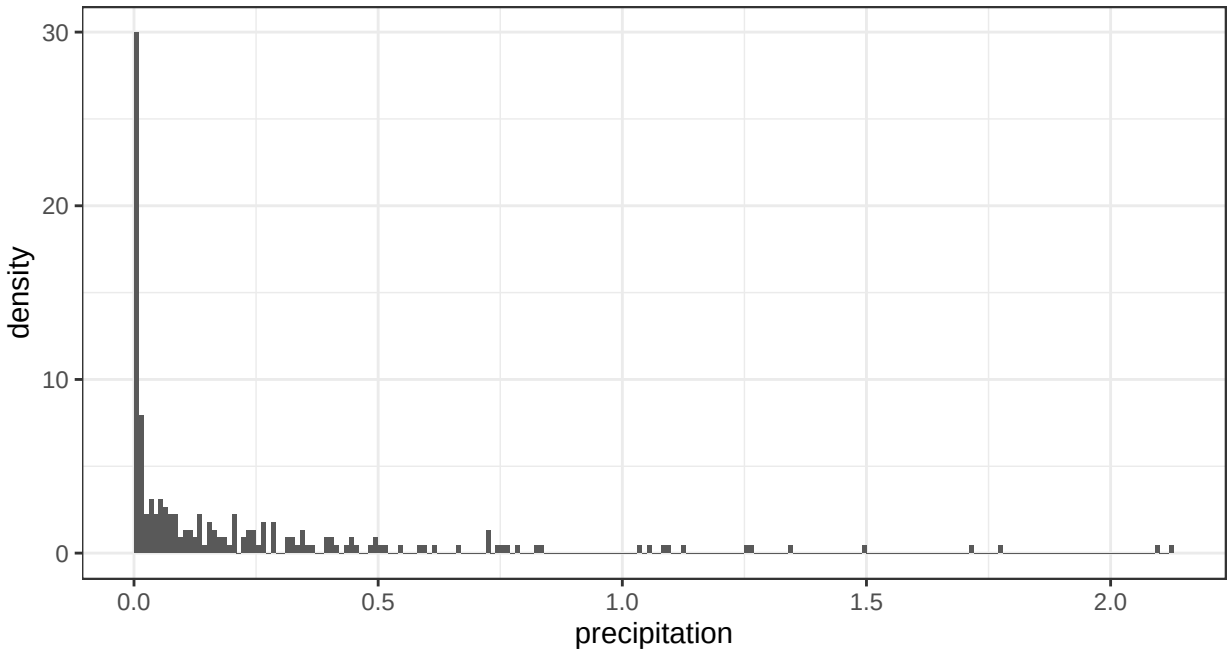
Stat 343: Numerically Maximizing the Likelihood via Stan

Rainfall in Illinois, all storms from 1960 - 1964

The code below reads in the data and makes an initial plot:



- Let's model these data with a Gamma distribution and estimate the parameters by maximum likelihood.
- We saw before that we will need to use some sort of numerical method to maximize the likelihood in order to obtain parameter estimates.

Overview of Stan

- Stan is a programming language that can be used to specify statistical models
- Most often, used for parameter estimation in Bayesian framework (coming up soon!)
- Also provides methods for maximizing the log-likelihood:
  - Newton's method
  - **L-BFGS** is the default: basically Newton's method, but uses an approximation to the Hessian matrix of the log-likelihood
- Stan code is compiled to C++ code, which is in turn compiled to an executable file
- There are interfaces to interact with Stan from R, Python, MATLAB, Julia, Stata, and the command line

For our purposes, there are 2 steps to using Stan:

