Lab 5: Prior selection and model reparameterization

Due: 11:59 pm on Friday, March 6

Turning in solutions

This lab is part of Homework 4. Solutions to the exercises, as well as the non-lab homework exercises are to be written up and uploaded to Gradescope as a PDF.

Getting started

You will need the following R packages. If you do not already have them installed, please do so first using the `install.packages` function.

```
require(rstan)
require(tidyverse)
require(rstanarm)
require(magrittr)
```

For this lab, you will need a few different stan files. Download all of them here:

- https://omelikechi.github.io/sta402spring26/labs/lab05/lab-05-cauchy_prior.stan;
- https://omelikechi.github.io/sta402spring26/labs/lab05/lab-05-flat_prior.stan;
- https://omelikechi.github.io/sta402spring26/labs/lab05/lab-05-log_odds.stan;
- https://omelikechi.github.io/sta402spring26/labs/lab05/lab-05-normal_prior.stan;
- https://omelikechi.github.io/sta402spring26/labs/lab05/lab-05-prob.stan; and
- https://omelikechi.github.io/sta402spring26/labs/lab05/lab-05-unif_prior.stan.

Download and make sure to save them in the same folder as the R script or R markdown file you are working from.

Prior selection

Suppose we have data Y, which we take to be normally distributed: $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$. Suppose further that we want to model μ_i as a linear function of a covariate/predictor x_i, using an intercept α and coefficient β. That is,

$$\mu_i = \alpha + \beta x_i.$$

Let's simulate data (with very small sample size) under this model.

```
create_df <- function(post_draws, prior_draws){
  post_draws <- data.frame(post_draws)
  post_draws$distribution <- "posterior"

  prior_draws <- data.frame(prior_draws)
  colnames(prior_draws) <- "alpha"
  prior_draws$distribution <- "prior"

  dat <- rbind(post_draws, prior_draws)
```

```
  return(dat)
}
```

```
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)
```

Note the true values of α and β. Once we observe the data $\{x_i, y_i\}$, 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, $\pi(\alpha) \propto 1$ and $\pi(\beta) \propto 1$. Let's look at how our posterior beliefs about α and β (especially α) act under these priors.

