Lab4

Bingying Liu

2/10/2020

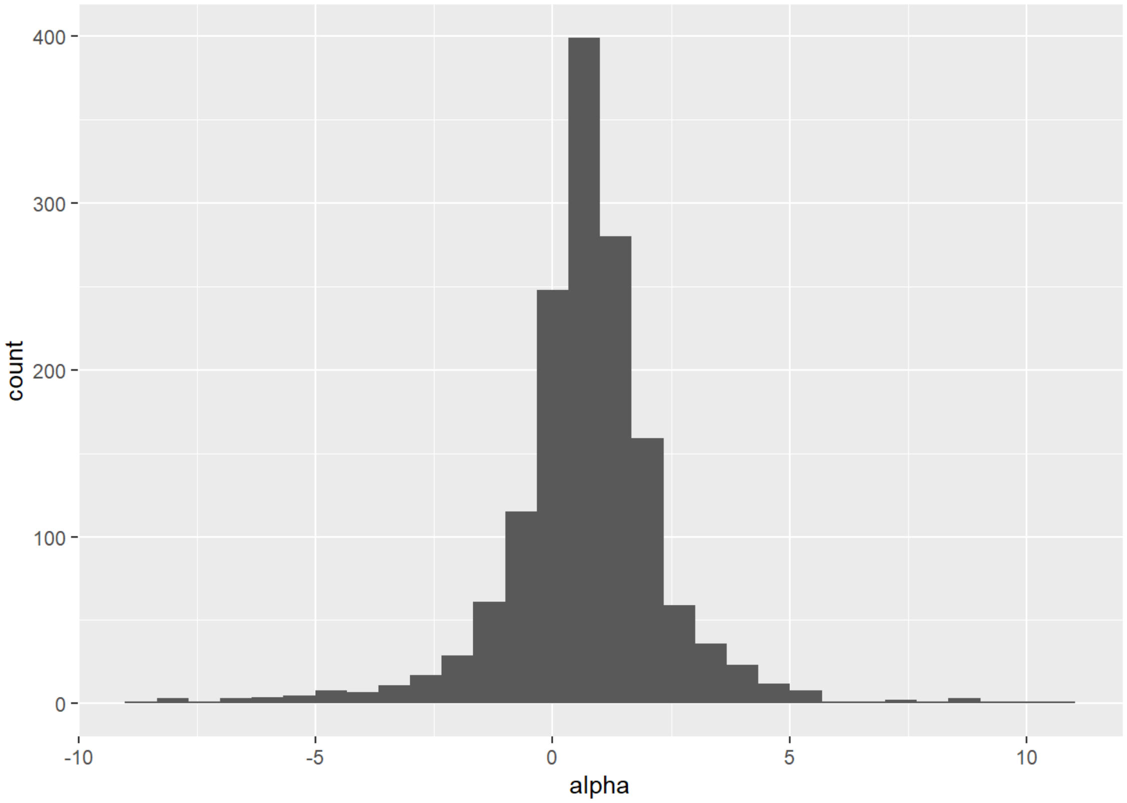
Exercise 1: Write down the posterior means of α and β. Give 95% credible intervals for each. Considering the amount of data we have, do the results seem surprising?

```
# Prior selection
set.seed(689934)
alpha <- 1
beta <- -0.25
sigma <- 1

N <- 5
x <- array(runif(N, 0, 2), dim=N)
y <- array(rnorm(N, beta * x + alpha, sigma), dim=N)

stan_dat <- list(y = y, x=x, N=N)
fit.flat <- stan(file = "lab-04-flat_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 2000, warmup =
500, seed=48)
alpha.flat <- as.matrix(fit.flat, pars = "alpha")
beta.flat <- as.matrix(fit.flat, pars = "beta")

ggplot(alpha.flat %>% as.data.frame, aes(x = alpha)) +
  geom_histogram(bins = 30)
```



```
print(fit.flat, pars = c("alpha"))
```

```
## Inference for Stan model: lab-04-flat_prior.
## 1 chains, each with iter=2000; warmup=500; thin=1;
## post-warmup draws per chain=1500, total post-warmup draws=1500.
##
##       mean se_mean   sd  2.5%   25%  50%  75% 97.5% n_eff Rhat
## alpha  0.7    0.13 1.74 -3.34 -0.01 0.75 1.46  4.05   172    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 10 23:39:47 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).
```

```
mean(alpha.flat) #0.6982
```

```
## [1] 0.6982131
```

```
mean(beta.flat) #0.3537
```

```
## [1] 0.3537682
```

```
quantile(alpha.flat, probs = c(0.025, 0.975)) #-3.34, 4.04
```

```
##      2.5%     97.5%
## -3.338598  4.049911
```

```
quantile(beta.flat, probs = c(0.025, 0.975)) #-2.27, 2.93
```