1. Define a statistical model in a separate .stan file.
2. From R, tell Stan to compile your model and do estimation.

Model Definition in Stan

- The following code is stored in a separate file called `gamma_model.stan`.
- It defines the **data** that are observed, the model **parameter**s, and the statistical **model**.
- Fundamentally, the file describes how to calculate the joint pdf of the data as a function of unknown parameters.

```
data {
  // sample size; a non-negative integer.  "data" because it is a known, observed quantity
  int<lower=0> n;

  // vector of n observations, each of which is a real number; rainfall amounts
  real x[n];
}

parameters {
  // the shape parameter for the Gamma distribution; a non-negative real number
  real<lower=0> alpha;

  // the scale parameter for the Gamma distribution; a non-negative real number
  real<lower=0> beta;
}

model {
  // each element of the vector x is modeled as following a Gamma(alpha, beta) distribution
  x ~ gamma(alpha, beta);
}
```

Equivalent model definition

The only change is in the model block, which makes more explicit what's happening behind the scenes:

- `target` is the value of the log of the pdf of the data (i.e., the log-likelihood function)
- At each step of optimization, `target` is re-initialized to 0
- The for loop then iterates through all observations in the data set `x[1]`, ..., `x[n]` and adds the log of the pdf for a gamma distribution with current parameter values `alpha` and `beta`
- Stan keeps track of all the contributions that were made to target and automatically differentiates the target with respect to parameters declared in the parameters block

```
data {
  // sample size; a non-negative integer
  int<lower=0> n;

  // vector of n observations, each of which is a real number; rainfall amounts
  real x[n];
}

parameters {
  // the shape parameter for the Gamma distribution; a non-negative real number
  real<lower=0> alpha;

  // the scale parameter for the Gamma distribution; a non-negative real number
  real<lower=0> beta;
}

model {
  // each element of the vector x is modeled as following a Gamma(alpha, beta) distribution
  for(i in 1:n) {
    target += gamma_lpdf(x[i] | alpha, beta);
  }
}
```

Performing optimization by calling Stan from R

We need to do the following tasks:

1. Read in our data
2. Set up a list with the "data" Stan needs to know about
3. Compile the Stan model definition to an executable
4. Call Stan to do the optimization
5. Extract the parameter estimates

All of the below (continuing on next page) is R code that is in a .Rmd or .R file.

```
# Load the rstan package
library(rstan)

# Read in data
il_storms <- bind_rows(
  read_csv("http://www.evanlray.com/data/rice/Chapter%2010/illinois60.txt", col_names = FALSE),
  read_csv("http://www.evanlray.com/data/rice/Chapter%2010/illinois61.txt", col_names = FALSE),
  read_csv("http://www.evanlray.com/data/rice/Chapter%2010/illinois62.txt", col_names = FALSE),
  read_csv("http://www.evanlray.com/data/rice/Chapter%2010/illinois63.txt", col_names = FALSE),
  read_csv("http://www.evanlray.com/data/rice/Chapter%2010/illinois64.txt", col_names = FALSE)
)
names(il_storms) <- "precipitation"

# Set up list with data Stan will need to know about
stan_data <- list(
  n = nrow(il_storms),
  x = il_storms$precipitation
)

# Compile the Stan model definition to an executable.  Takes a few seconds to run.
gamma_model_compiled <- stan_model(file = "gamma_model.stan")

# Call Stan to do optimization
gamma_fit <- optimizing(gamma_model_compiled,
  data = stan_data,
  seed = 8742,
  init = "random"
)
```

Here's a look at the return object, which is a list with 4 components:

- par is a named vector of parameter estimates
- value is the value of the log-likelihood at the maximum, after dropping constants that don't involve the parameters
- return_code is 0 if everything went well in the optimization procedure, otherwise an error code to be sad about
- we will discuss theta_tilde later

```
gamma_fit
```

```
## $par
##  alpha   beta
## 0.4408 1.9643
##
## $value
## [1] 185.3
##
## $return_code
## [1] 0
##
## $theta_tilde
##       alpha  beta
## [1,] 0.4408 1.964
```

Let's make a plot!

```
ggplot(data = il_storms, mapping = aes(x = precipitation)) +
  geom_histogram(center = 0.005, binwidth = 0.01, mapping = aes(y = ..density..)) +
  stat_function(fun = dgamma,
    args = list(shape = gamma_fit$par["alpha"], rate = gamma_fit$par["beta"]),
    color = "orange") +
  theme_bw()
```