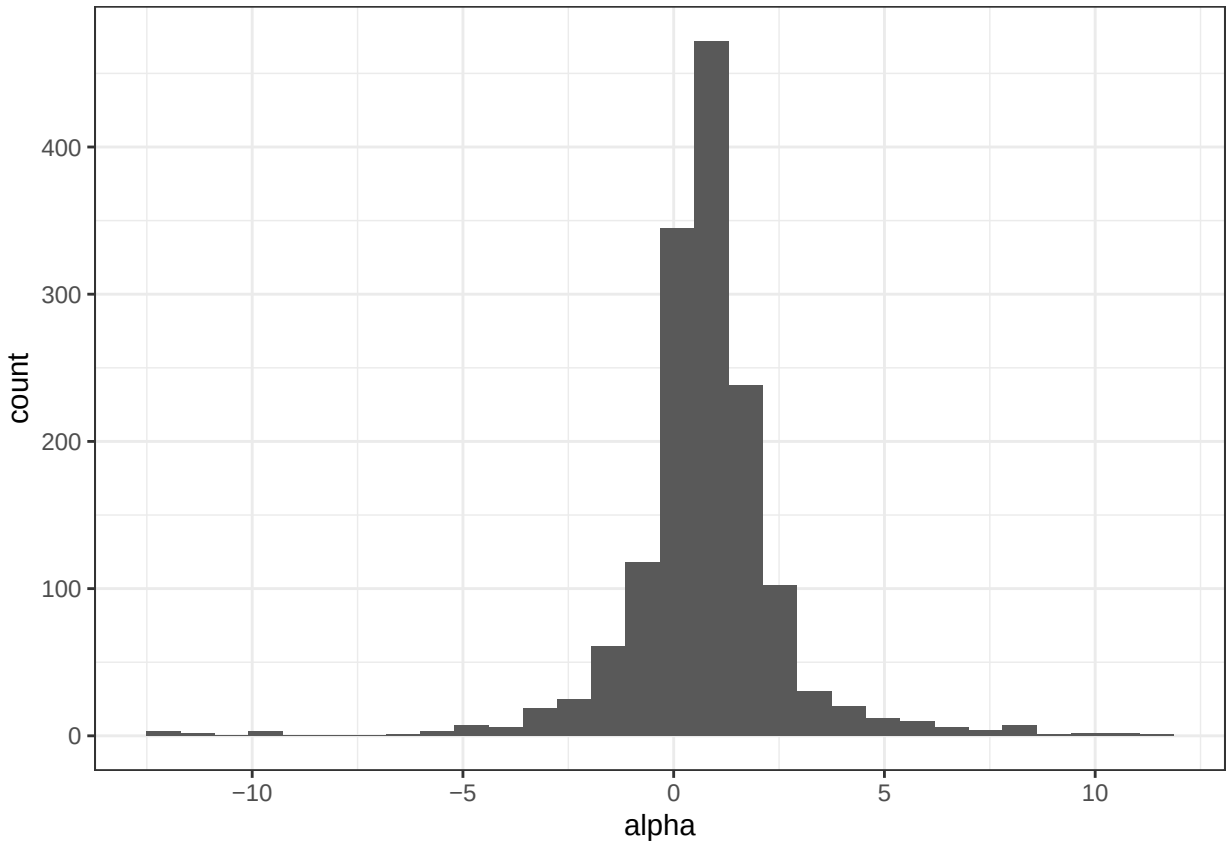
```
stan_dat <- list(y = y, x=x, N=N)
fit.flat <- stan(file = "lab-05-flat_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 2000
```

```
## Warning: There were 2 divergent transitions after warmup. See
## https://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
```

```
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: anon_model.
## 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 0.75    0.11 1.9   -3 -0.03 0.73 1.5   5.1   283    1
##
## Samples were drawn using NUTS(diag_e) at Thu Jan  8 12:04:54 2026.
## 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?

---

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?

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.

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 $\alpha \sim Unif(-10, 10)$

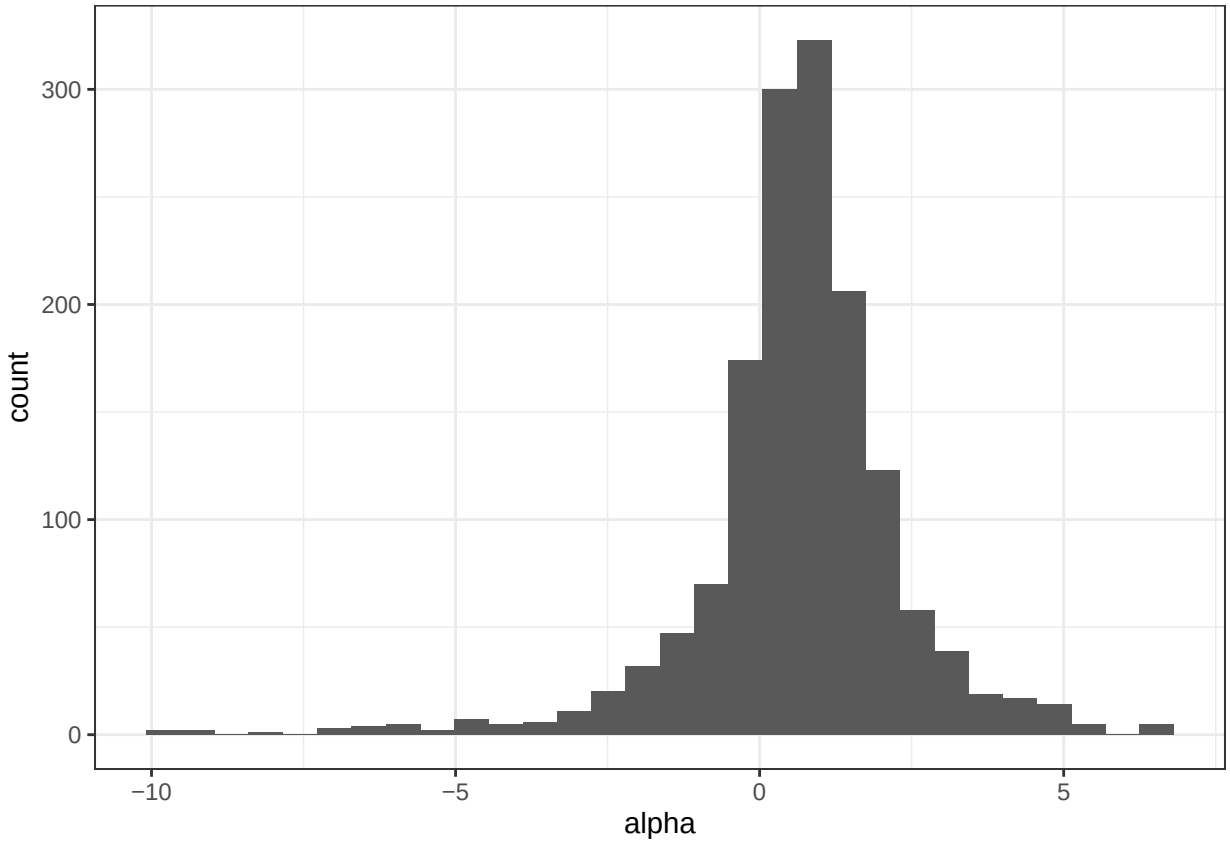
```
stan_dat <- list(y = y, x=x, N=N, lb = -10, ub = 10)
fit.unif <- stan(file = "lab-05-unif_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 2000
```

```
## Warning: There were 1 divergent transitions after warmup. See
## https://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
```

```
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)
```




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

```
## Inference for Stan model: anon_model.
## 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 0.64    0.11 1.7 -3.3 0.01 0.72 1.4   4.1   241    1
##
## Samples were drawn using NUTS(diag_e) at Thu Jan  8 12:05:08 2026.
## 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 might specify just how informative we would like our prior beliefs to be.

