Homework 08 - STAT440

Joseph Sepich (jps6444)

11/01/2020

```
set.seed(42)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(MASS)
```

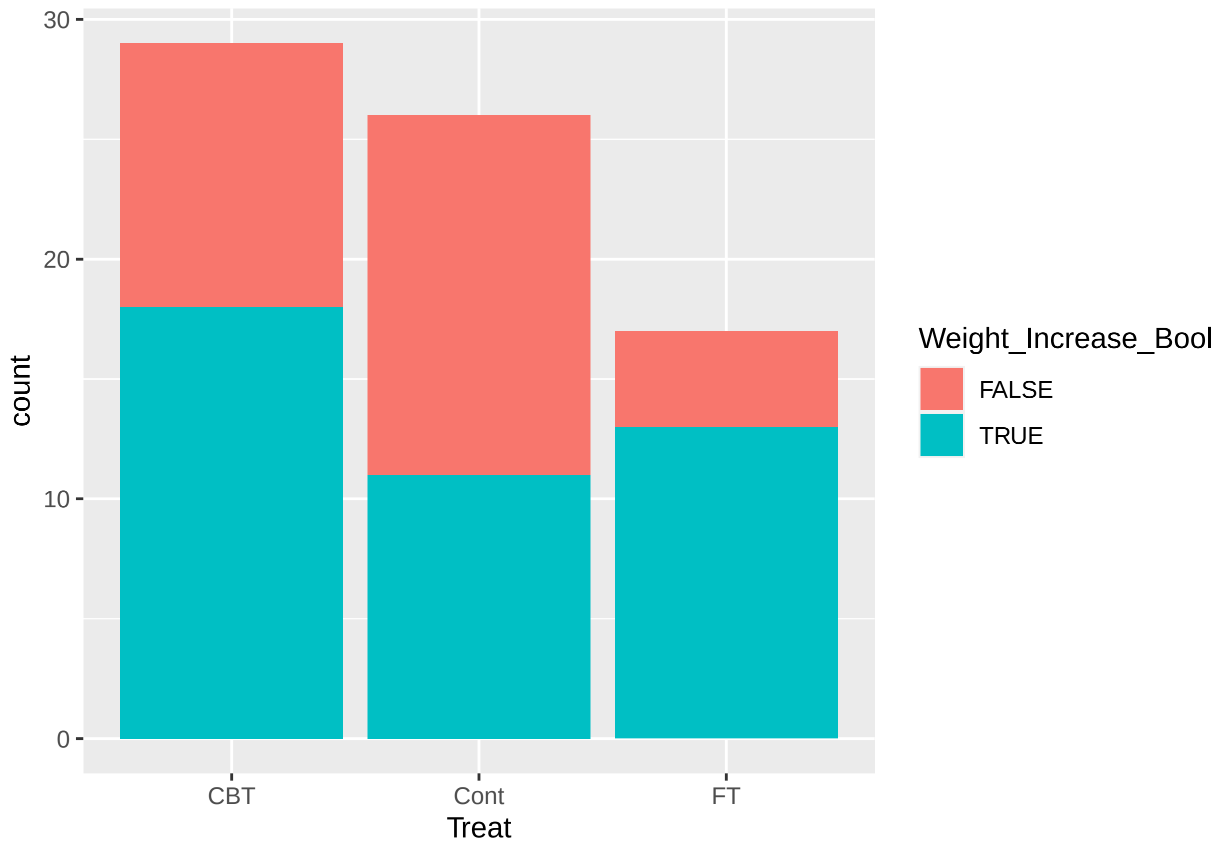
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
df <- anorexia
```

Problem 1

```
df$Weight_Increase <- as.numeric(df$Postwt > df$Prewt)
df$Weight_Increase_Bool <- df$Postwt > df$Prewt
```

