STA 360 Lab 6: Prior selection and model reparameterization

Isaac Fan

12 March, 2021

Preliminaries

Please submit a pdf copy of this lab on Sakai by Friday, March 12th at 11:59 PM. Exercises 3 and 5 will be graded for completion. You do not need to do the bonus question for exercise 5, but you should (it's a good exercise).

The T-Distribution

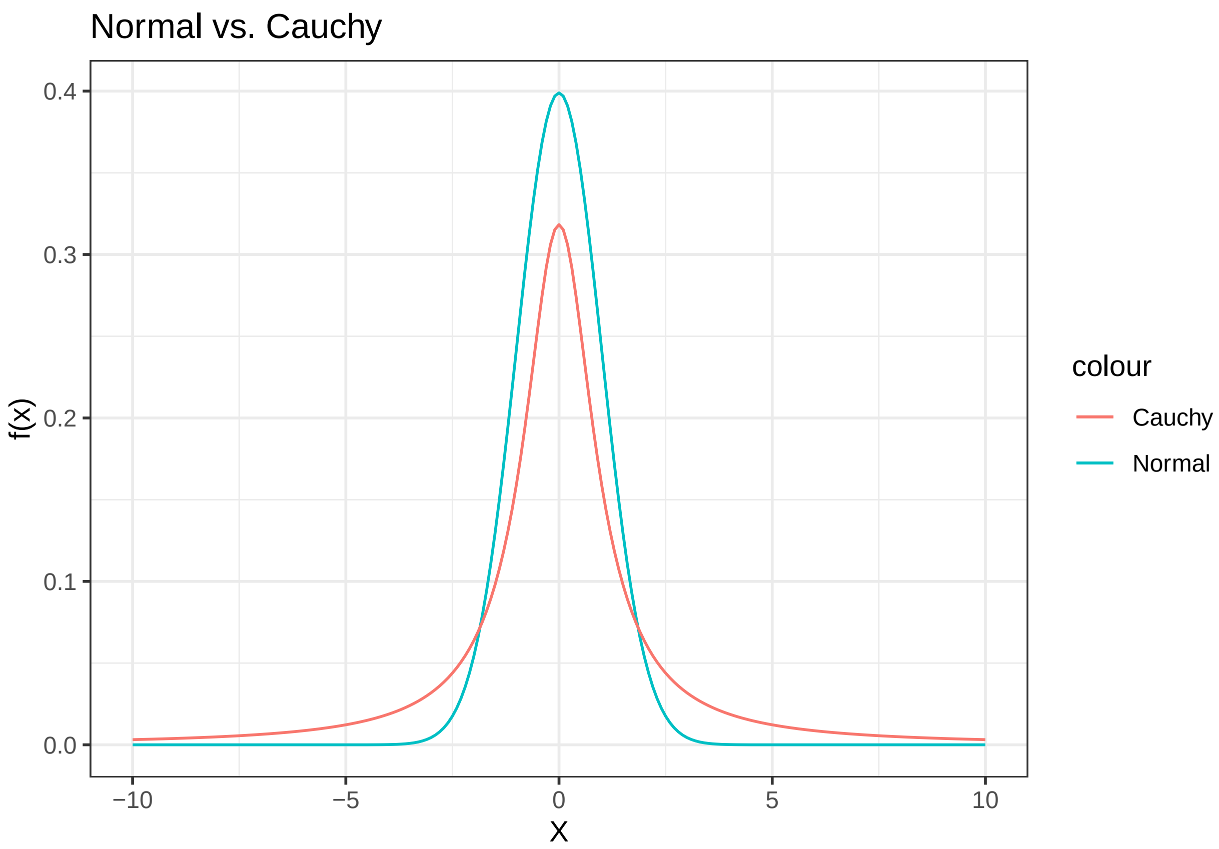
The t-distribution with parameters $k > 0$, $\mu \in \mathbb{R}$, and $\sigma > 0$ has density function

$$f(t) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{k\pi}\sigma}\left(1 + \frac{1}{k}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-(k+1)/2}$$

for all $t \in \mathbb{R}$ and denoted $t_k(\mu, \sigma)$.

- The t-distribution has support $\mathbb{R}$, like the normal distribution.
- It has heavier tails than a normal distribution.
- Moments (ie, integrals $E[T^m]$) are not defined (infinite) for $m \geq k$.
- $k = 1$ is a special case, called the Cauchy distribution. The Cauchy distribution has infinite mean and variance.
- The t-distribution arises from a normal-inverse gamma model (HW 4, exercise 1).

```
grid <- seq(-10, 10, by = 0.1)
dens <- data.frame(X = grid, "Normal" = dnorm(grid, 0, 1), "Cauchy" = dcauchy(grid, 0, 1))

ggplot(dens, aes(x = X)) + geom_line(aes(x = X, y = Normal, color = "Normal")) + geom_line(aes(x = X, y
```



Exercise 1 (Not for completion)

For what $k > 0$ does a t-distribution have finite variance?

k > 2.

$$Var[T] = E[T^2] - (E[T])^2$$

Prior selection

Suppose we have data Y, which we take to be Normally distributed: $y_i \sim N(\mu, \sigma^2)$. Suppose further that we want to model μ as a linear function of a fixed covariate x using an intercept α and coefficient β:

$$\mu = \alpha + \beta x$$

We call this model a *simple linear regression model.*

```
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)
```

Both α and β are unknown, so we would like to perform inference on them. However, we only have 5 data points. Given so few data points, we can be quite sure that the resulting posteriors will be sensitive to the choice of priors.