Weakly informative priors

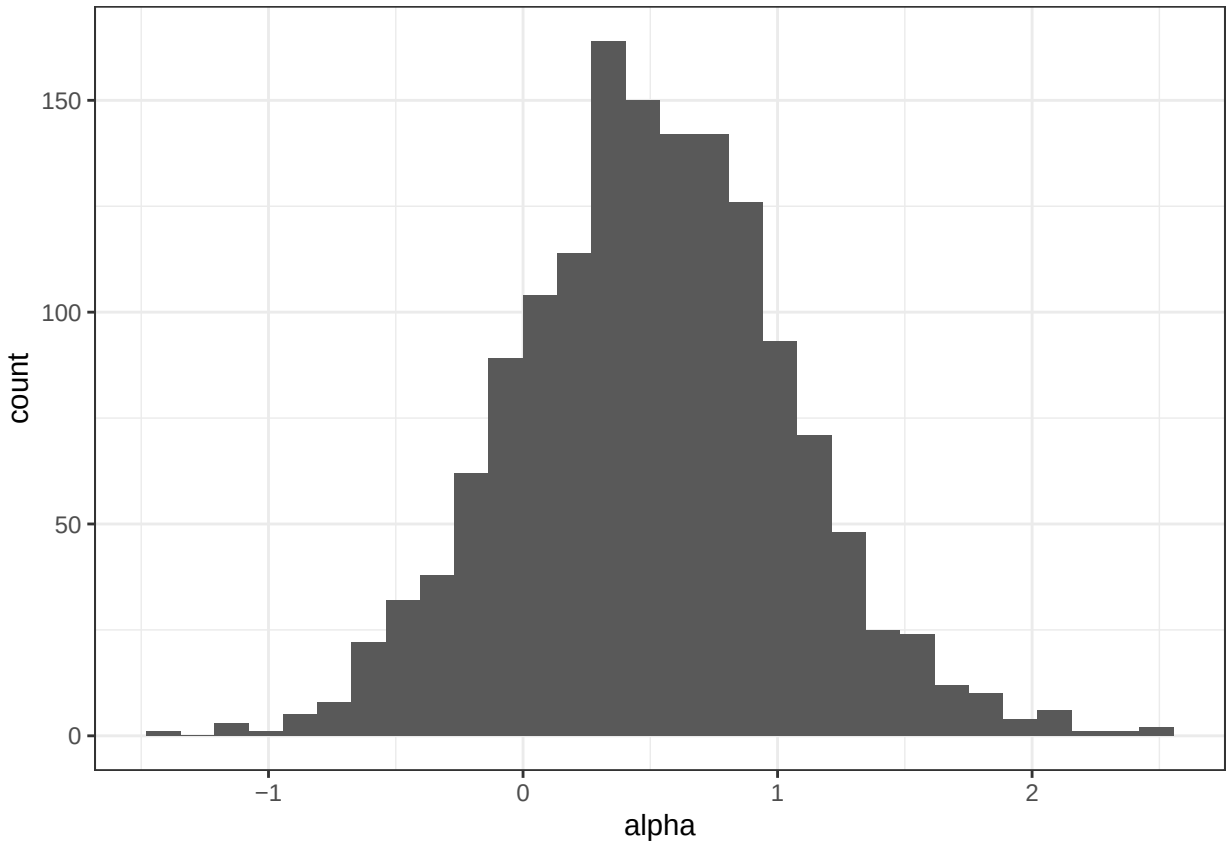
We often must consider the scale of the parameters we wish to estimate. In applied problems where we know how to interpret the parameters, the scale is easier to identify. We may consider a weakly informative prior to be such that if there is a reasonably large amount of data, the likelihood will dominate 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 prrior $\alpha \sim Unif(0, 1)$, 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 consider a $N(0, 1)$ prior.

```
stan_dat <- list(y = y, x=x, N=N)
fit.norm <- stan(file = "lab-05-normal_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 20
alpha.norm<- as.matrix(fit.norm, pars = c("alpha"))
```

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




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

```
## Inference for Stan model: anon_model.
## 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 0.51    0.02 0.54 -0.55 0.15 0.5 0.85   1.6   585    1
##
## Samples were drawn using NUTS(diag_e) at Thu Jan  8 12:05:22 2026.
## 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).
```

---

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?

