Final Project

Math 6480

Given the data provided, use Frequentist and Bayesian approaches to estimate the percentage of domestic and international students that see a therapist.

Introduction

We are interested in estimating the proportion of domestic and international students who have ever been in therapy. From the data gathered we will attempt to estimate separate percentages for domestic and international students and test whether they are different.

Due to the sensitivity of the question the data was gathered by a randomized response experiment and provided via a text file. Students were given a quarter and an index card. They were asked to flip the coin and if it landed tails up to answer "yes" on the card. If the coin landed heads up they are asked to truthfully answer if they have ever sought therapy. Each student will be asked to truthfully identify their nationality.

Additionally, this analysis is to be carried out using a Bayesian and Frequentist methods to analyze the data. A comparison of their methods and results will be included. The data is provided in table 1 in the appendix.

Methods and Results

Frequentist Approach

Several approaches were considered and are briefly discussed below.

1) Difference of Binomial Proportions

We felt that independence could be justified, because the international students are generally from different cultures than the domestic students. We believed that just being at the same school, or in the same culture as the domestic students wouldn't radically alter the embedded cultural norms of international students.

That is, we believed that if an international student was unlikely to seek therapy, because it wasn't acceptable in their society, then being in a temporary and different surrounding was not likely to change their attitude on seeking therapy. Similarly, we believed that domestic students aren't likely to change their views on therapy, just by being in close proximity or interacting with international students.

Unfortunately, the sample size provided is likely too small to yield credible results, so it was not pursued.

2) Fisher's Exact Test

Given the binary variables and small sample size this test seemed appropriate. Unfortunately, it does not provide a means of estimating the domestic and international proportions.

3) Bootstrapping / Resampling With Replacement

Bootstrapping was chosen as a means to overcome the limitations present in the above approaches to analyze the data. Being unfamiliar with the topic of Bootstrapping, initial sources of information and

guidance were the Bootstrapping pages of Wikipedia [1] and the Institute for Digital Research and Education at UCLA [2].

Bootstrapping with replacement was chosen to more simulate and estimate a difference in binomial proportions. A discussion of estimating a difference of binomial proportions and sampling with and without replacement was very helpful with this technique [3]. The article offered Minitab macros and r code to facilitate simulation. The r code, modified for this analysis, is provided in the appendix.

The following plots show the effect of resampling with replacement for various iteration values.