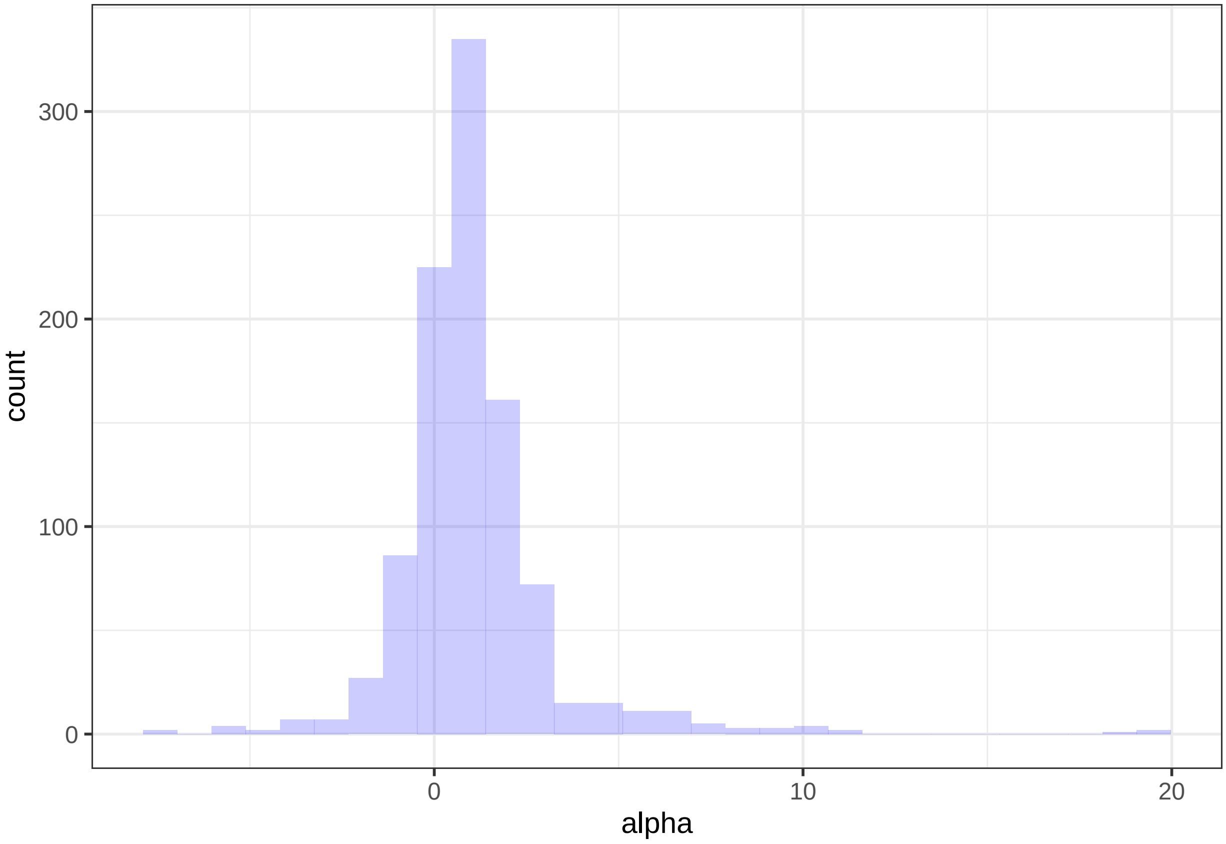
Flat priors

One possible choice of prior is a flat prior for α and β. That is, $p(\alpha) \propto 1$ and $p(\beta) \propto 1$. Let's look at how our posterior beliefs about α and β act under these priors.

```
stan_dat <- list(y = y, x=x, N=N)
fit.flat <- stan(file = "flat_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 1100, warmu
```

```
## Warning in readLines(file, warn = TRUE): incomplete final line found on 'C:
## \Users\isaac\OneDrive\Documents\STA360 Lab 6\flat_prior.stan'
```

```
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, alpha = 0.2, fill = "blue")
```



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

```
## Inference for Stan model: flat_prior.
## 1 chains, each with iter=1100; warmup=100; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=1000.
##
##       mean se_mean   sd  2.5%  25%  50%  75% 97.5% n_eff Rhat
## alpha 1.05    0.18 2.15 -2.21 0.09 0.78 1.63  6.48   141 1.01
##
## Samples were drawn using NUTS(diag_e) at Fri Mar 12 21:40:50 2021.
## 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).
```

Notice how the posterior for α is quite diffuse–there is much uncertainty about what α is. While the true value for α is 1, what is the posterior mean? What is the 95% credible interval?

Exercise 2 (Not for completion)

Compute 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?

```
quantile(alpha.flat, c(.025, .975))
```

```
##      2.5%     97.5%
## -2.207415  6.478666
```

```
mean(alpha.flat)
```

```
## [1] 1.052244
```

```
quantile(beta.flat, c(.025, .975))
```

```
##     2.5%    97.5%
## -3.33850  3.24487
```

```
mean(beta.flat)
```

```
## [1] 0.1901685
```

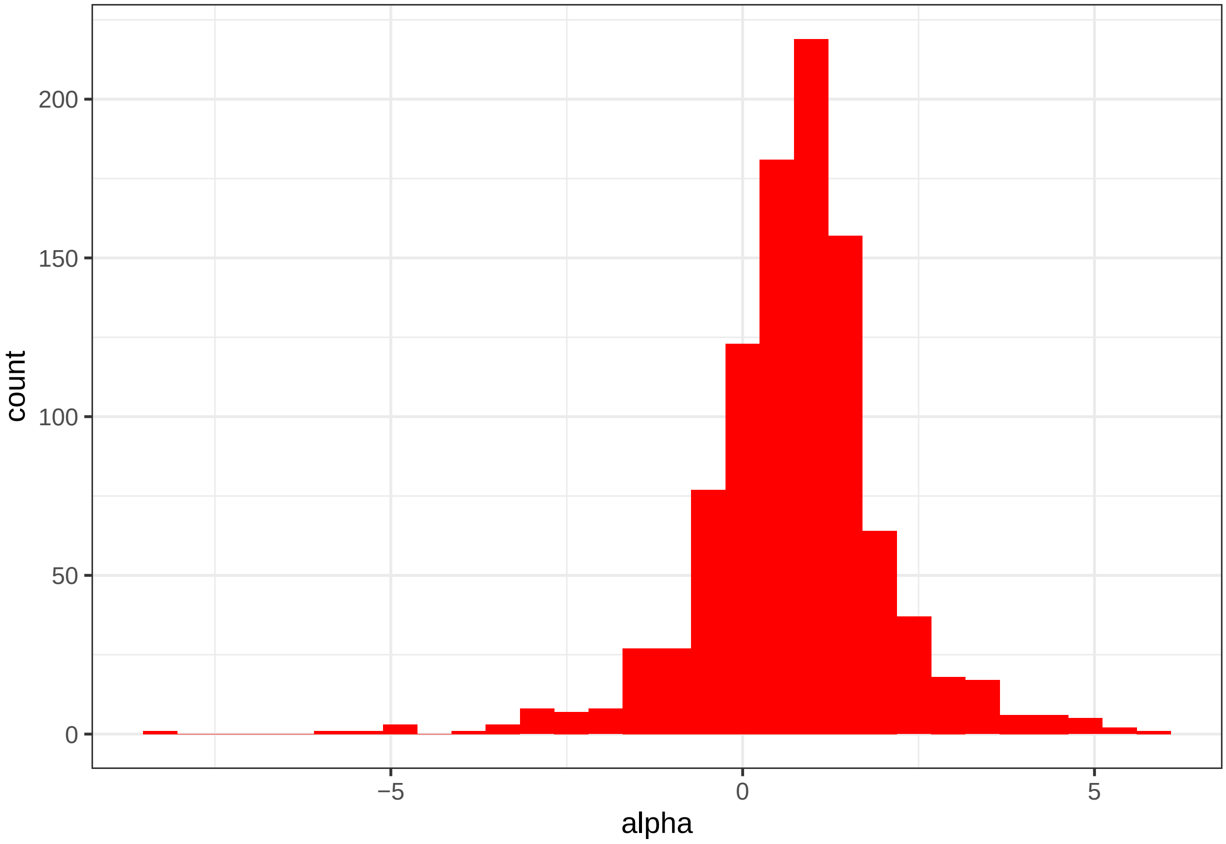
By doing inference with a flat/diffuse prior, we might have thought we were using the least prior information possible. However, flat priors may actually bias our estimates for a parameter by allowing the posterior to be pulled towards extreme and unlikely values, as evidenced above.

