Exercise 3.2.1

Saurabh Steixner-Kumar
(social): - -

Contents

1 Question

Exercise 3.2.1 Compare the data sets k1 = 8, n1 = 10, k2 = 7, n2 = 10 and k1 = 80, n1 = 100, k2 = 70, n2 = 100. Before you run the code, try to predict the effect that adding more trials has on the posterior distribution for δ.

2 Comments/Solution

Looking at the posterior distribution from the previous exercises we know that adding more data will make the posterior estimate narrower and less uncertain. Now we can also run the model script of the two datasets and observe these effects. Have a look at the plots section below.

In different words: When you have more information (i.e., high n) the posteriors–for the individual rates, as well as for the difference between them that is of interest—become more peaked. This means that you are more certain about what values for the difference are plausible, and what values are not.

The model used to calculate the required values and the plots is scripted below. Copy/pasting the given code will generate the same result on your own machine.

3 Code

3.1 libraries

The libraries required for the script and the plots.

```
# clears workspace
rm(list=ls())
```

```r
#load libraries
library(rstan)
library(ggplot2)
library(patchwork)
```

3.2 Data

The data required for this particular stan model.

```r
# data  initialization
k1 <- 8;n1 <- 10;k2 <- 7;n2 <- 10
# to be passed on to Stan
stan_data <- list(k1 = k1, n1 = n1, k2 = k2, n2 = n2)
#
k1 <- 80;n1 <- 100;k2 <- 70;n2 <- 100
# to be passed on to Stan
stan_data_1 <- list(k1 = k1, n1 = n1, k2 = k2, n2 = n2)
```

3.3 Stan code

Stan code, that can be written in R as such or in a separate new file with stan extension.

```r
write("// Stan code here in this section

// Inferring delta through theta1 and theta2
data {
  int<lower=1> n1;
  int<lower=1> n2;
  int<lower=0> k1;
  int<lower=0> k2;
}
parameters {
  real<lower=0,upper=1> theta1;
  real<lower=0,upper=1> theta2;
}
transformed parameters {
  real<lower=-1,upper=1> delta;
  delta = theta1 - theta2;
}
model {
  // Prior Distribution for Rate Theta
  theta1 ~ beta(1, 1);
  theta2 ~ beta(1, 1);
  // Observed Counts
  k1 ~ binomial(n1, theta1);
  k2 ~ binomial(n2, theta2);
} // ",

"3_2_1.stan")
```

3.4 code in R to run stan

Running stan through R (with the required input parameters).

```
myinits <- list(
  list(theta1=.1,theta2=.9),  # chain 1 starting value
  list(theta1=.9,theta2=.1))  # chain 2 starting value

# parameters to be monitored:
parameters <- c("delta", "theta1", "theta2")

# The following command calls Stan with specific options.
# For a detailed description type "?stan".
mod_fit <- stan(file="3_2_1.stan",
                data=stan_data,
                init=myinits,  # If not specified, gives random inits
                pars=parameters,
                iter=2000,
                chains=2,
                thin=1,
                warmup=100,  # Stands for burn-in; Default = iter/2
                seed=123  # Setting seed; Default is random seed
)
mod_fit_1 <- stan(file="3_2_1.stan",
                data=stan_data_1,
                init=myinits,  # If not specified, gives random inits
                pars=parameters,
                iter=2000,
                chains=2,
                thin=1,
                warmup=100,  # Stands for burn-in; Default = iter/2
                seed=123  # Setting seed; Default is random seed
)
```