```
df %>%
    ggplot(aes(x=Treat)) +
    geom_bar(aes(fill=Weight_Increase_Bool))
```



Problem 2

Simplifying our data into a simply binary variable removes the weight/value of the data that comes with analyzing the difference in weight from pre to post treatment. This could be both good and bad. Sometimes we like to use this information to help quantify how significant of a change occurs; however since we are going to be looking between groups, and mostly care if one group had a significant increase versus another, we care more about comparing the amount of change between groups. In this way removing the value of differences of each person is good, because people are all different, so the extremity of change can vary a lot, so analyzing with this binary variable will not be greatly affected by outliers.

Another reason to use a binary variable is we know what distribution it comes from, a bernoulli distribution. This can help us perform various analysis. On the other hand we do not know what kind of distribution the pairwise differences may have, even if we could guess.

Problem 3

The conjugate prior we will use for the Bernoulli distribution is the Beta distribution. This makes our prior and posterior functions a beta distribution while our likelihood function is a Bernoulli distribution:(

$$\pi(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}\theta^{\beta-1}}{B(\alpha\beta)}$$

$$f(X_i|\theta) = \theta^{X_i}(1-\theta)^{1-X_i}$$

The parameters in the posterior become:

$$\alpha \rightarrow \alpha + \Sigma X_i$$
$$\beta \rightarrow \beta + n - \Sigma X_i$$

Problem 4

We will use the initial prior parameters of $\alpha = \beta = 1$. Since higher α values pull the probability near $\theta = 0$ up and β does the same for near $\theta = 1$ and vice versa for lower, initializing for $\alpha = \beta = 1$ gives the density of the parameter theta a uniform distribution, so our prior belief is that it is equally likely to be anywhere from 0 to 1. Now let's calculate the value of the posterior parameters:

```
# prior params
alpha_prior <- 1
beta_prior <- 1

# CBT Posterior
obs_cbt <- df %>%
    filter(Treat == 'CBT')
sum_obs <- sum(obs_cbt$Weight_Increase)
alpha_cbt <- alpha_prior + sum_obs
beta_cbt <- beta_prior + length(obs_cbt$Weight_Increase) - sum_obs
print(paste0('CBT: alpha ',alpha_cbt,', beta ', beta_cbt))
```

```
## [1] "CBT: alpha 19, beta 12"
```

```
# Cont Posterior
obs_cont <- df %>%
    filter(Treat == 'Cont')
sum_obs <- sum(obs_cont$Weight_Increase)
alpha_cont <- alpha_prior + sum_obs
beta_cont <- beta_prior + length(obs_cont$Weight_Increase) - sum_obs
print(paste0('Cont: alpha ',alpha_cont,', beta ', beta_cont))
```

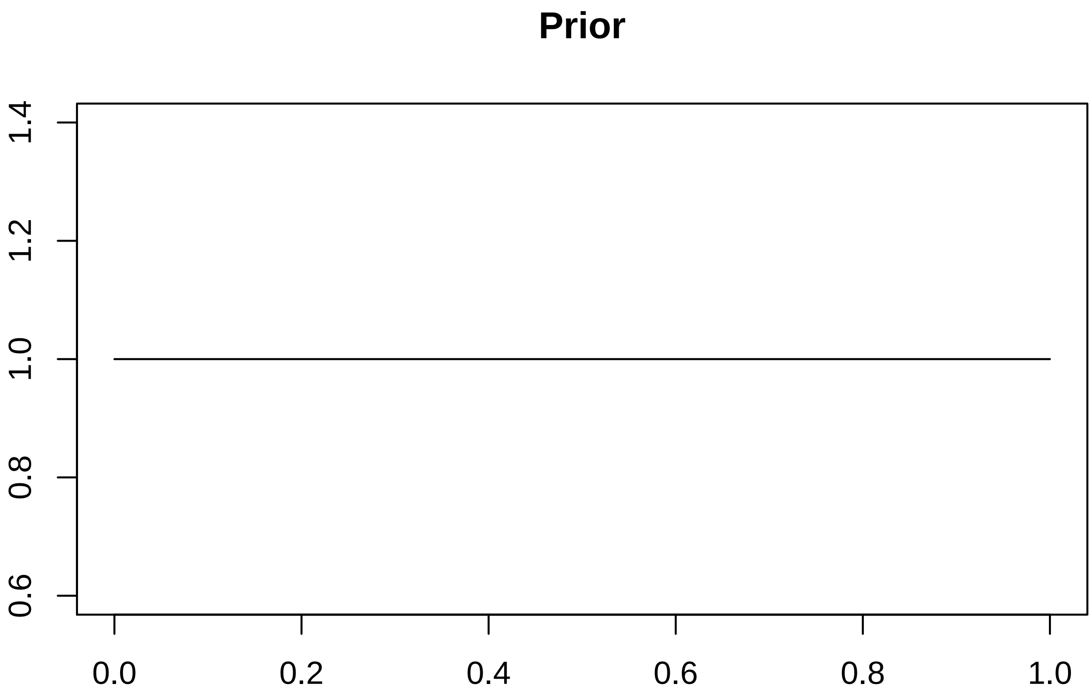
```
## [1] "Cont: alpha 12, beta 16"
```