Diving a bit deeper, consider another flat prior: $\alpha \sim Unif(a, b)$. Under this prior, we are saying that we believe $a \leq \alpha \leq b$. We exhibit this in the following code, with the prior $\alpha \sim Unif(-10, 10)$

```r
stan_dat <- list(y = y, x=x, N=N, lb = -10, ub = 10)
fit.unif <- stan(file = "unif_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 1100, warmu
```

```
## Warning in readLines(file, warn = TRUE): incomplete final line found on 'C:
## \Users\isaac\OneDrive\Documents\STA360 Lab 6\unif_prior.stan'

## Warning: There were 44 divergent transitions after warmup. See
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.

## Warning: Examine the pairs() plot to diagnose sampling problems

## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess

## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quant
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

```r
alpha.unif <- as.matrix(fit.unif, pars = c("alpha"))
beta.unif <- as.matrix(fit.unif, pars = c("beta"))

ggplot(alpha.unif %>% as.data.frame, aes(x = alpha)) +
  geom_histogram(bins = 30, alpha = 02, fill = "red")
```



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

```
## Inference for Stan model: unif_prior.
## 1 chains, each with iter=1100; warmup=100; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=1000.
##
##       mean se_mean   sd  2.5%  25% 50%  75% 97.5% n_eff Rhat
## alpha 0.74     0.1 1.32 -2.14 0.15 0.8 1.38   3.6   187 1.01
##
## Samples were drawn using NUTS(diag_e) at Fri Mar 12 21:42:06 2021.
## 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).
```

While the posterior mean under this Uniform prior is closer to the true value, the posterior is still very spread out. So we have seen that a diffuse or flat prior is *not* necessarily non-informative, and in these cases is actually extremely informative! Diffuse priors inherently spread probability mass across large regions of parameter space. We often assume that the data will overwhelm the prior, so a diffuse prior will let the data dominate posterior inference. However as showcased here, having only a small amount of observed data may allow the diffuse prior to become informative. Therefore, it would be wise to make the conscious choice to have an informative prior. However, we have some leeway with how informative we want our priors to be.

Weakly informative priors

We often must consider the scale of the parameters we wish to estimate. In applied problems where we know how to interpret the parameters, the scale is easier to identify. We may consider a weakly informative prior to be such that if there is a reasonably large amount of data, the likelihood will dominate the posterior and the prior is not important. This sort of prior ought to rule out unreasonable parameter values, but is not so strong as to rule out possible values which might make sense. As a general rule, it is wise to not use hard constraints unless the bounds represent true constraints (ex. bounding a prior for a variance parameter below by 0). As an example, we might think that α could be between 0 and 1. Instead of setting the prior $\alpha \sim Unif(0, 1)$, it would be wise to use a Normal(0.5, 1) prior instead. In the following, we consider some weakly informative priors for the parameters. The data we have are on unit scale, so we consider priors also on the unit scale.

Light-tailed

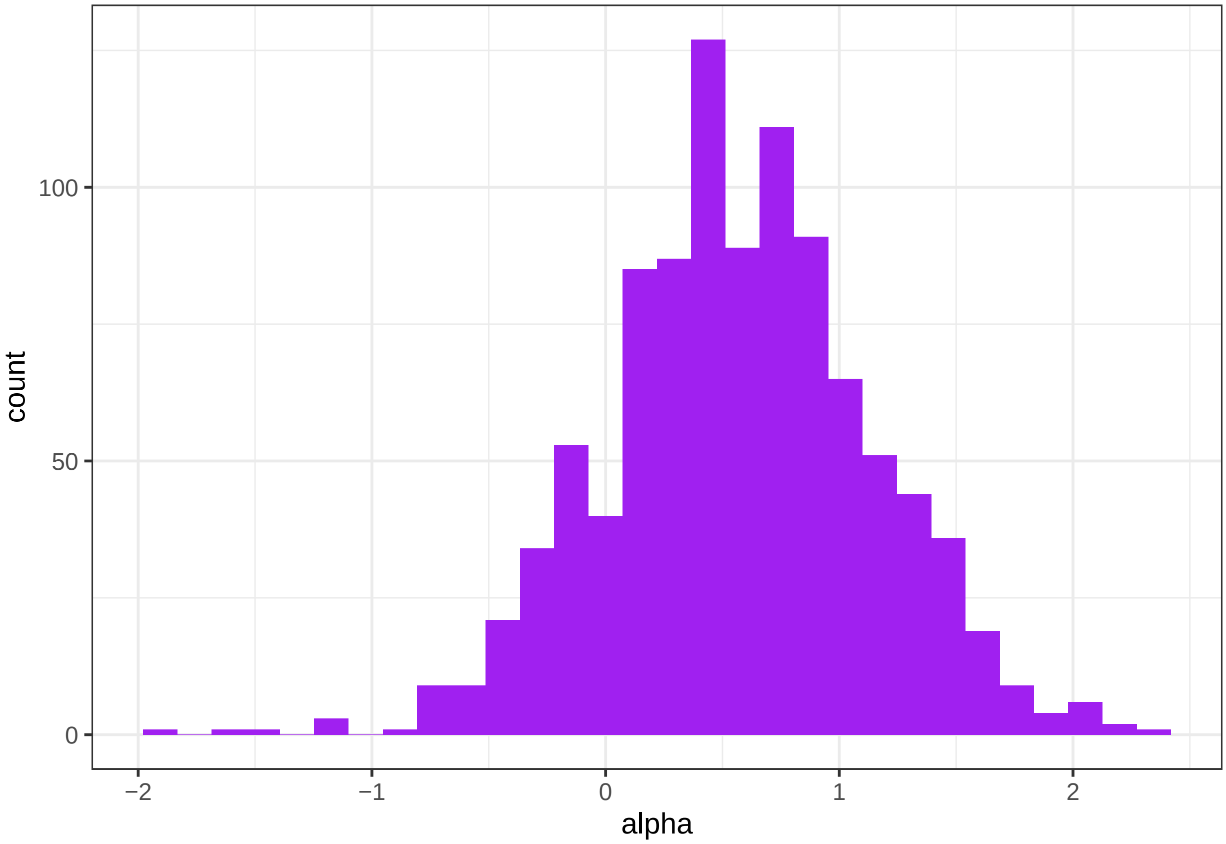
In this section, we fit the model under the priors

$$\alpha, \beta \overset{iid}{\sim} \mathcal{N}(0, 1).$$