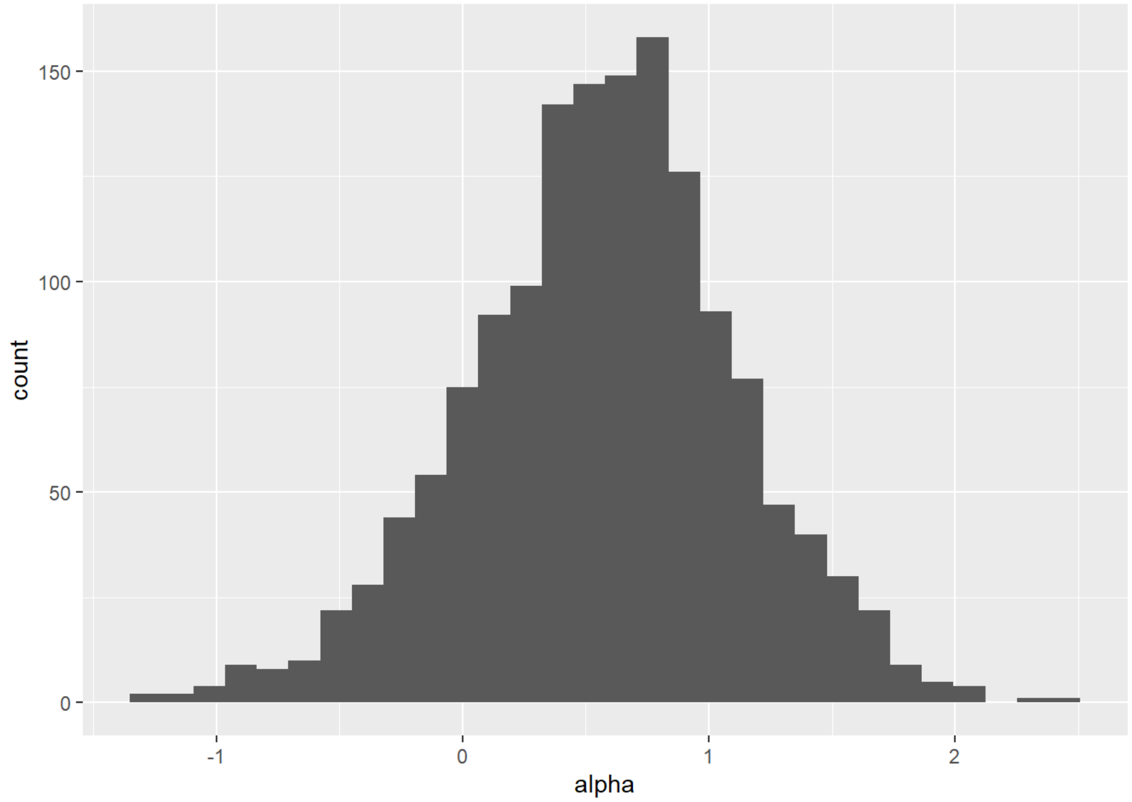
```
##      2.5%     97.5%
## -2.271038  2.939878
```

The posterior means of α and β are 0.6982 and 0.3537 respectively and 95% credible intervals for α and β are $[-3.34, 3.04]$ and $[-2.27, 2.93]$ respectively. The results are surprising because we assume flat prior is “noninformative”, but it’s quite informative and pull the posterior towards extreme and unlikely values that can bias towards inference.