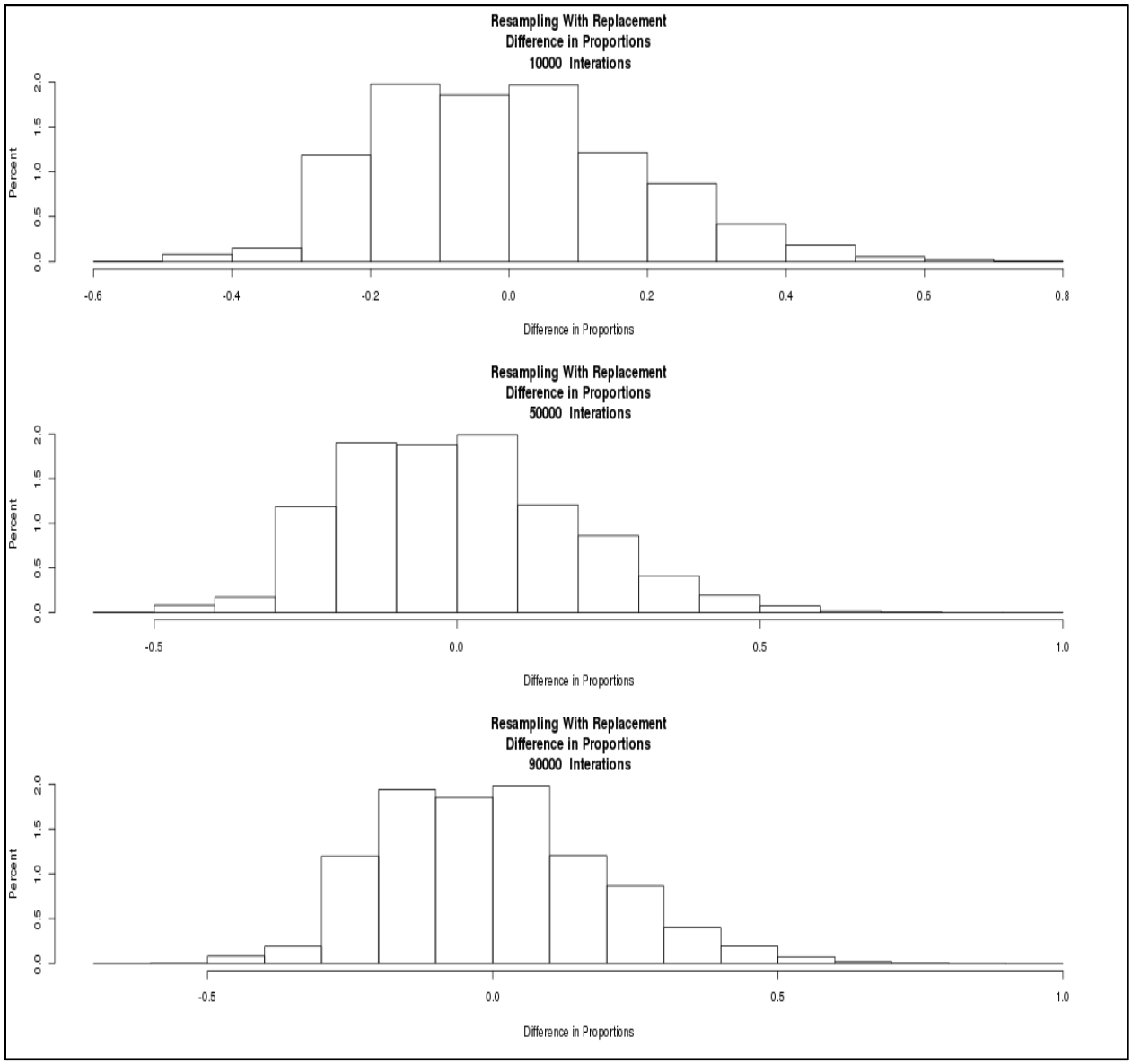

Figure 1: Frequentist Plot of Simulations

The simulations increase from 10,000, to 50,000, and lastly to 90,000 iterations. With increased iterations the range of value decreases slightly, but the difference in proportions clearly straddle the value of no difference in value.

The following table lists details of the resampling simulation.

Iterations	10,000	50,000	90,000
Two-sided p-value	1.00	1.00	1.00
Mean Domestic	0.81796	0.81782	0.81969
Standard Deviation Domestic	0.17106	0.17326	0.17200
Mean International	0.81798	0.81858	0.81875
Standard Deviation International	0.09395	0.09304	0.09334

The two-sided p-value result is one for each of the three resampling iterations. This is not surprising given that all of the domestic students responded "yes" to the question. Further data is needed to overcome this deficiency.

The estimated proportions of domestic and international students who have sought therapy are nearly identical for all resampling iterations.

The plots above and the table values strongly suggest that there is no significant difference between the proportion of domestic and international students who have seen a therapist.

An interval estimate for the estimated international proportion less the estimated domestic proportion is as follows:

5%	95%
-0.29411	0.34117

The interval is generally centered at zero, with the international proportion as much as 29% less than, or almost 35% greater than, the domestic proportion. The zero value provides a strong indication that there is no significant difference between the respective population proportions.

Bayesian Approach

The Bayesian approach we took to analyzing the data is also to compare two binomial proportions. We approached the analysis with the use of two different priors. One prior was a non-informative prior and the second was an informative prior based on our prior beliefs about the location of the domestic and international student proportions.

Beta priors were used for both the non-informative prior and the informative prior. Given the project "hint", they were then incorporated into the posterior log function to estimate the joint posterior distribution for the difference of the logarithm of two binomial variables, as follows:

$$\log_{\left(\frac{1+p.i}{2}\right)} 13 + \log_{\left(\frac{1-p.i}{2}\right)} 4 + lbeta(\alpha.i, \beta.i) - \log_{\left(\frac{1+p.d}{2}\right)} 5 - lbeta(\propto.d, \beta.d)$$

Initially, there was difficulty with convergence using the laplace function from the LearnBayes R package. The following Logit function was used to transform the domestic and international proportion values to the real line.

$$\Theta_d = logit\left(\frac{p.d}{1-p.d}\right), \Theta_i = logit\left(\frac{p.i}{1-p.i}\right)$$

This was following the suggestions found in chapter five of Bayesian Computation with R.

i) Non-informative Prior

The non-informative prior was simply Beta(1,1) for both domesticate and international students. The first step was to estimate starting values and apply laplace in LearnBayes to estimate the joint posterior logit distribution for the domestic and international proportions.

The starting values of (0.99, 0.75) was chosen for the domestic and international proportions respectively. The output for laplace is as follows:

```
> fit.1 = laplace(post.log.fun, start.1, data.1)
> fit.1
$mode
[1] -18.6670583   0.1176503

$var
       [,1]     [,2]
[1,] 25588634 0.0000000
[2,]       0 0.6821494

$int
[1] 4.366081

$converge
[1] TRUE
```

The output suggest the domestic and posterior logit values are roughly -18.67 and 0.12, which are roughly 0.0 and 0.53 domestic and international proportion values.

The output of the laplace values are used as input to a Metropolis random walk to simulate from the joint posterior distribution.

```
> prop.1 = list(var=fit.1$var, scale=.2)
> max.iter.p.d = 30000
> fit.1.2 = rwmetrop(post.log.fun, prop.1, start.1, max.iter.p.d, data.1)
> fit.1.2$accept
```

[1] 0.2882667

The acceptance rate is close to the desired acceptance rate for the algorithm. Although the following density plot appears reasonable, the algorithm does not closely approximate the laplace output.

