with the smaller sample compared with the posterior distribution and mean from the larger sample?

Commented [AK4]: 2 extra credit points

```
set.seed(1234) # For reproducibility!
gvhda <- gvhd[which(gvhd$treatment=='A'),]
gvhdsub <- gvhda[sample(1:nrow(gvhda), 10, replace=FALSE), ]
```



Prior Mean:

```
> sum(p_grid  * prior)/sum(prior)
[1] 0.2413793
```

Posterior Mean:

```
> sum(p_grid  * posterior_As)
[1] 0.2308007
```

With a smaller data set, there is more weight put into the prior and the range of probable parameter values remains relatively unchanged (as in the plot above). But, as the number of observations in the data set grows, the likelihood has more weight and in effect concentrates the prior distribution into a narrower range of probable parameter values as seen in the posterior distribution (see plots for larger A and B samples).