4 Outputs

4.1 Model summary

In order of definition.

```
## Inference for Stan model: 3_2_1.
## 2 chains, each with iter=2000; warmup=100; thin=1;
## post-warmup draws per chain=1900, total post-warmup draws=3800.
##
##          mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
## delta    0.08    0.00 0.18  -0.26  -0.04   0.09   0.20   0.42  3786    1
## theta1   0.75    0.00 0.12   0.48   0.67   0.77   0.84   0.94  3576    1
## theta2   0.67    0.00 0.13   0.39   0.58   0.68   0.76   0.90  3664    1
## lp__   -15.44    0.02 1.04 -18.22 -15.85 -15.13 -14.69 -14.42  1732    1
##
## Samples were drawn using NUTS(diag_e) at Thu Nov 05 21:15:42 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

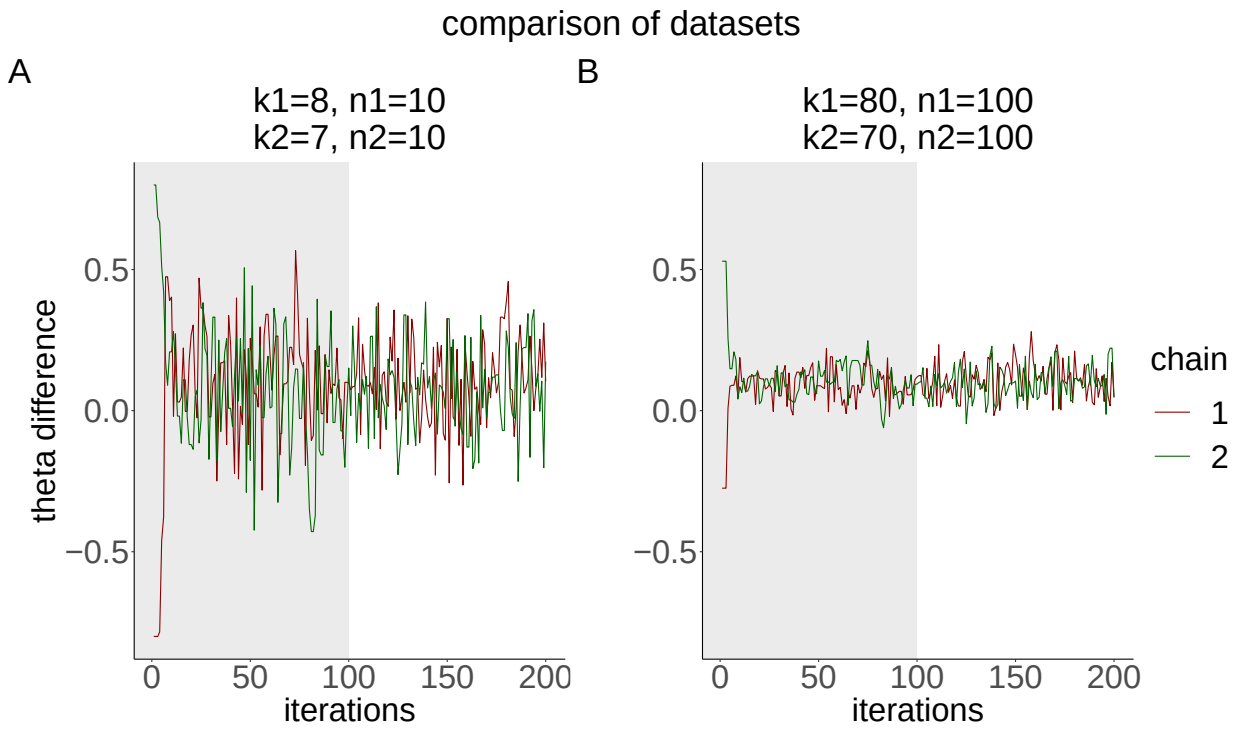
```
## Inference for Stan model: 3_2_1.
## 2 chains, each with iter=2000; warmup=100; thin=1;
## post-warmup draws per chain=1900, total post-warmup draws=3800.
```