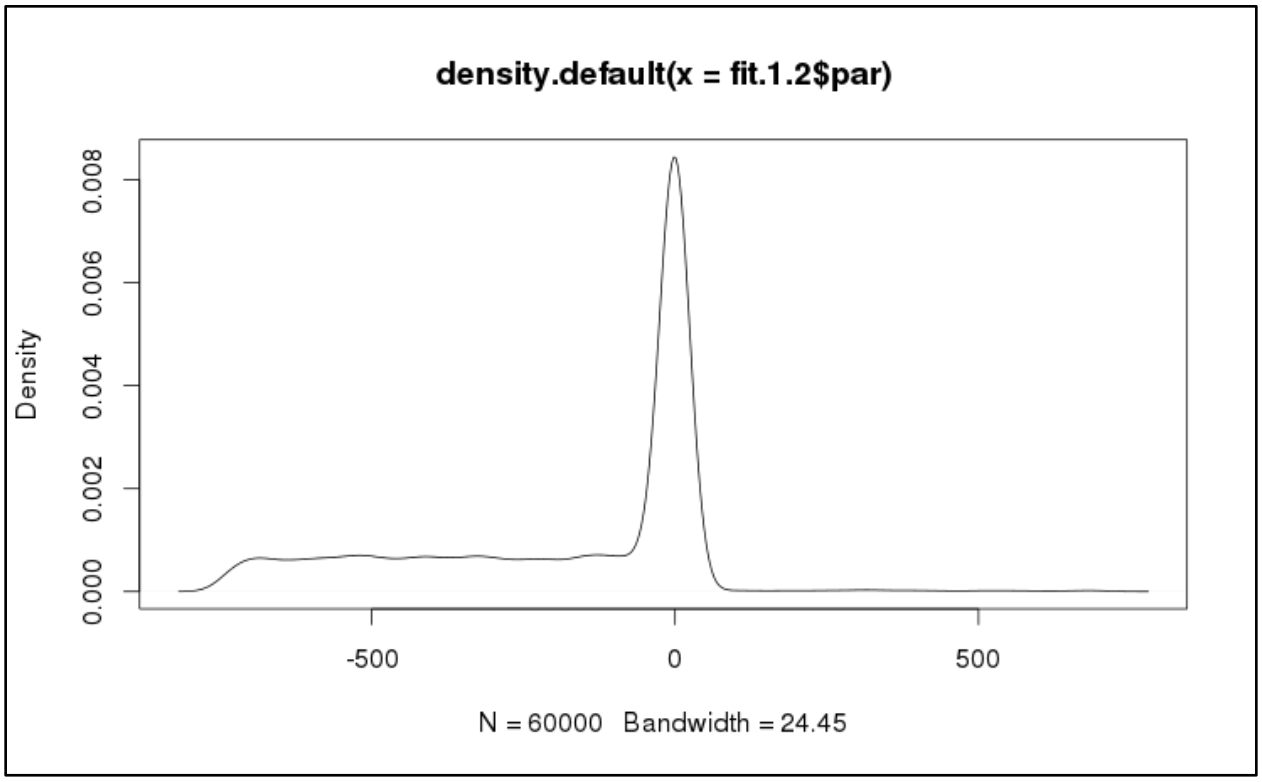

Figure 2: rwmetrop Density

This can be seen by comparing the means and standard deviations of the function laplace with the values from the MCMC algorithm.

```
> cbind(c(fit.1$mode), modal.sds)
                        modal.sds
[1,] -18.6670583      5058.5209544
[2,]   0.1176503         0.8259233

> cbind(post.means, post.sds)
      post.means          post.sds
[1,] -354.3477818       247.242136
[2,]   -0.4411714         1.446949
```

Unfortunately, we have tried numerous starting points and data values and although we can get a reasonable acceptance rate from the algorithm, we cannot find a combination of values to produce closer means or standard deviation, nor acceptable trace plots as shown below.