```
# FT Posterior
obs_ft <- df %>%
    filter(Treat == 'FT')
sum_obs <- sum(obs_ft$Weight_Increase)
alpha_ft <- alpha_prior + sum_obs
beta_ft <- beta_prior + length(obs_ft$Weight_Increase) - sum_obs
print(paste0('FT: alpha ',alpha_ft,', beta ', beta_ft))
```

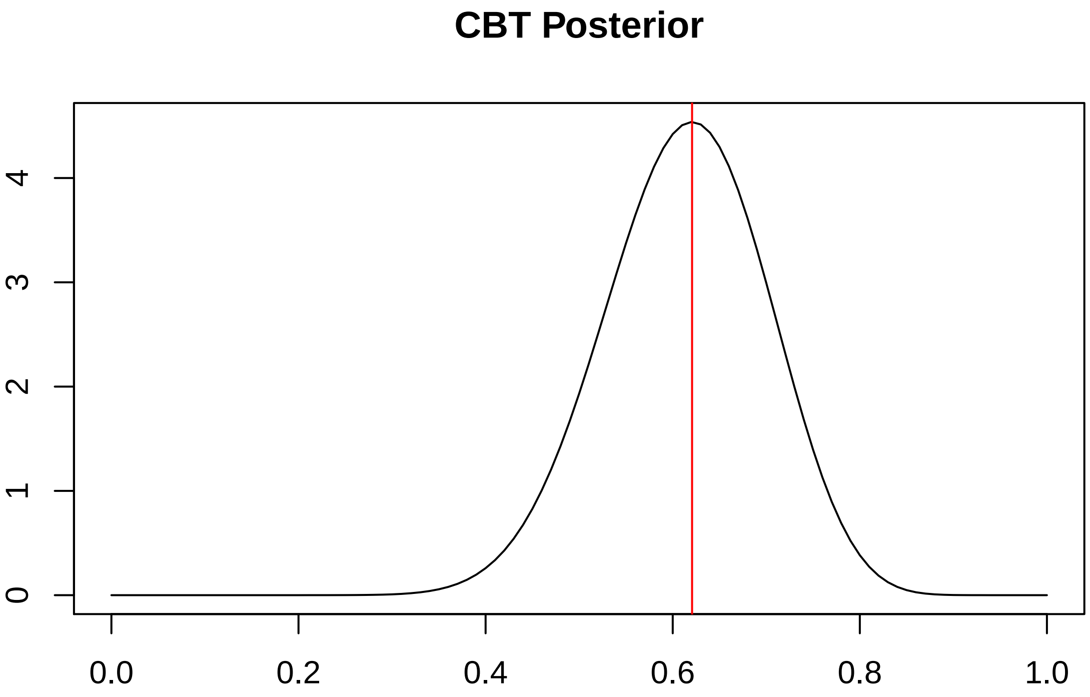
```
## [1] "FT: alpha 14, beta 5"
```

Problem 5

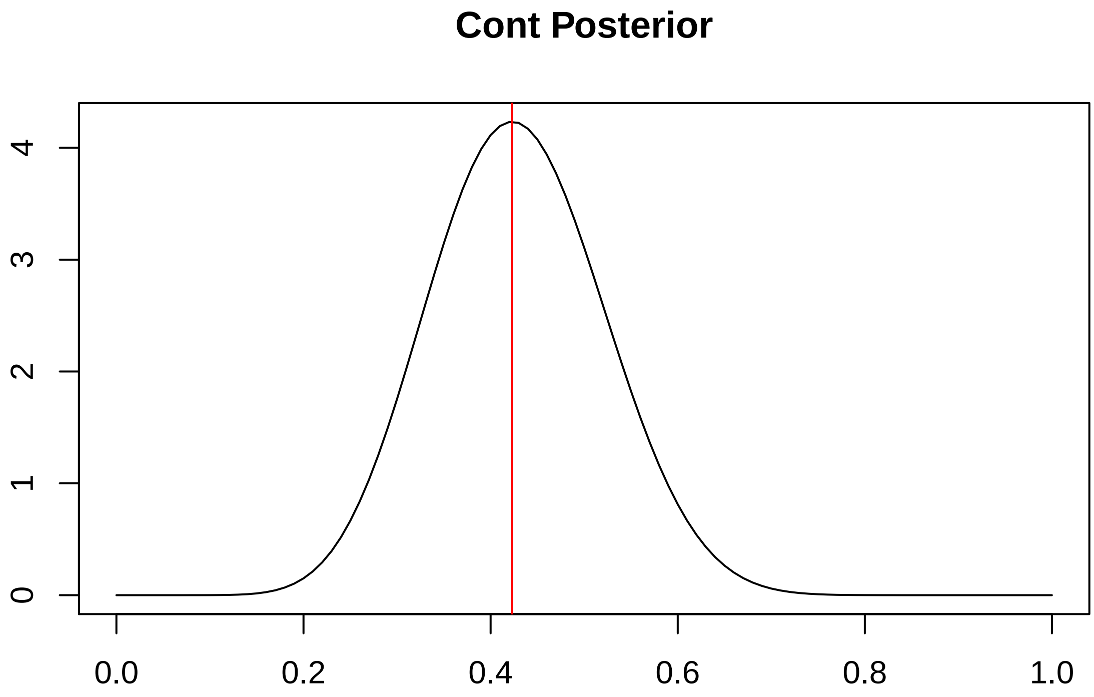