Exercise 2: Compute the posterior means of α and β. Give 95% credible intervals for each. How does the posterior inference under this N(0,1) prior compare to the diffuse priors above? How informative is this weakly informative prior?

```
## Reasonable scale
# Light-tailed
stan_dat <- list(y = y, x=x, N=N)
fit.norm <- stan(file = "lab-04-normal_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 2000, warmup
 = 500, seed=49)
alpha.norm<- as.matrix(fit.norm, pars = c("alpha"))
beta.norm <- as.matrix(fit.norm, pars = c("beta"))

ggplot(alpha.norm %>% as.data.frame, aes(x = alpha)) +
  geom_histogram(bins = 30)
```



```
print(fit.norm, pars = c("alpha"))
```

```
## Inference for Stan model: lab-04-normal_prior.
## 1 chains, each with iter=2000; warmup=500; thin=1;
## post-warmup draws per chain=1500, total post-warmup draws=1500.
##
##       mean se_mean   sd  2.5%  25%  50% 75% 97.5% n_eff Rhat
## alpha 0.57    0.02 0.55 -0.55 0.23 0.59 0.9  1.64   499    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 10 23:41:07 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).
```

```
mean(alpha.norm) #0.569
```

```
## [1] 0.5688131
```

```
mean(beta.norm) #0.397
```

```
## [1] 0.396668
```

```
quantile(alpha.norm, probs = c(0.025, 0.975)) #-0.548, 1.637
```

```
##       2.5%      97.5%
## -0.5482955  1.6374156
```

```
quantile(beta.norm, probs = c(0.025, 0.975)) #-0.426, 1.251
```

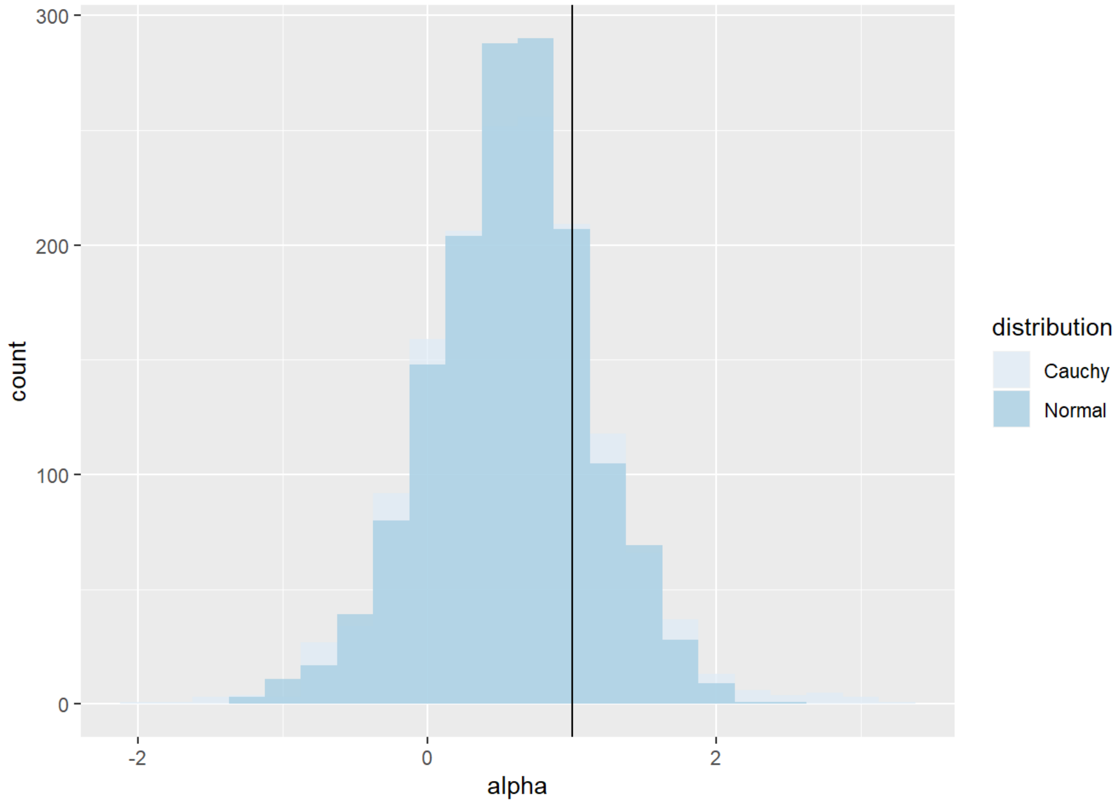
```
##       2.5%      97.5%
## -0.4265907  1.2518907
```