```
stan_dat <- list(y = y, x=x, N=N)
fit.norm <- stan(file = "normal_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 1100, war
```

```
## Warning in readLines(file, warn = TRUE): incomplete final line found on 'C:
## \Users\isaac\OneDrive\Documents\STA360 Lab 6\normal_prior.stan'
```

```r
alpha.norm<- as.matrix(fit.norm, pars = c("alpha"))
beta.norm <- as.matrix(fit.norm, pars = "beta")

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



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

```
## Inference for Stan model: normal_prior.
## 1 chains, each with iter=1100; warmup=100; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=1000.
##
##       mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## alpha 0.57    0.03 0.57 -0.5 0.21 0.56 0.93  1.66   398    1
##
## Samples were drawn using NUTS(diag_e) at Fri Mar 12 21:43:29 2021.
## 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 3 (For completion)

Compute the posterior means of α and β. Give 95% credible intervals for each. How does the posterior inference under this $\mathcal{N}(0,1)$ prior compare to the diffuse priors above?

```
quantile(alpha.norm, c(.025, .975))
```

```
##       2.5%      97.5%
## -0.4994739  1.6639191
```

```
mean(alpha.norm)
```

```
## [1] 0.5677155
```

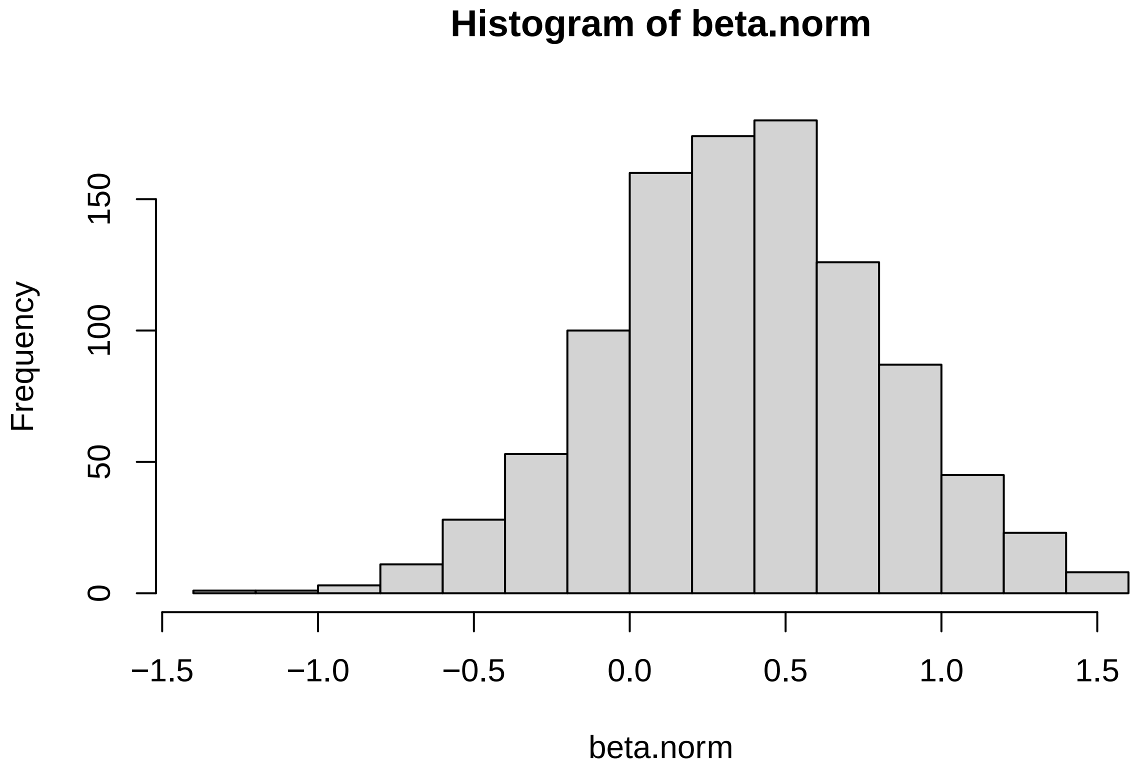
```
quantile(beta.norm, c(.025, .975))
```