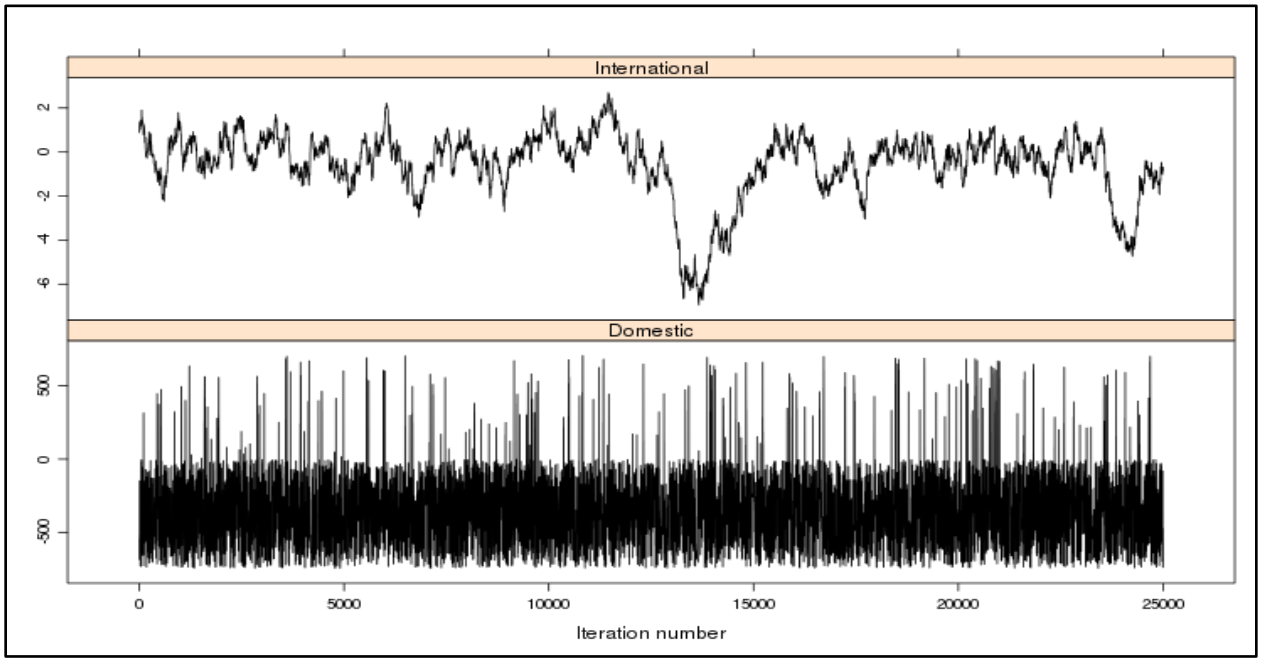

Figure 3 Trace Plots Uniform Priors

The following autocorrelation plots suggest a tremendous amount of dependence for the international proportion.

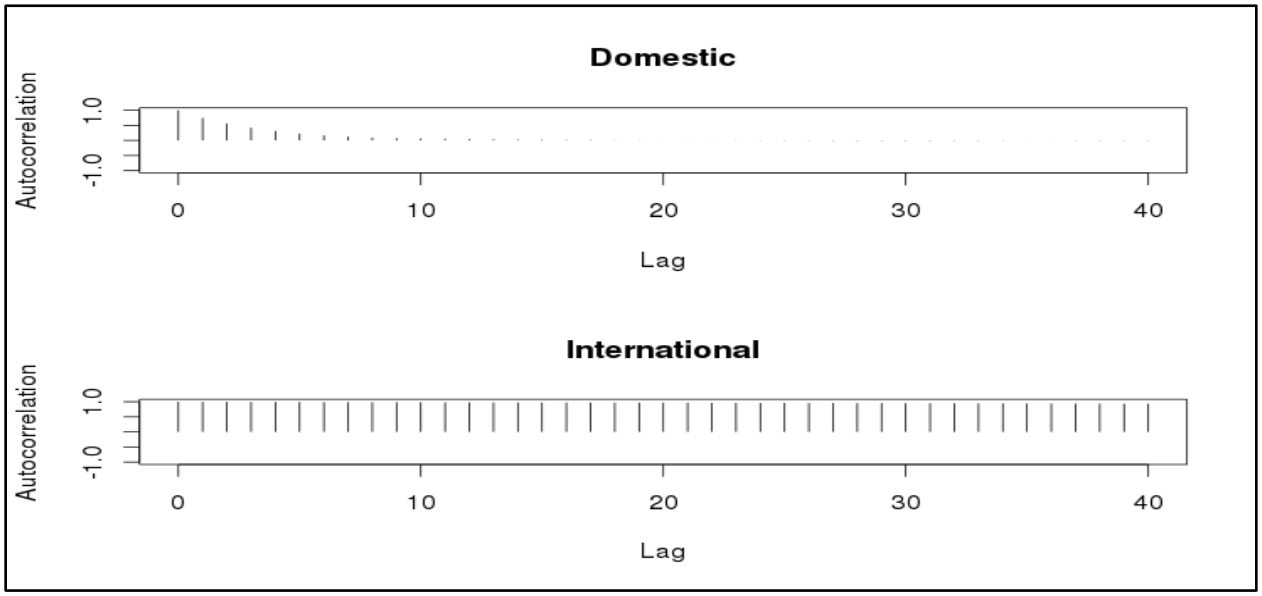

Figure 4: Autocorrelation Plots Uniform Priors

A lack of adequate fit is corroborated by the following contour plot.