The posterior means of α and β are 0.569 and 0.397 respectively and 95% credible intervals for α and β are $[-0.548, 1.637]$ and $[-0.426, 1.251]$ respectively. Given that only 5 datapoints are simulated, the posterior is strongly affected by the weakly informative priors. However, since these priors were chosen to be coherent with prior information, a prior-dominated posterior yields reasonable inferences and performs much better than diffuse prior. Weakly informative prior is more informative than flat prior.

```
# heavier-tailed
stan_dat <- list(y = y, x=x, N=N)
fit.cauchy <- stan(file = "lab-04-cauchy_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 2000, warmup
= 500, seed=55)
alpha.cauchy<- as.matrix(fit.cauchy, pars = c("alpha"))

plot_dat <- create_df(alpha.norm, alpha.cauchy) %>%
  mutate(distribution = if_else(distribution == "posterior", "Normal","Cauchy"))

ggplot(plot_dat, aes(alpha, fill = distribution)) +
  geom_histogram(binwidth = 0.25, alpha = 0.7, position = "identity")+
  geom_vline(xintercept = alpha) +
  scale_fill_brewer()
```



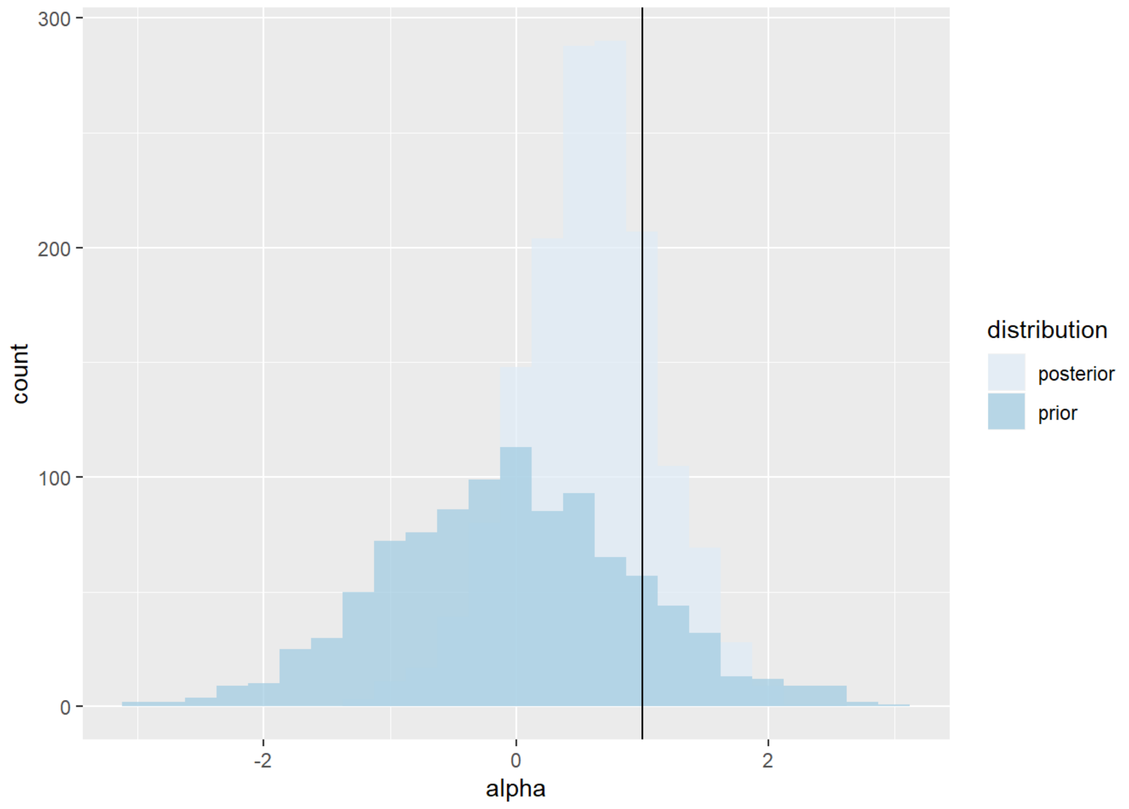
Exercise 3: Would you say that a Cauchy prior is more or less informative than a Normal prior (assume that their inter-quartile ranges are comparable)?