```
##       2.5%      97.5%
## -0.4981612  1.2539506
```

```
mean(beta.norm)
```

```
## [1] 0.3661443
```

```
hist(beta.norm)
```

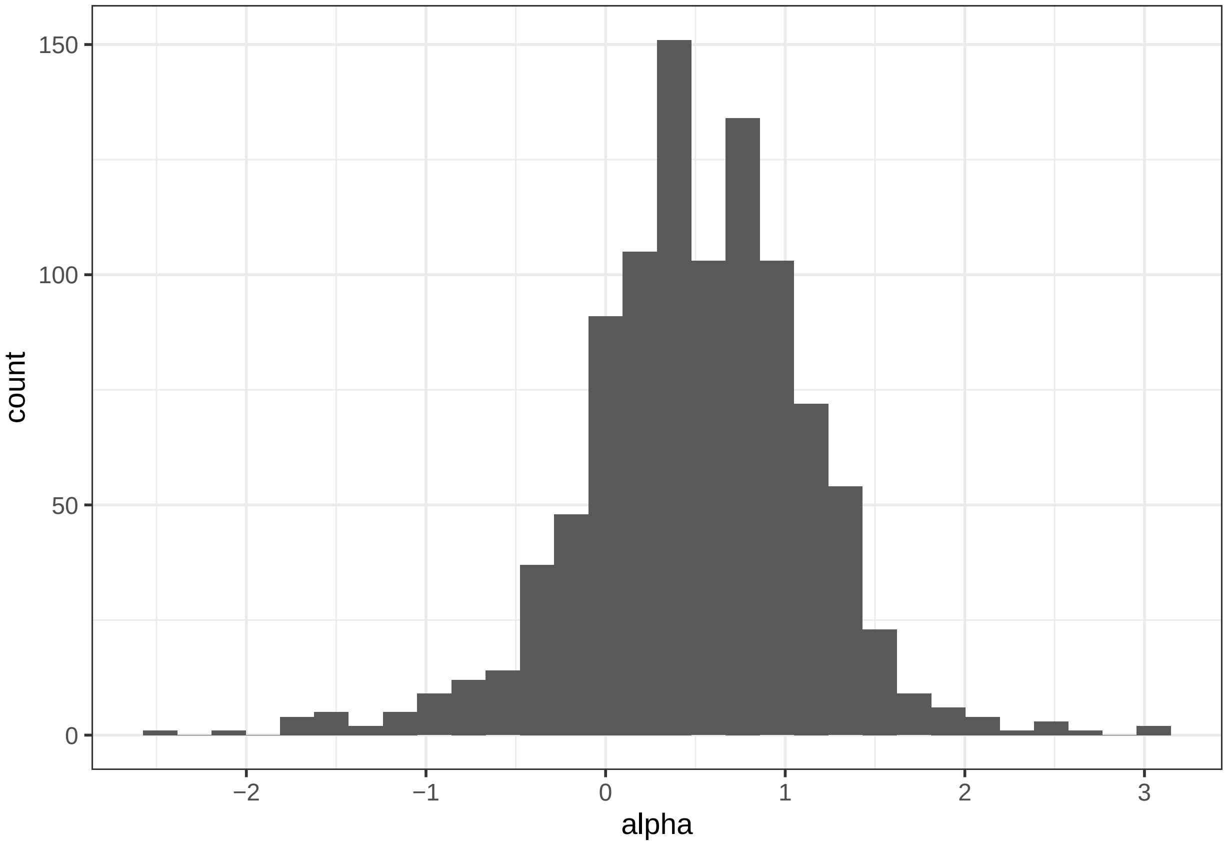


The approximation is much closer to the true values of alpha and beta, and the credible intervals are much narrower.

Heavy-tailed

Now, we will set $\alpha, \beta \overset{iid}{\sim} \text{Cauchy}(0, 1)$. Recall that the Cauchy has heavier/fatter tails than the Normal distribution.

```
stan_dat <- list(y = y, x=x, N=N)
fit.cauchy <- stan(file = "cauchy_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 1100, wa

alpha.cauchy<- as.matrix(fit.cauchy, pars = c("alpha"))

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



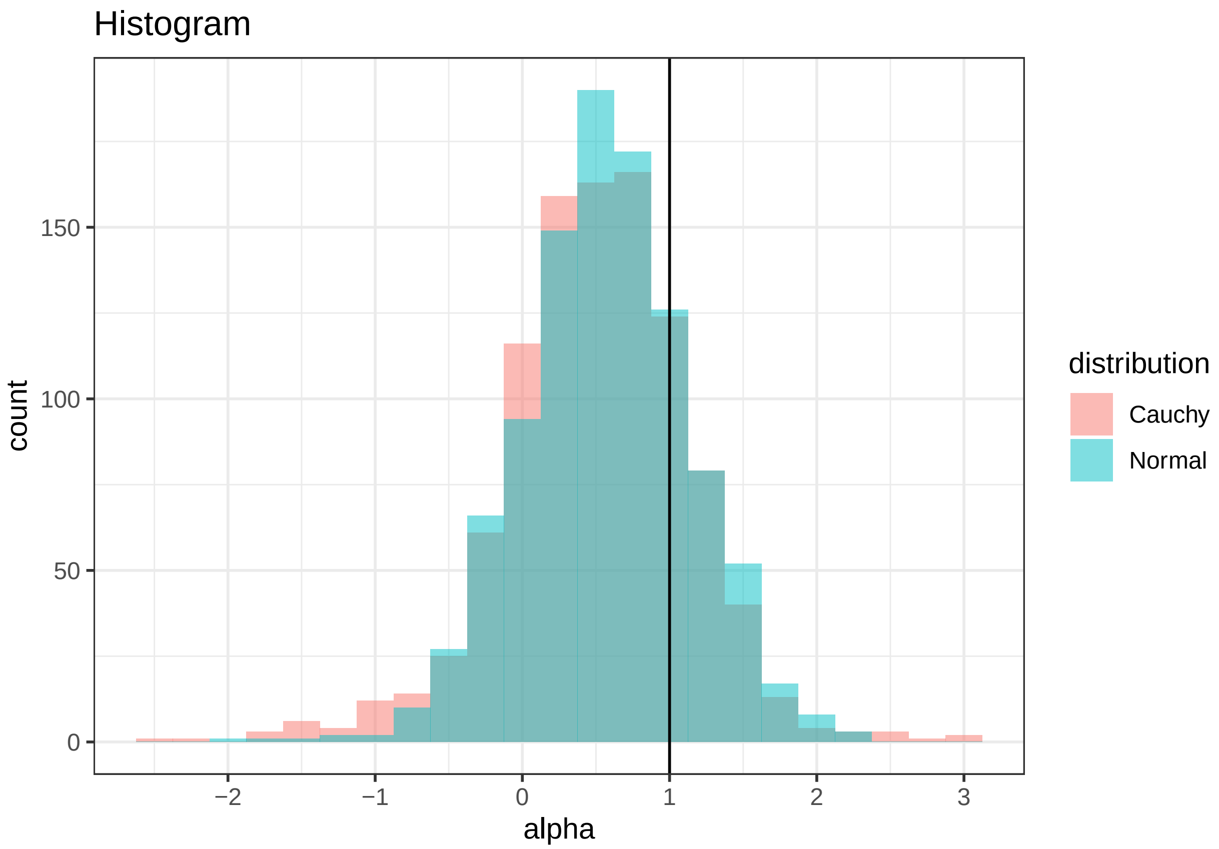
```
print(fit.cauchy, pars = c("alpha"))
```

```
## Inference for Stan model: cauchy_prior.
## 1 chains, each with iter=1100; warmup=100; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=1000.
```