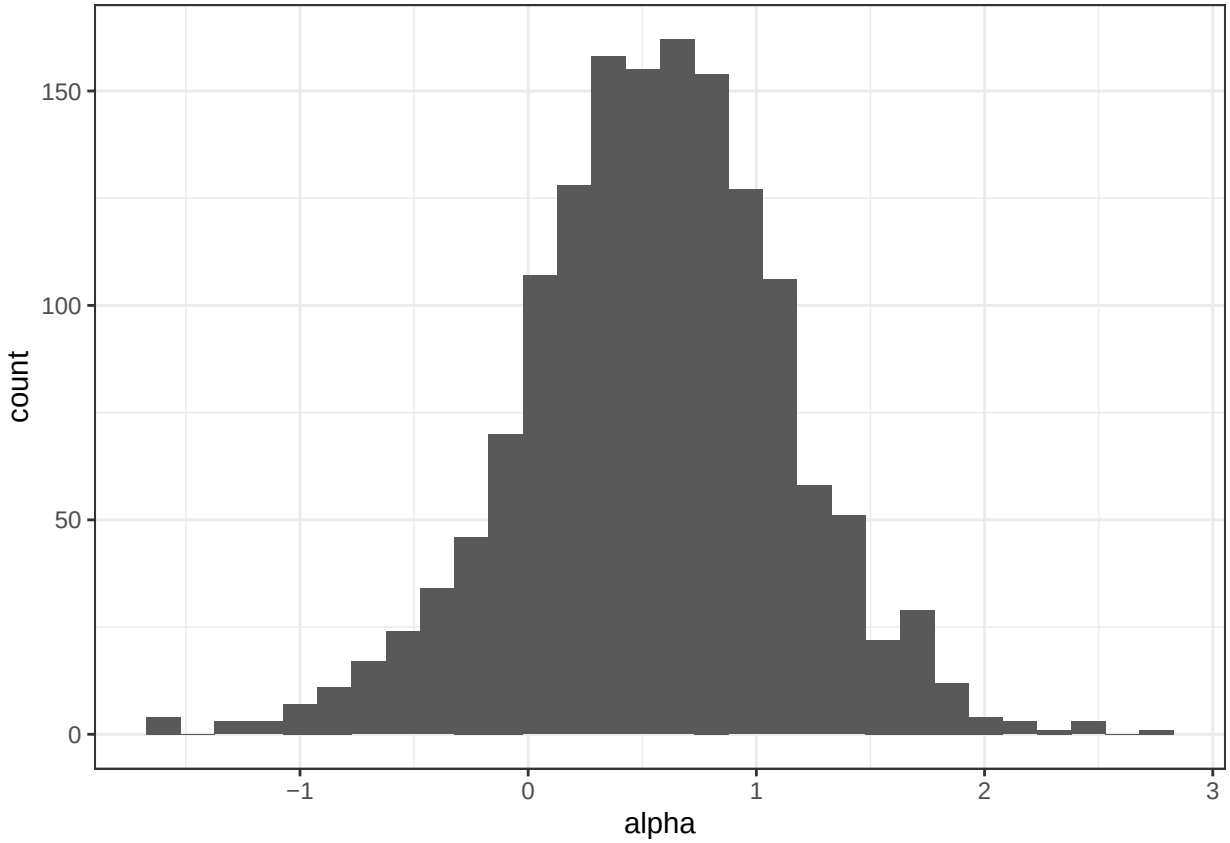
Heavy-tailed

The Cauchy distribution (the Cauchy(0,1) distribution is just the t-distribution with df=1) has heavier/fatter tails than the Normal distribution.

```
stan_dat <- list(y = y, x=x, N=N)
fit.cauchy <- stan(file = "lab-05-cauchy_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 2
```

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

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



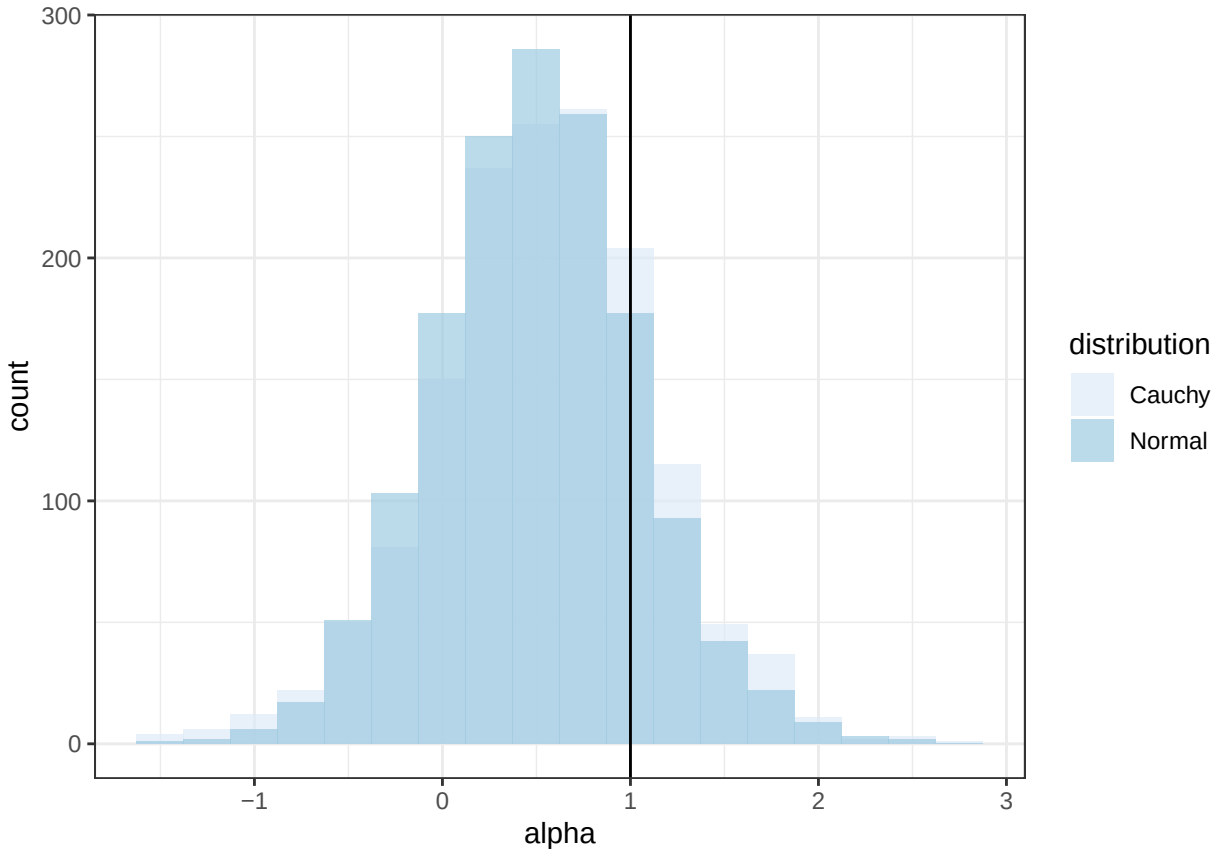

```r
print(fit.cauchy, pars = c("alpha"))
```

```
## Inference for Stan model: anon_model.
## 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 0.55    0.03 0.58 -0.68 0.19 0.56 0.93   1.7   471    1
##
## Samples were drawn using NUTS(diag_e) at Thu Jan  8 12:05:36 2026.
## 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 plot displays the posteriors for α under these two priors