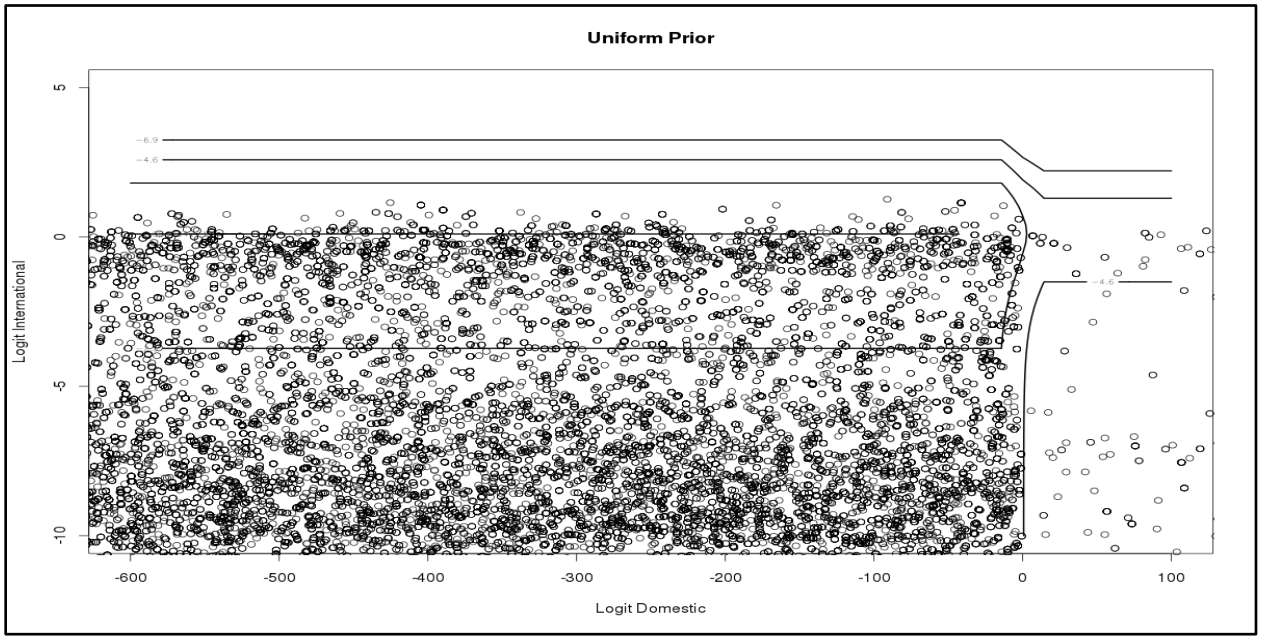

Figure 5: Contour Plot Uniform Prior

iI) Informative Prior

The informative prior was based on Beta(0.5,0.5) for both domesticate and international students. This was after a dozen or more attempts to avoid the following errors message when using the laplace function.

```
Error in solve.default(fit$hessian) :
  Lapack routine dgesv: system is exactly singular
```

and

```
Error in optim(mode, logpost, gr = NULL, ..., hessian = TRUE, control = list(fnscale = -1)) :
  non-finite finite-difference value [2]
```

The error messages could only be avoided when using a flatter prior. Although, Beta(0.5, 0.5) is relatively "U" shaped, which surprised us that it was successful when used with laplace.

After settling on Beta(0.5, 0.5) priors, the next step was to estimate starting values and apply laplace in LearnBayes to estimate the joint posterior logit distribution for the domestic and international proportions. The starting values of (0.99, 0.75) were used again for the domestic and international proportions respectively. The output for laplace is as follows:

```
> fit.2 = laplace(post.log.fun, start.2, data.2)
> fit.2
$mode
```

```
[1] -2.0792506  0.1176088

$var
          [,1]      [,2]
[1,] 2.812298 0.0000000
[2,] 0.000000 0.6821651

$int
[1] -5.271151

$converge
[1] TRUE
```

The output suggest the domestic and posterior logit values are roughly -2.08 and 0.12, which are roughly 0.11 and 0.53 domestic and international proportion values.

The output of the laplace values are used as input to a Metropolis random walk to simulate from the joint posterior distribution.

```
> prop.2 = list(var=fit.2$var, scale=750)
> max.iter.p.i = 30000
> fit.2.2 = rwmetrop(post.log.fun, prop.2, start.2, max.iter.p.i, data.2)
> fit.2.2$accept
[1] 0.2143333
```

The acceptance rate is close to the desired acceptance rate for the algorithm. Although the following density plot appears reasonable, the algorithm does not closely approximate the laplace output.