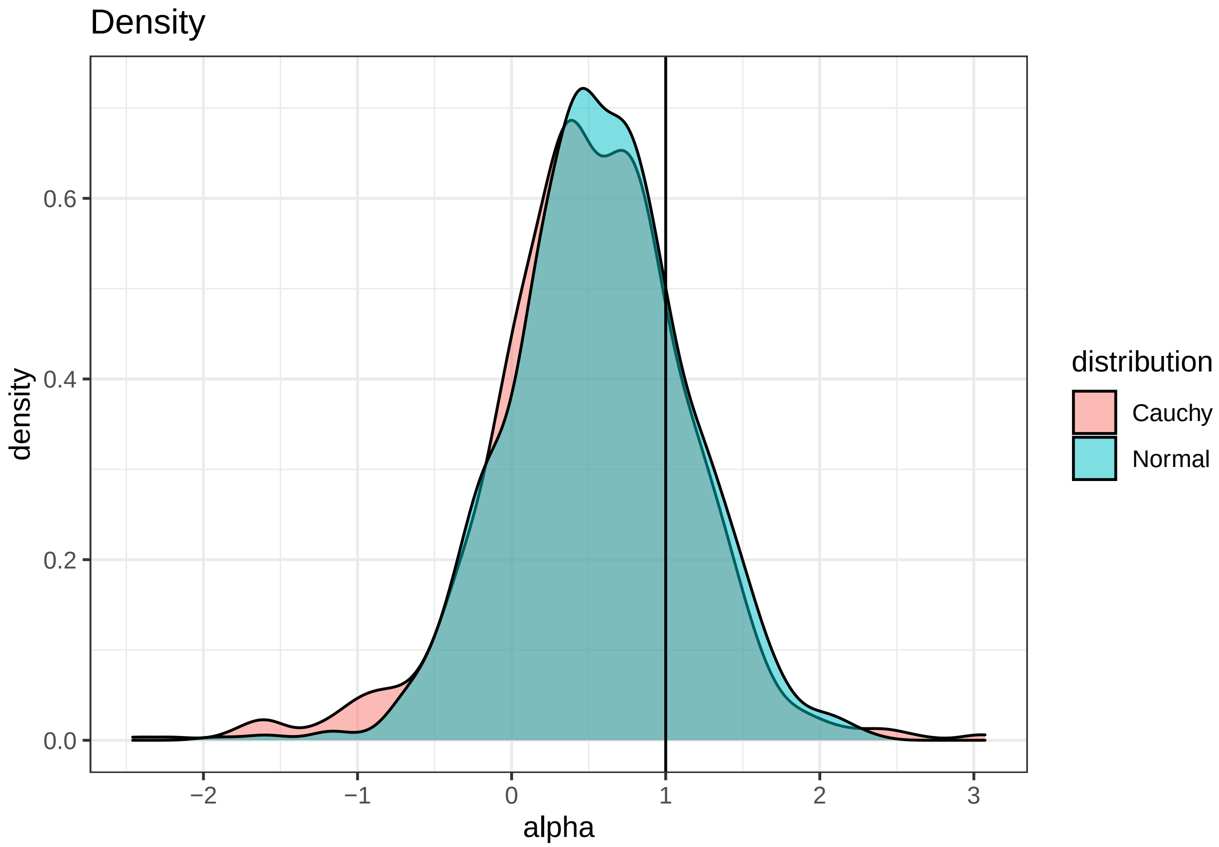
```
##
##       mean se_mean   sd  2.5%  25%  50% 75% 97.5% n_eff Rhat
## alpha  0.5    0.04 0.64 -0.89 0.14 0.51 0.9  1.63   251    1
##
## Samples were drawn using NUTS(diag_e) at Fri Mar 12 21:44:38 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

The following plots display the posteriors for α under these two priors (normal and Cauchy):

```
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 = 0.5, position = "identity")+
  geom_vline(xintercept = alpha) + labs(title = "Histogram")
```



```
ggplot(plot_dat, aes(alpha, fill = distribution)) +
  geom_density(alpha = 0.5, position = "identity")+
  geom_vline(xintercept = alpha) + labs(title = "Density")
```



The Cauchy prior allocates higher probabiity mass to extreme values as compared to the Normal prior, while still concentrating most of the posterior mass for α within a desired scale.

Exercise 4 (Not for completion)

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 would say the inter-quartile range generated by a Normal prior is more informative because it gives less weight to extremes, meaning the inter-quartile range will be narrower. This will give a better idea of what the true value of beta and alpha are. ***

Sensitivity to prior selection

In the previous example, the Normal and the Cauchy priors for α performed relatively similarly. The true value of α was 1 and the priors we used were weakly centered around 1, so we happened to choose a good scale for the parameter. Let's examine what happens when this is not the case. We will simulate new data now with $\alpha = 10$ instead of 1. We will also double the number of data points to 10.