```r
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()
```



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.

---

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)?

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.

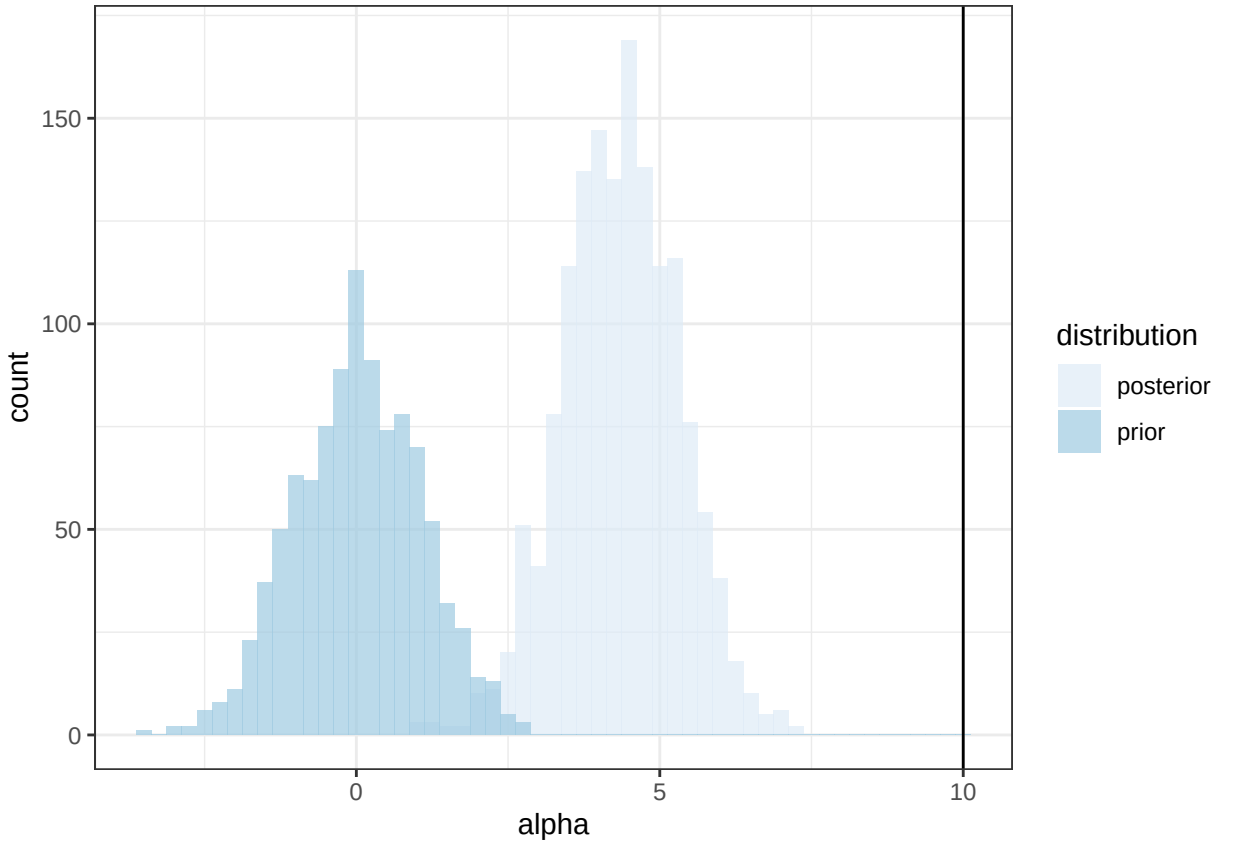
```
alpha <- 10
N <- 10
x <- runif(N, 0, 2)
y <- rnorm(N, beta * x + alpha, sigma)
```

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 = "lab-05-normal_prior.stan", data = stan_dat, chains = 1, refresh = 0, iter = 20
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()
```