```
##
##           mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff
## delta     0.10    0.00 0.06   -0.02    0.06    0.10    0.14    0.22  3801
## theta1    0.79    0.00 0.04    0.71    0.77    0.80    0.82    0.87  3720
## theta2    0.70    0.00 0.05    0.60    0.67    0.70    0.73    0.78  4058
## lp__   -115.58    0.02 1.06 -118.38 -116.01 -115.24 -114.82 -114.53  1798
##        Rhat
## delta     1
## theta1    1
## theta2    1
## lp__      1
##
## Samples were drawn using NUTS(diag_e) at Thu Nov 05 21:15:42 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

4.2 Plots

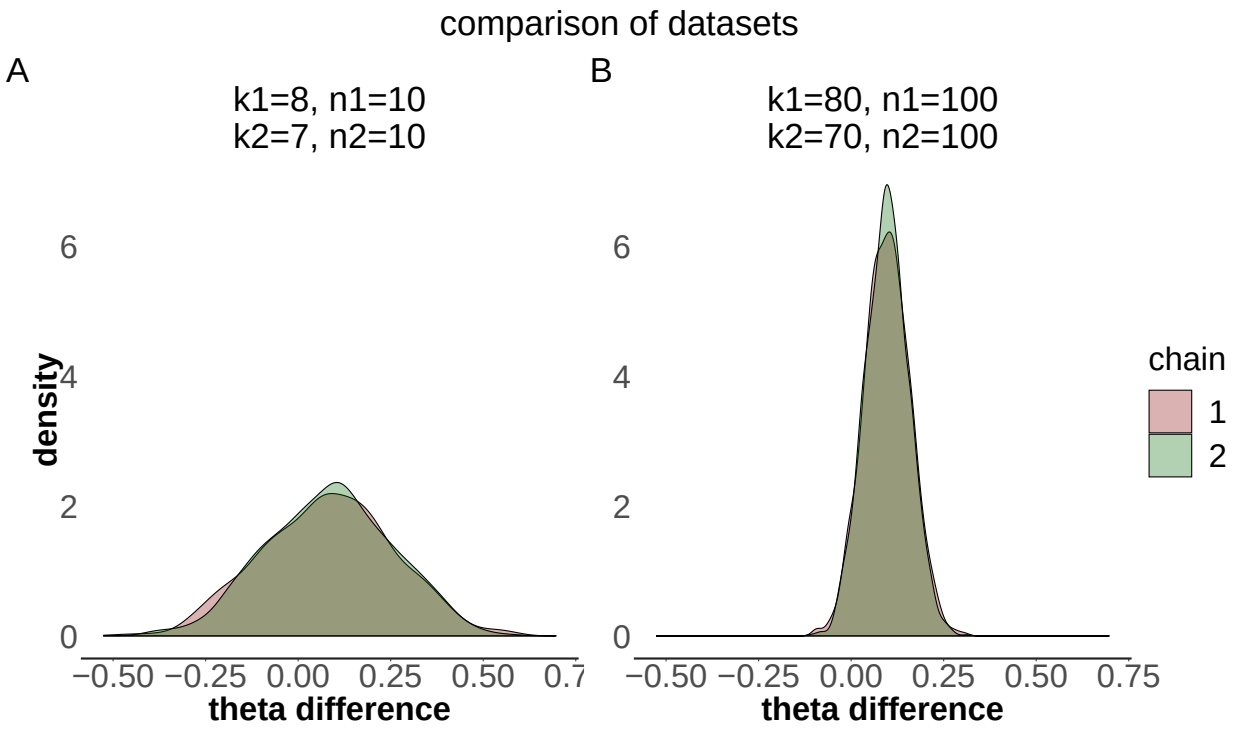
4.2.1 Plot (chains)

The initial movement of the chains are shown here (including the warmup phase). The two chains begin from the initial starting points of as defined in the input parameters of the stan model.



4.2.2 Plot (posterior)

The plot of the δ values per chain superimposed on each other.


comparison of datasets
A
k1=8, n1=10
k2=7, n2=10
B
k1=80, n1=100
k2=70, n2=100
density
theta difference
theta difference
chain
1
2
0
2
4
6
−0.50
−0.25
0.00
0.25
0.50
0.75