```
alpha<-10
N <- 10
```

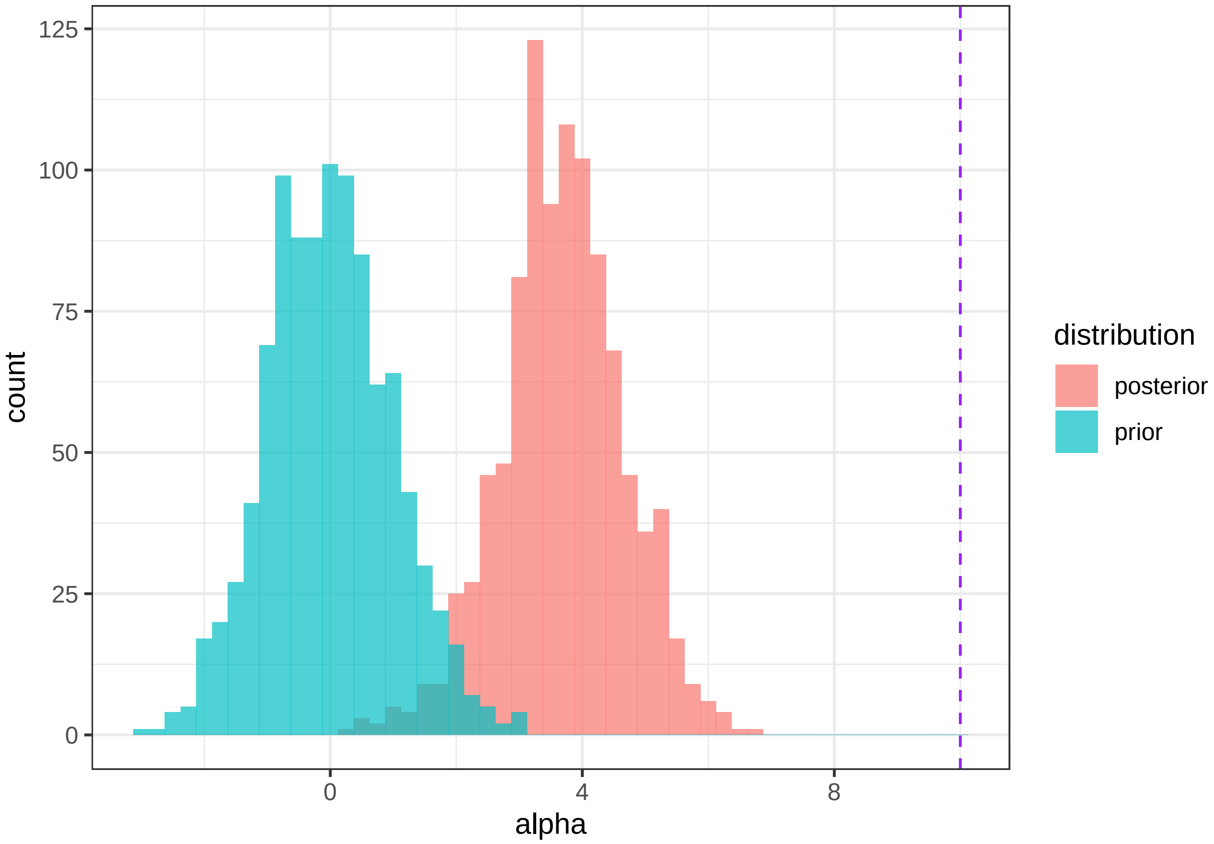
```
x <- runif(N, 0, 2)
y <- rnorm(N, beta * x + alpha, sigma)
```

Considering first the same Normal prior as above: $\alpha \sim N(0, 1)$.

```
stan_dat <- list(y = y, x=x, N=N)
fit.norm <- stan(file = "normal_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 1100, warn
```

```
## Warning in readLines(file, warn = TRUE): incomplete final line found on 'C:
## \Users\isaac\OneDrive\Documents\STA360 Lab 6\normal_prior.stan'
```

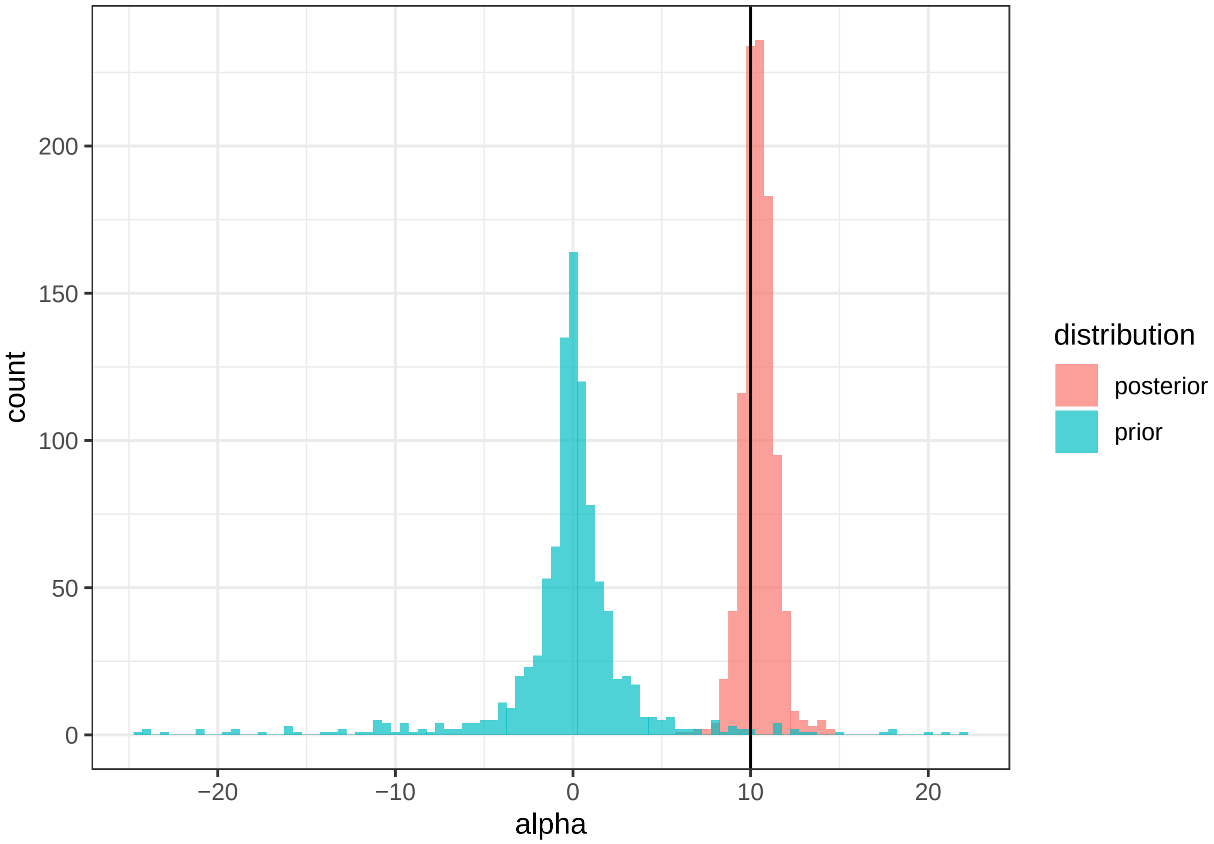
```
alpha.norm<- as.matrix(fit.norm, pars = c("alpha"))
prior_draws <- rnorm(1000, 0, 1)
plot_dat <- create_df(alpha.norm, prior_draws)

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



Here the prior is once again weakly-informative, but notice how the prior is dominating the posterior. The posterior is extremely sensitive to the choice of our prior, so much so that the we do not observe posterior values close to the true α at all. Instead, the posterior is concentrated around the upper extremes of the prior.