I think a Cauchy prior is less informative than a Normal prior because Cuachy is more diffused and has heavier tails than Normal, while the shape of Normal is more concentrated

```
## 5. Normal prior: α~N(0,1).
stan_dat <- list(y = y, x=x, N=N)
fit.norm <- stan(file = "lab-04-normal_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 2000, warmup
 = 500, seed=49)
alpha.norm<- as.matrix(fit.norm, pars = c("alpha"))

prior_draws <- rnorm(1000, 0, 1)
plot_dat <- create_df(alpha.norm, prior_draws)

ggplot(plot_dat, aes(alpha, fill = distribution)) +
  geom_histogram(binwidth = 0.25, alpha = 0.7, position = "identity")+
  geom_vline(xintercept = alpha) +
  scale_fill_brewer()
```

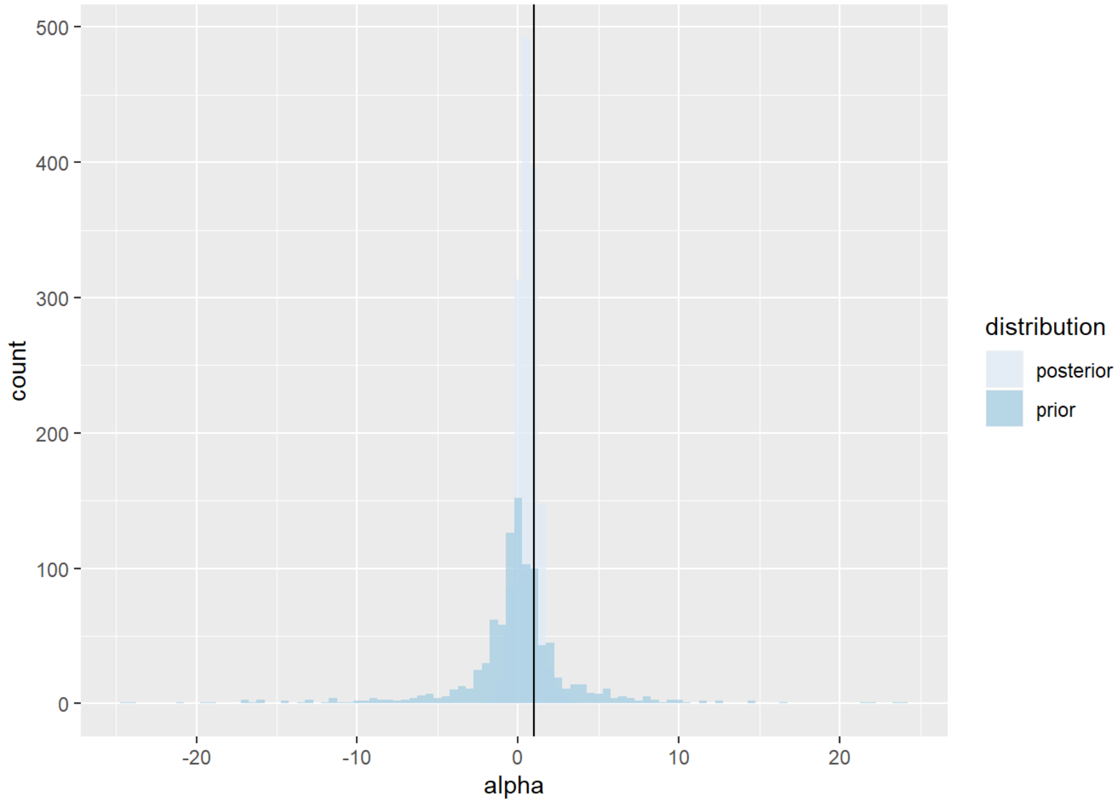


```
# how the prior is dominating the likelihood. The posterior is extremely sensitive to
# the choice of our prior, so much so that we don't observe posterior values close to the true alpha at all.
# Instead, posterior is concentrated around the upper extremes of the prior

## 6. Cauchy prior
stan_dat <- list(y = y, x=x, N=N)
fit.cauchy <- stan(file = "lab-04-cauchy_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 2000, warmup
= 500, seed=55)
alpha.cauchy<- as.matrix(fit.cauchy, pars = c("alpha"))

prior_draws <- rcauchy(1000, 0, 1)
prior_draws <- prior_draws[abs(prior_draws) < 25]
plot_dat <- create_df(alpha.cauchy, prior_draws)

ggplot(plot_dat, aes(alpha, fill = distribution)) +
  geom_histogram(binwidth = .5, alpha = 0.7, position = "identity")+
  geom_vline(xintercept = alpha) +
  scale_fill_brewer()
```