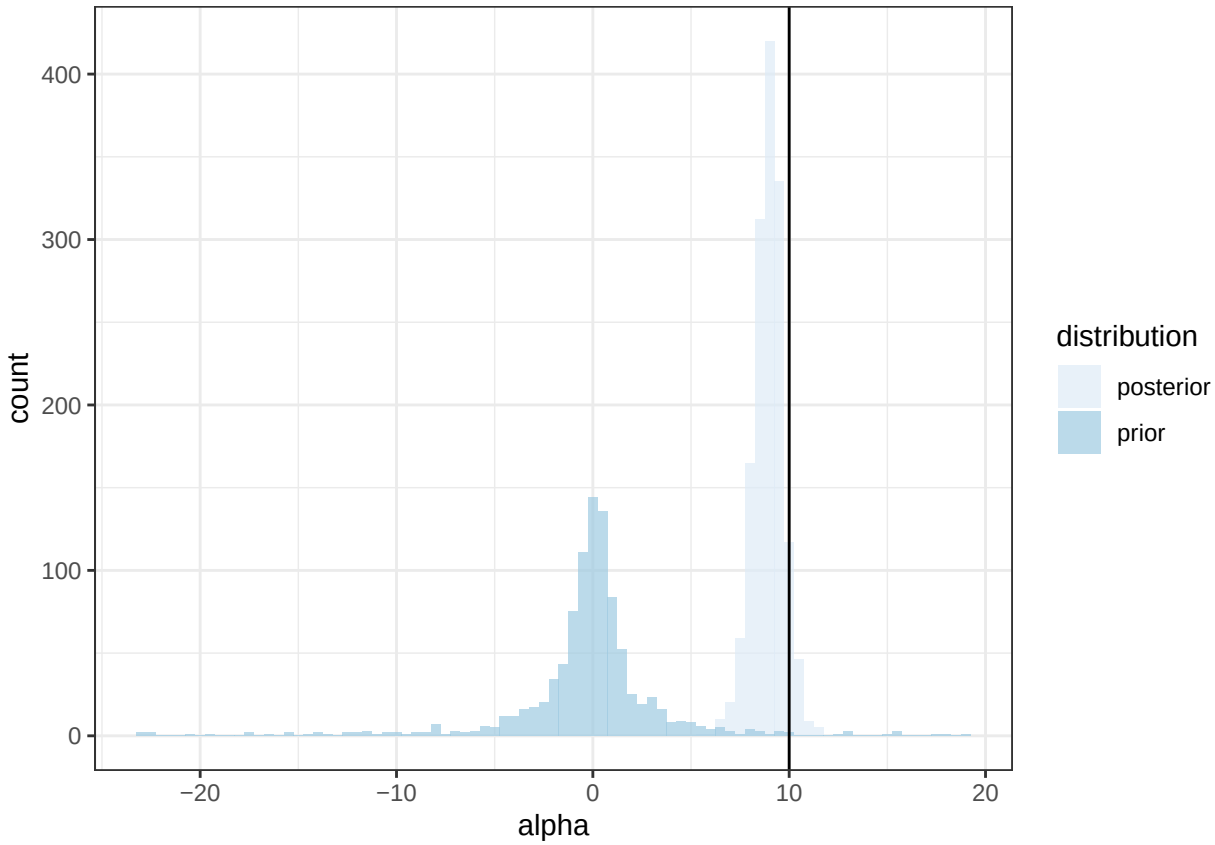
Here the prior is once again weakly-informative, but notice how the prior is dominating the likelihood. The posterior is extremely sensitive to the choice of our prior, so much so that the we do not observe posterior values close to the true α at all. Instead, the posterior is concentrated around the 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 = "lab-05-cauchy_prior.stan",data = stan_dat, chains = 1, refresh = 0, iter = 2
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()
```



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.

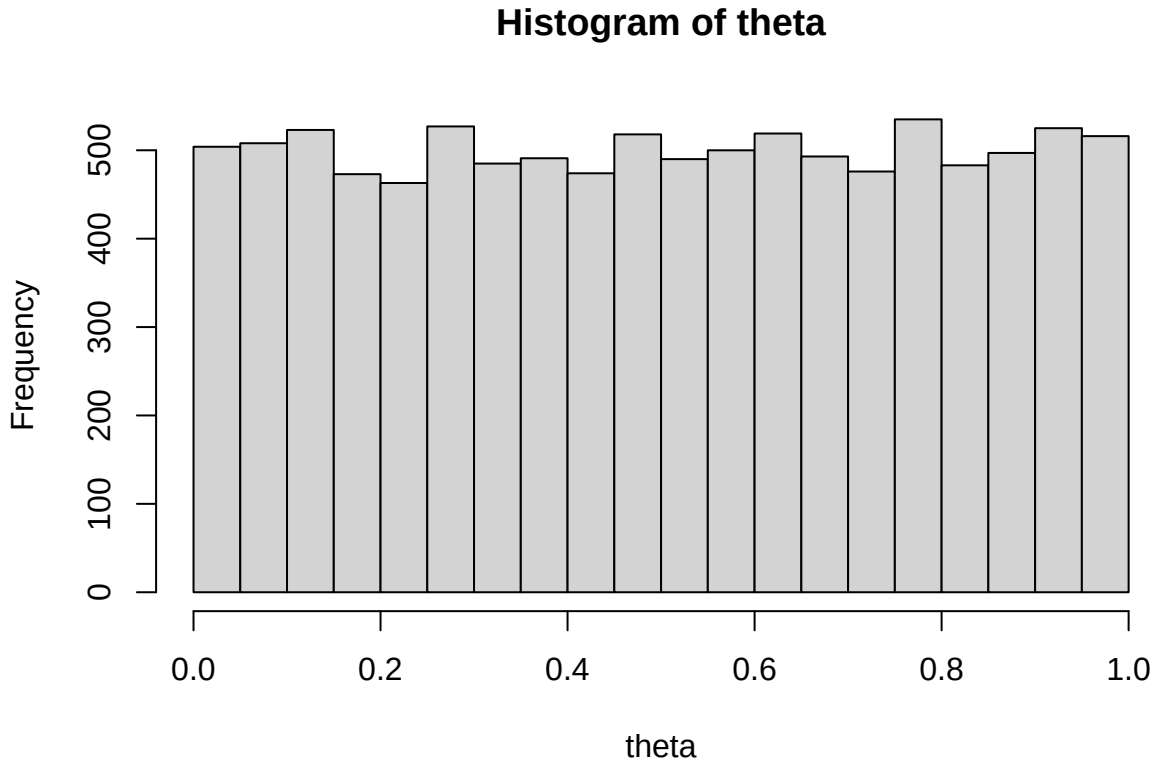
---

4. What happens as we increase the number of observations?
5. When might we prefer to use lighter- versus heavier-tailed priors?
6. How might we determine a reasonable scale for the prior?

Model reparameterization

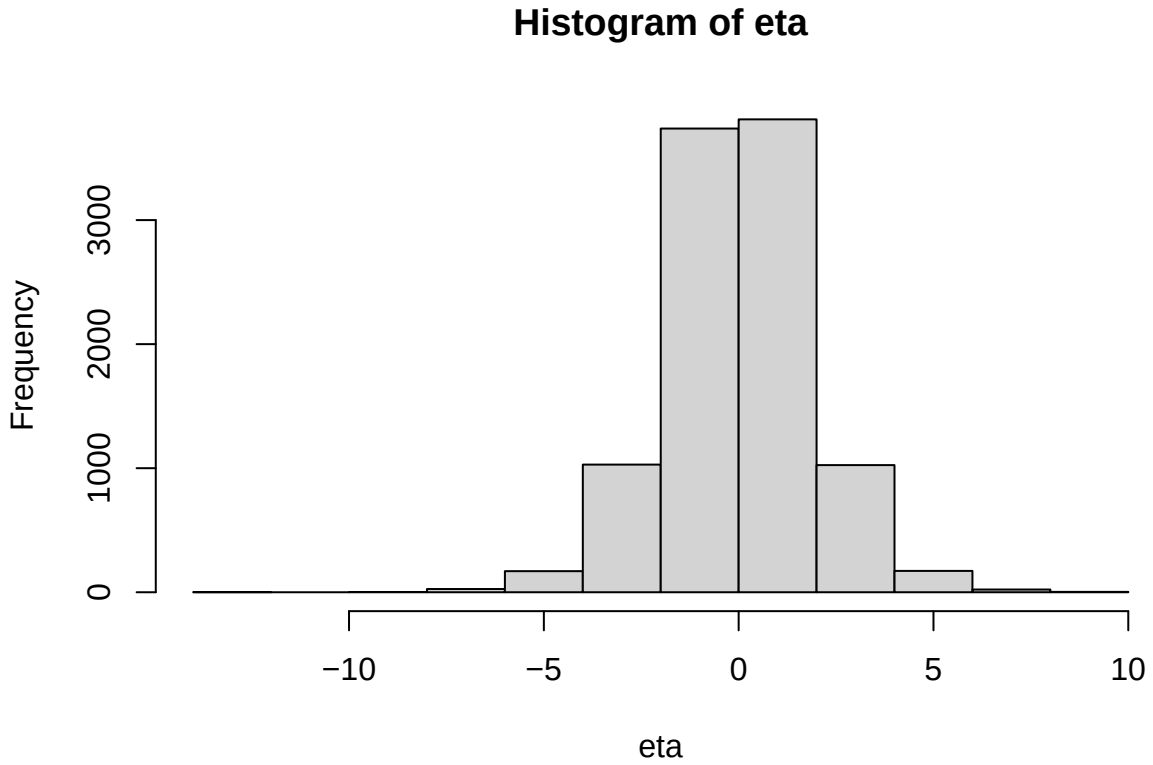
Pierre-Simon, marquis de Laplace was interested in modeling the rate of female births in Paris in the 18th century: $\theta \in [0, 1]$. Laplace had access to birth records in Paris between 1745 and 1770 with a total of n entries, and so a reasonable model for $X =$ number of female entries could be $X \sim Binomial(n, \theta)$, with n known and $\theta \in [0, 1]$. Laplace had no reason to believe that a certain value of θ was more likely than another, so he placed a Uniform(0,1) (or Beta(1,1)) prior on θ. We may think this prior $\theta \sim Unif(0, 1)$ to be uninformative–after all, it gives equal probability to every value in the range 0 and 1:

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



However, what if we consider the log-odds ratio of rate of female birth: $\eta = \log \frac{\theta}{1-\theta}$. By the same logic as above, Laplace should not have any reason to believe one value for the log-odds η_1 to be any more probable than another value η_2. Is this the case?

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



Now we see that the distribution of the log-odds of θ is most certainly not uniform. We can map directly between values θ and η, but the prior distributions do not reflect the same type of uncertainty.

Let us pretend that the true birth rate of females in Paris was 0.3, and let us simulate some data with $N = 10$ observations.