```
# all same prior
curve(dbeta(x,shape1=alpha_prior,shape2=beta_prior),from=0,to=1,xlab="",ylab="",main="Prior")
```



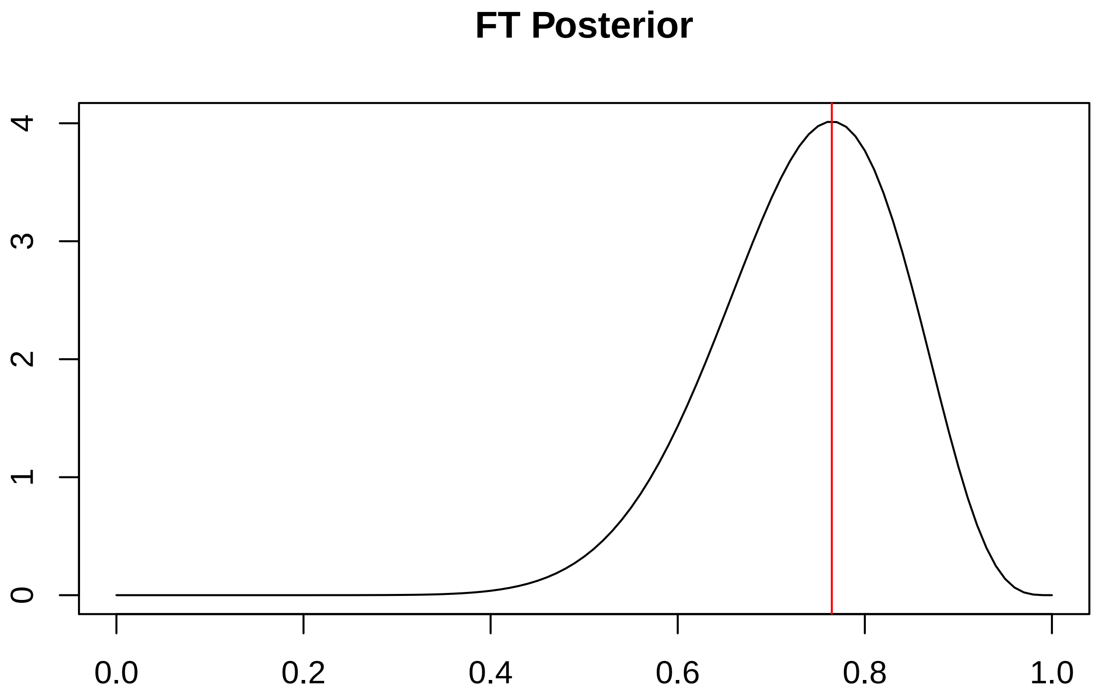
```
# CBT Posterior
curve(dbeta(x,shape1=alpha_cbt,shape2=beta_cbt),from=0,to=1,xlab="",ylab="",main="CBT Posterior")
abline(v=sum(obs_cbt$Weight_Increase/length(obs_cbt$Weight_Increase)), col='red')
```