Exercise 4: What happens as we increase the number of observations?

As we increase the number of observations, data (likelihood) has more and more influence compared to prior. Therefore, likelihood could dominate posterior as observations are large enough.

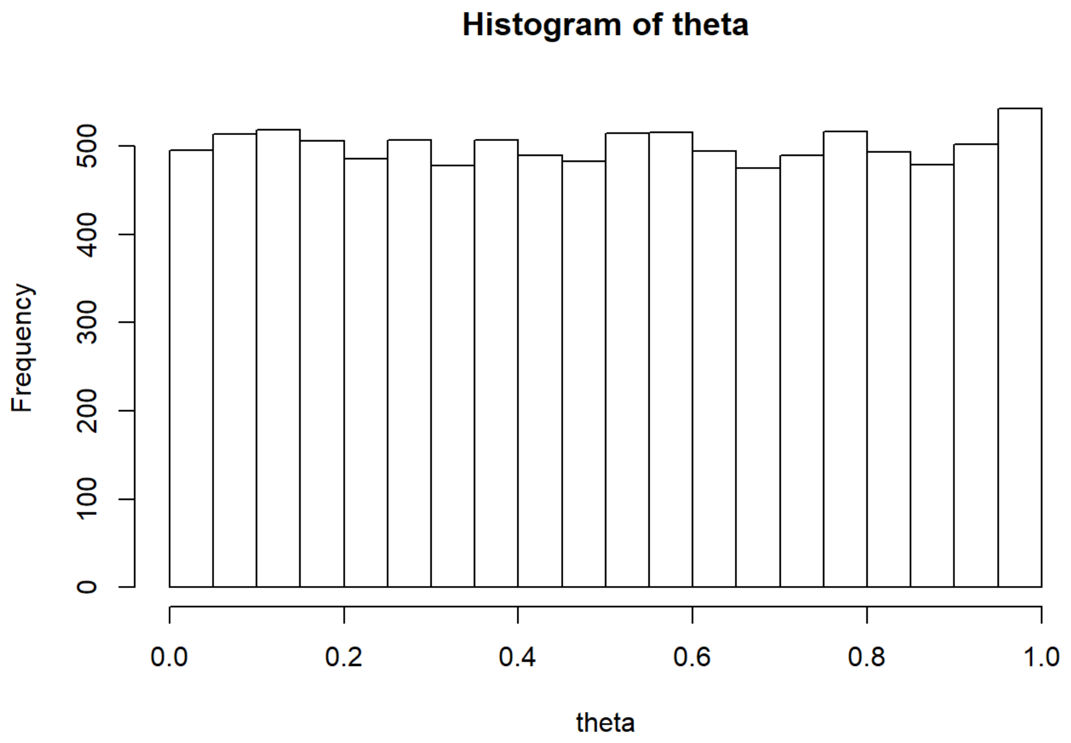
Exercise 5: When might we prefer to use lighter- versus heavier-tailed priors?

We might perfer to use lighter-tailed priors when the scale is well-chosen since it can be coherent with the data value. Also, lighter-tailed priors penalizes parameter values above certain scales, preventing the posterior from assigning any nontrivial probability mass to the more extreme values favoured by the likelihood.