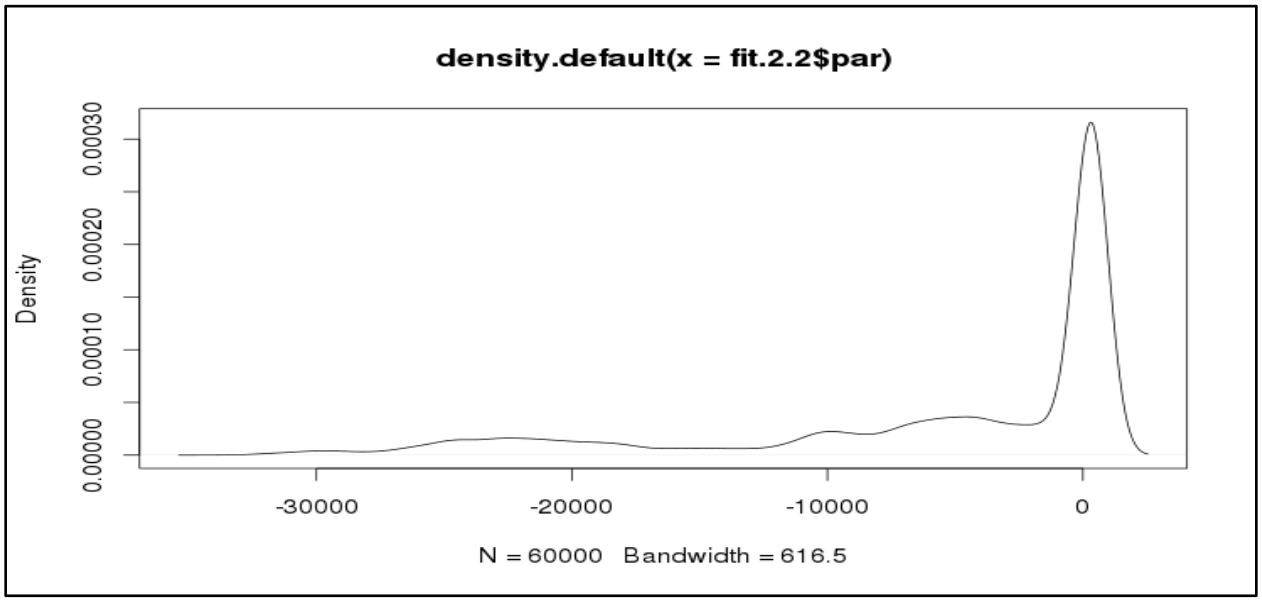

Figure 6 rwmetrop Density

This can be seen by comparing the means and standard deviations of the function laplace with the values from the MCMC algorithm.

```
> cbind(c(fit.2$mode), modal.sds)
                        modal.sds
[1,] -2.0792506         1.6769906
[2,]  0.1176088         0.8259329
> cbind(post.means, post.sds)
     post.means         post.sds
[1,]    338.8915        211.7615
[2,] -10904.5547        8538.1000
```

Unfortunately, we again have tried numerous starting points and data values and although we can get a reasonable acceptance rate from the algorithm, we cannot find a combination of values to produce closer means or standard deviation, nor acceptable trace plots as shown below.

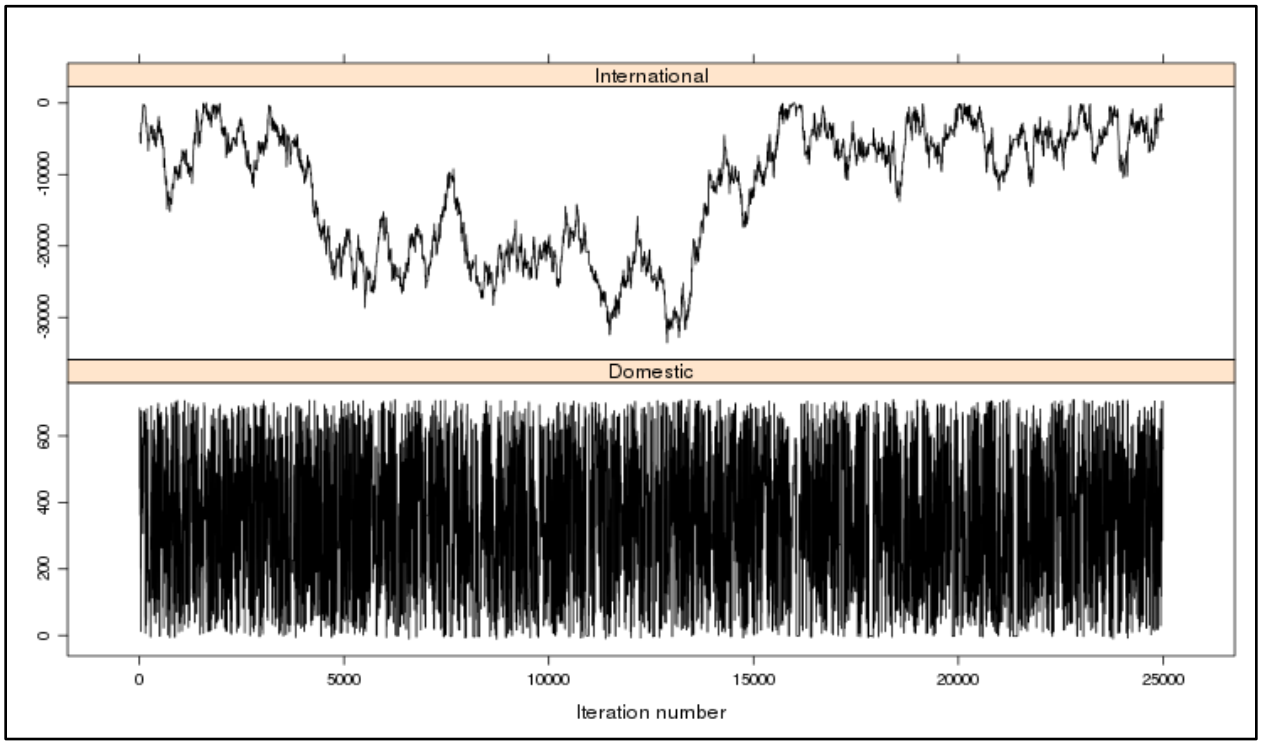

Figure 7: Trace Plot Informed Priors

The following autocorrelation plots suggest a tremendous amount of dependence for the international proportion.