```
# Cont Posterior
curve(dbeta(x,shape1=alpha_cont,shape2=beta_cont),from=0,to=1,xlab="",ylab="",main="Cont Posterior")
abline(v=sum(obs_cont$Weight_Increase/length(obs_cont$Weight_Increase)), col='red')
```



```
# FT Posterior
curve(dbeta(x,shape1=alpha_ft,shape2=beta_ft),from=0,to=1,xlab="",ylab="",main='FT Posterior')
abline(v=sum(obs_ft$Weight_Increase/length(obs_ft$Weight_Increase)), col='red')
```



Based on the plots the CBT group had a weight increase around $\theta = 0.6$, which means roughly 60% of the cases increased. Similarly the Cont group was around $\theta - 0.4$ and the FT group was around $\theta = 0.7$.

Problem 6

```
M <- 1000

CBT <- rbeta(M, alpha_cbt, beta_cbt)
Cont <- rbeta(M, alpha_cont, beta_cont)
FT <- rbeta(M, alpha_ft, beta_ft)

diff_cbt <- CBT - Cont
diff_ft <- FT - Cont

alpha_conf <- 0.05
cred_cbt <- quantile(diff_cbt, probs=c(alpha_conf/2, 1-alpha_conf/2))
cred_ft <- quantile(diff_ft, probs=c(alpha_conf/2, 1-alpha_conf/2))

print(paste0('Credible interval CBT - Cont: ', cred_cbt))
```

```
## [1] "Credible interval CBT - Cont: -0.0744525642945931"
## [2] "Credible interval CBT - Cont: 0.409060391213553"
```

```r
print(paste0('Credible interval FT - Cont: ', cred_ft))
```

```
## [1] "Credible interval FT - Cont: 0.0276021849805676"
## [2] "Credible interval FT - Cont: 0.545706587075375"
```

The credible interval for the difference between the FT and Control treatment does not contain 0. This means the data shows evidence that the success rate of weight increase in the FT group outpaces the success rate of the weight increase in the control group with 95% credibility. The CBT interval does not display this, so we do not have enough evidence in the data to support that the success rate of the weight increase in the CBT treatment group is better than the success rate of weight increase in the control group.