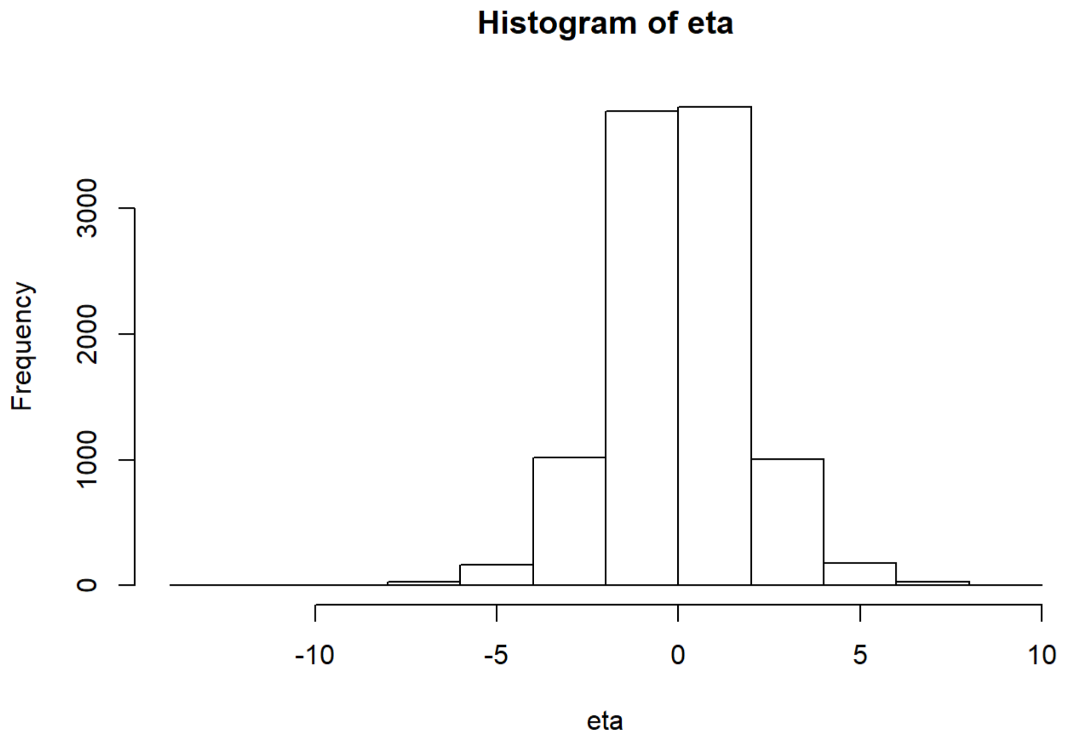
Exercise 6: How might we determine a reasonable scale for the prior?

If we have historical/prior knowledge about certain data, we could choose prior around scale of known data range. However, if we’re completely unsure about the data, it might not be possible to determine a reasonable scale.

```
## Model Reparameterization
theta <- runif(10000,0,1)
hist(theta)
```



```
logit <- function(x){
  ret <- log(x/(1-x))# finish function for log odds
  return(ret)
}
eta <- logit(theta)
hist(eta)
```



```
# pretend the true birth rate of females in Paris was 0.3, simulate N=10 obs
set.seed(123);
theta <- 0.3;
N <- 10;
y <- rbinom(N, 1, theta)

theta.mle <- sum(y)/N

stan_dat <- list(y = y,N=N)
fit.bayes.prob <- stan(file = "lab-04-prob.stan", data = stan_dat, refresh = 0, iter = 2000)
print(fit.bayes.prob, pars = c("theta", "eta"))
```

```
## Inference for Stan model: lab-04-prob.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##        mean se_mean   sd  2.5%   25%   50%  75% 97.5% n_eff Rhat
## theta  0.41    0.00 0.14  0.17  0.31  0.41 0.51  0.68  1468    1
## eta   -0.38    0.02 0.61 -1.58 -0.79 -0.36 0.05  0.75  1477    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 10 23:43:38 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).
```

Exercise 7: What would we expect to be the posterior mode of our samples? Calculate the posterior mode theoretically, and compare it to the estimated mode from the posterior samples.