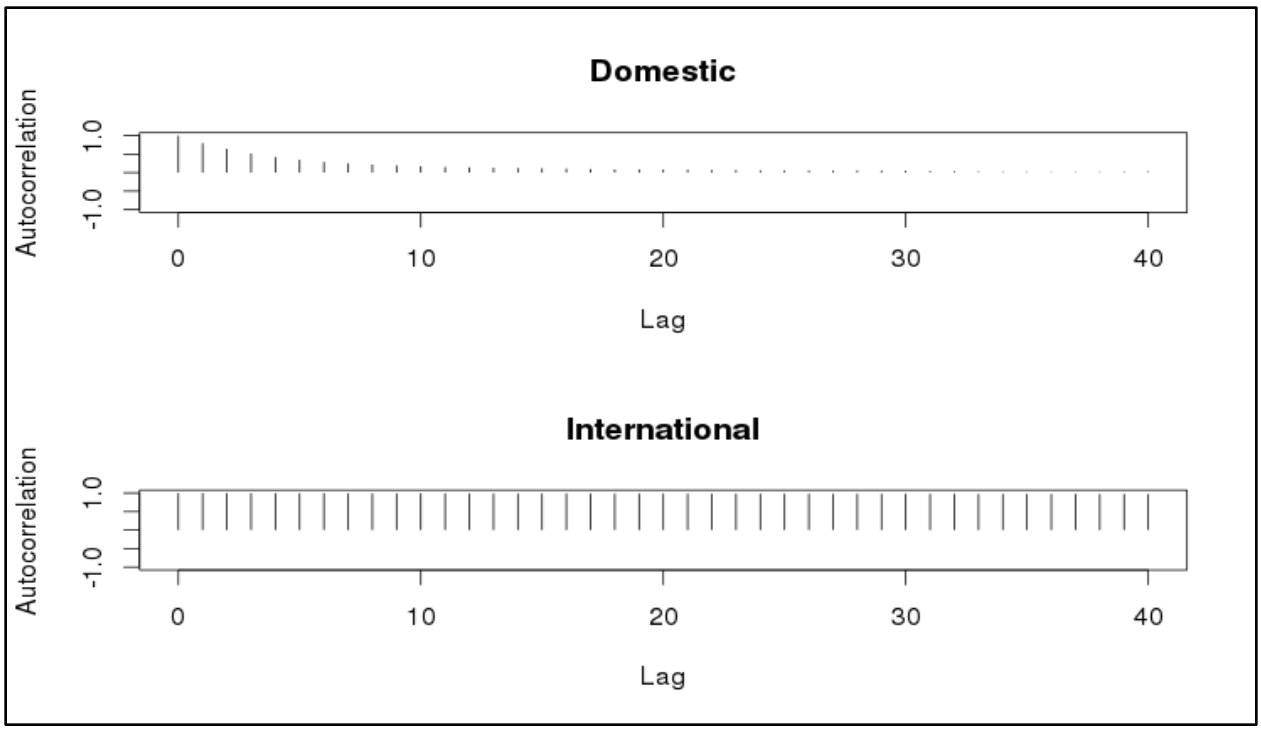

Figure 8: Autocorrelation Plot Informed Priors

A lack of adequate fit is corroborated by the following contour plot.