What if we use a heavier-tailed distribution like the Cauchy?

```
stan_dat <- list(y = y, x=x, N=N)
fit.cauchy <- stan(file = "cauchy_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 1100, wa

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)
```



Notice how under this Cauchy(0,1) prior, the posterior is able to concentrate around the true $\alpha = 10$. The heavy tails of the Cauchy allow the posterior to move beyond the scale occupied by the prior. From this histogram, it is much clearer that the prior we chose was probably inappropriate and conflicts with the data.

Exercise 5 (For completion)

Now, we will fit the model with a standard t-distribution prior with k degrees of freedom. That is,

$$\alpha, \beta \overset{iid}{\sim} t_k(0, 1).$$

For the rest of this problem, let $k = 5$.

```r
k <- 5
stan_dat <- list(y = y, x=x, N=N, k = k)
fit.t <- stan(file = "t_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 1100, warmup = 100

alpha.t<- as.matrix(fit.t, pars = c("alpha"))
beta.t<- as.matrix(fit.t, pars = c("beta"))
```

(a) How does the shape of the t_5 distribution compare to the $\mathcal{N}(0, 1)$? How about to the Cauchy distribution?

The t_5 distribution has fatter tails than both the normal and cauchy distribution.

(b) Compute credible intervals and posterior mean estimates for α and β under this model.

```r
quantile(alpha.t, c(.025, .975))
```

```
##      2.5%     97.5%
##  8.001421 11.719995
```

```r
mean(alpha.t)
```

```
## [1] 9.998621
```

```r
quantile(beta.t, c(.025, .975))
```

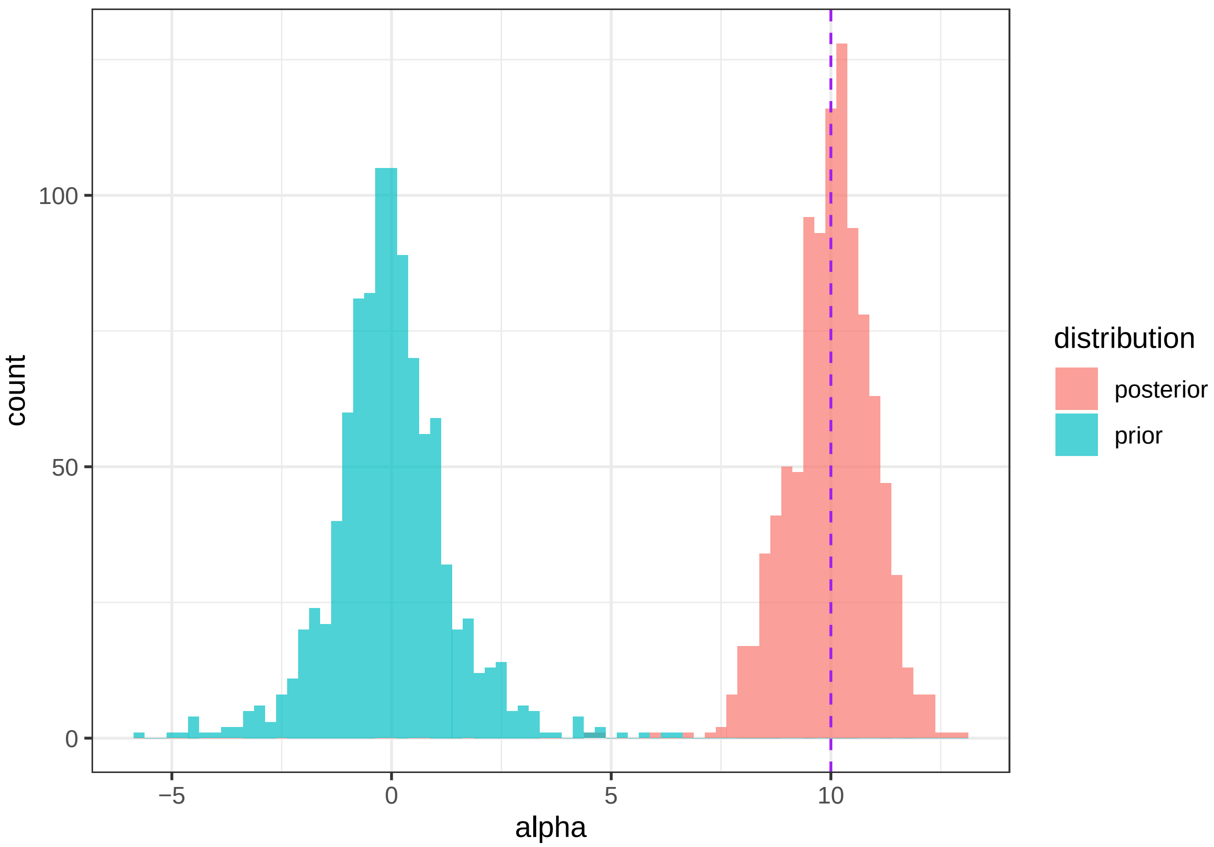
```
##      2.5%     97.5%
## -1.477689  1.447985
```

```r
mean(beta.t)
```

```
## [1] -0.03650662
```

(c) Make a plot of the posterior distribution of α overlaid on top of its prior distribution.

```r
prior_draws <- rt(1000, 5)
plot_dat <- create_df(alpha.t, prior_draws)

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



(d) Give two reasons why you might use a t-distribution prior instead of a normal prior.

It is more robust because it has fatter tails, meaning that a t-distribution allows for the posterior to be a good estimate despite a bad prior belief.

A t-distribution is also preferable when there is a small sample size because it is weaker than a normal prior. This means the prior is less likely to over power the sampled data in the given posterior.

(e) (BONUS) Under our model (but assume β, σ^2 are known), is $\alpha \sim t_k(0,1)$ a conjugate prior? Hint:

$$f(\alpha) \propto \left(1 + \frac{1}{k}\alpha^2\right)^{-(k+1)/2}.$$

---

*Adapted from this Rstan tutorial