Since $sum(y) = 4$ and $N = 10$, we get a posterior distribution $p(\theta|y) = Beta(\theta|5, 7)$, and we know that if $\theta\ Beta(5, 7)$, then θ has a mode of

$$\theta = \frac{a + \sum y_i - 1}{a + b + N - 2} = \frac{1 + 4 - 1}{1 + 1 + 10 - 2} = \frac{4}{10} = 0.4$$

We can calculate the posterior mode given the names of the data variables.

```
model.prob <- stan_model("lab-04-prob.stan")
fit.pmode.prob <- optimizing(model.prob, data=c("N", "y"))
fit.pmode.prob$par
```

```
##      theta        eta
##  0.4000013 -0.4054598
```

The calculated posterior mode is the same as estimated mode from the posterior.

Exercise 8: Is this prior proper? Does it result in a proper posterior? If so, under which conditions?

Since $\theta \sim Beta(0, 0)$, we derive the posterior as following

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{-1}(1-\theta)^{-1} \\ &= \theta^{y-1}(1-\theta)^{n-y-1} \\ &\propto Beta(y, n-y) \end{aligned}$$

If $y > 0$ and $n - y > 0$, then this would result in a proper posterior since it has the kernel of beta distribution. Otherwise, the posterior doesn't satisfy the requirement of beta distribution, thus resulting in improper posterior.

Exercise 9: If we set a uniform prior for π, what is the induced prior on $\theta = \frac{\pi}{(1+\pi)}$? Is this prior proper?

Since $f_y(y) = f_x(g^{-1}(y))|\frac{d}{dy}(g^{-1}(y)|$ and $f_\pi(\pi) = 1$ (uniform distribution), therefore

$$\begin{aligned} f_\theta(\theta) &= f_\pi(\frac{\theta}{1-\theta})|\frac{d}{d\theta}\frac{\theta}{1-\theta}| \\ &= |1\frac{1-\theta+\theta}{(1-\theta)^2}| \\ &= \frac{1}{(1-\theta)^2} \end{aligned}$$

$\int_0^1 f_\theta(\theta)d\theta = \int_0^1 \frac{1}{(1-\theta)^2}d\theta = \infty$, therfore, this prior is improper.

Exercise 10: Derive the Fisher Information of the Binomial sampling model under the success probability parameterization. From this expression, state the Jeffreys prior for θ. Write your own Stan file called jeffreys.stan to produce samples from the model under your derived Jeffreys prior. Use the existing .stan files as a guide. Use the code chunks above as a guideline to run Stan from R and print the results obtained from your sampler.

Since for likelihood,

$$\begin{aligned} p(y|\theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\ log(p(y|\theta)) &= log\binom{n}{y} + ylog(\theta) + (n-y)log(1-\theta) \\ \frac{\partial}{\partial\theta}logp(y|\theta) &= \frac{y}{\theta} - (n-y)(1-\theta)^{-1} \\ \frac{\partial}{\partial^2\theta}logp(y|\theta) &= -y\theta^{-2} - (n-y)(1-\theta)^{-2} \\ \text{Since } E(y) &= n\theta \\ E(\frac{\partial}{\partial^2\theta}logp(y|\theta)) &= \frac{-n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = -n\theta^{-1}(1-\theta)^{-1} \\ -E(\frac{\partial}{\partial^2\theta}logp(y|\theta)) &= n\theta^{-1}(1-\theta)^{-1} \end{aligned}$$

$$\begin{aligned} p(\theta) &\propto \sqrt{I(\theta)} \\ &\propto \sqrt{n}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\ &\propto Beta(\frac{1}{2},\frac{1}{2}) \end{aligned}$$

```
fit.jeffreys <- stan(file = "jeffreys.stan", data = stan_dat, refresh = 0, iter = 1000)
print(fit.jeffreys, pars = c("theta"))
```

```
## Inference for Stan model: jeffreys.
## 4 chains, each with iter=1000; warmup=500; thin=1; 
## post-warmup draws per chain=500, total post-warmup draws=2000.
## 
##       mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## theta 0.05       0 0.02 0.01 0.03 0.04 0.06  0.09   706    1
## 
## Samples were drawn using NUTS(diag_e) at Mon Feb 10 23:44:56 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).
```