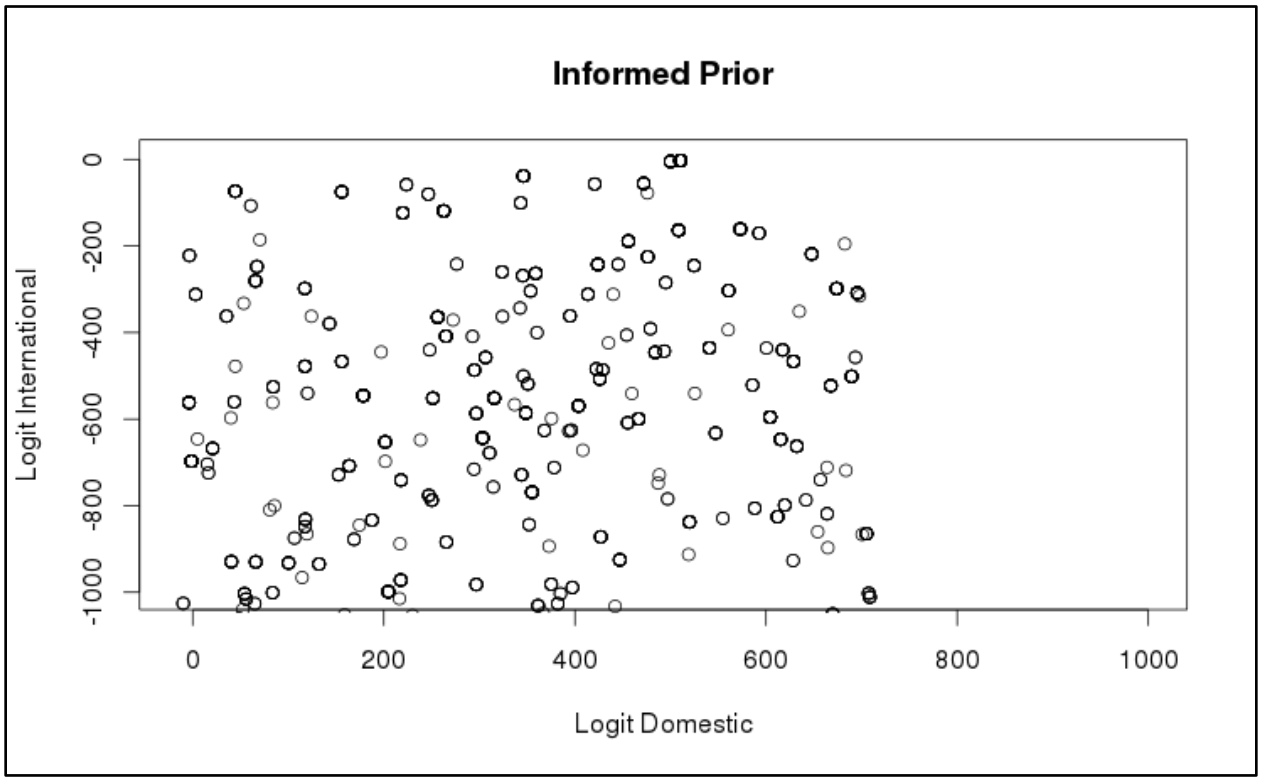

Figure 9: Contour Plot Informed Priors

iii) Bayes Factor

The following is the calculation of the Bayes Factor for first model, uniform priors, and the second model, informative priors.

```
> logmarg = c(fit.1$int, fit.2$int)
> exp(logmarg[1] - logmarg[2])
[1] 15324.86
```

The Bayes Factor value was calculated by the use of the post.log.fun function, shown below, used to calculate the joint posterior distribution for domestic and international logit values.

```
post.log.fun = function(theta, data)
{
 p.dom = exp(theta[1])/(1+exp(theta[1]))
 p.int = exp(theta[2])/(1+exp(theta[2]))

 alpha.dom = data[1]
 beta.dom  = data[2]
 alpha.int = data[3]
 beta.int  = data[4]

 # domestic
 log.dom.p =
  5*log(1 + p.dom) - 5*log(2) +
  (alpha.dom - 1) * log(p.dom) +
  lbeta(alpha.dom, beta.dom)

 # international
 log.int.p =
  13*log(1 + p.int) - 13*log(2) +
  4*log(1 - p.int) - 4*log(2) +
  lbeta(alpha.int, beta.int)

 # in keeping with freqentist resampling, int - dom
 val = log.int.p - log.dom.p
 return(val)
}
```

The value of 15,324 was calculated following the examples of section 8.6 in the course text. It large positive value suggests that there is large support for the first model, that of non-informative priors.

Conclusion

Both the Frequentist and Bayesian approaches came to the conclusion that is unlikely that the proportions of domestic and international students seeking therapy are very different.

Due to the small sample size we chose to employ a simulation based resampling approach to estimate the distribution of differences in domestic and international proportions. A plot of the distribution for several iteration values was produced. The simulation estimated roughly 0.82 for both the domestic and international proportions, and a confidence interval suggested zero was a likely value for their difference.

The Bayesian approach developed a posterior log function for the difference in binomial proportions, and used this function with the LearnBayes laplace function to estimate the mode for the joint posterior logit distribution. The output values for the laplace function were then used as a starting point for the Random Walk Metropolis Algorithm. Although we could generate acceptable acceptance rates for the algorithm, we were unable to find appropriate choices for starting and scale values for both uniform and informed priors. The trace plots of the international values did not look randomly generated like white noise and the autocorrelation plots showed strong dependency. We were not able to overcome this deficiency in the time allowed.

Both Frequentist and Bayesian methods have benefits and drawbacks. When the assumptions are satisfied for a Frequentist method, then it may be easier to communicate results to a non-technical audience. We feel that the Bayesian approach to statistical inference is more unified.

We have enjoyed this course and have learned a lot. It was a bit of a psychological shift to start to think in priors, but once one is comfortable with the idea of using priors to estimate parameters, the feeling is rather liberating. The idea of priors seems to be a more natural way to model. In a similar manner to when one is introduced to hierarchical modeling. Initially, hierarchical modeling feels like unnecessary detail intended to confuse rather than illuminate, but when you become more comfortable with it you realize that you're actually more accurately modeling the problem space.

References

[1] Bootstrapping (statistics). (2013, December 9). In Wikipedia, The Free Encyclopedia. Retrieved 19:40, December 16, 2013, from http://en.wikipedia.org/w/index.php?title=Bootstrapping_(statistics)&oldid=585276212

[2] R Library: Introduction to bootstrapping

www.ats.ucla.edu/stat/r/library/boostrap.html

[3] Using Resampling to Compare Two Proportions

http://www.public.iastate.edu/~wrstephe/Resampling_Final.pdf

Appendix

Table 1: Student Data

	Yes	No	Total by Student
International	13	4	17
Domestic	5	0	5
Total by Answer	18	4	22

R Code

The r code for the Frequentist and Bayesian approaches is provided by the associated HTML project file submitted with this project.