```
set.seed(123);
theta <- 0.3;
N <- 10;
y <- rbinom(N, 1, theta)
```

We know that the MLE for θ is the proportion of successes:

```
theta.mle <- sum(y)/N
```

Success probability parameterization

In this simple Beta-Binomial model, the parameter of interest is θ and we want to generate samples from the posterior $f(\theta|y)$. Take a look into the `prob.stan` file and see that we explicitly declare θ to be the parameter of interest with a Uniform(0,1) (or Beta(1,1)) prior. The following chunk of code runs performs posterior inference with this data model and prior.

```
stan_dat <- list(y = y,N=N)
fit.bayes.prob <- stan(file = "lab-05-prob.stan", data = stan_dat, refresh = 0, iter = 2000)
print(fit.bayes.prob, pars = c("theta", "eta"))
```

```
## Inference for Stan model: anon_model.
## 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  0.42    0.00 0.14  0.16  0.32  0.42 0.51  0.69  1510    1
```

```
## eta   -0.37    0.02 0.61 -1.64 -0.76 -0.33 0.05  0.79  1510    1
##
## Samples were drawn using NUTS(diag_e) at Thu Jan  8 12:05:52 2026.
## 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).
```

So we get a posterior mean $E[\theta|y]$ of 0.42 (verify this is what we would expect!).

---

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.

Log-odds parameterization

What about a uniform prior belief on the log odds η? First, recognize that the support of η is the entire real line, so a uniform prior belief on η corresponds to an infinite-support flat prior (as detailed above).

If we transformed variables back to probabilities, we would see that the induced prior on θ would be a $\text{Beta}(0, 0)$.

---

8. Is this prior proper, that is, does it integrate to 1? Does it result in a proper posterior, that is, does it integrate to 1? If so, under which conditions?

Let's try to sample from this model:

```
fit.logodds <- stan(file = "lab-05-log_odds.stan", data = stan_dat, refresh = 0, iter = 2000)
print(fit.logodds, pars = c("theta", "eta"))
```

```
## Inference for Stan model: anon_model.
## 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  0.40    0.00 0.15  0.14  0.29  0.39  0.50  0.70  1562    1
## eta   -0.46    0.02 0.68 -1.84 -0.88 -0.44 -0.01  0.83  1534    1
##
## Samples were drawn using NUTS(diag_e) at Thu Jan  8 12:06:05 2026.
## 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).
```

Odds parameterization

Consider yet another parameterization of the binomial sampling model in terms of the (un-logged) odds $\pi = \frac{\theta}{1-\theta}$.

---

9. If we set a uniform prior for π, what is the induced prior on $\theta = \frac{\pi}{(1+\pi)}$ (do a quick review of functions of random variables if you do not remember)? Is this prior proper, that is, does it integrate to 1?

From our work above, we realize that a uniform prior belief about π does not necessarily translate into uniform priors for transformations of π.

Jeffreys' prior

Jeffreys (1961) suggested a default procedure for specifying a prior distribution for any parameter in the sampling model, which he designed to be invariant under transformations:

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta)$ is the Fisher Information *of the sampling model* $p(x|\theta)$:

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2 \log\left(p(x|\theta)\right)}{\partial \theta^2}\right]$$

and the expectation is taken with respect to the sampling distribution $p(x|\theta)$. It is important to recognize that because the Fisher Information is determined by the sampling model, the Jeffreys prior is also determined by the sampling model. Specifying a Jeffreys prior for this model would thus be useful.

This is meant as a quick introduction to Jeffreys prior; we will dive deeper into this in the main lectures.

Acknowledgement

This lab was created by Jordan Bryan and Becky Tang. A part of it was adapted from this tutorial.