In lecture 16, we looked at precipitation amounts in Madison County (at Morrisville station). We found that the Weibull distribution had a good fit to the monthly precipitation amounts.

We found that the MLEs for the Weibull distribution were

$$\hat{a} = 2.1871$$
$$\hat{\sigma} = 3.9683$$

and

$$-\mathcal{L}(\{\hat{a}, \hat{\sigma}\}|\mathbf{x}) = 2166.496$$

is the realized negative log-likelihood. Note this means that the log-likelihood is

$$\mathcal{L}(\{\hat{a}, \hat{\sigma}\}|\mathbf{x}) = -2166.496,$$

and the usual likelihood is

$$L(\{\hat{a}, \hat{\sigma}\}|\mathbf{x}) = e^{[\mathcal{L}(\{\hat{a}, \hat{\sigma}\}|\mathbf{x})]} \approx= e^{-2166.496},$$

which `R` cannot differentiate from 0.

1. Someone asked "why Weibull?" in class. That is, why wouldn't we use another right-skewed distribution like the Gamma (see Lecture 15), or the Log-Normal (see Lecture 17).

   (a) Compute the MLEs for these data using a Gamma distribution.

```
    library(tidyverse)

#Data and Data Cleaning
data <- read_csv("agacis.csv")

## Rows:  115 Columns:  14
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr (13):  Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, Annual
## dbl  (1):  Year
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

data.long <- data |>
  dplyr::select(-Annual) |>                     # Remove annual column
  pivot_longer(cols = c(Jan, Feb, Mar, Apr,    # pivot the column data into one col
                        May, Jun, Jul, Aug,
                        Sep, Oct, Nov, Dec),
               values_to = "Precipitation",    # store the values in Precipitation
               names_to = "Month") |>          # store the months in Month
  mutate(Precipitation = case_when(Precipitation == "M" ~ NA_character_,
                                   TRUE                 ~ Precipitation))|>
  mutate(Precipitation = as.numeric(Precipitation))

#MLE for Gamma
llgamma <- function(data, par, neg=F){
  alpha <- exp(par[1])
  beta <- exp(par[2])

  loglik <- sum(log(dgamma(x=data, shape=alpha, rate=beta)), na.rm=T)
```

```
  return(ifelse(neg, -loglik, loglik))
}
mle_gamma <- optim(par = c(1,1),
              fn = llgamma,
              data=data.long$Precipitation,
              neg=T)
alpha=mle_gamma$par[1]
beta=mle_gamma$par[2]
```

(b) Compute the MLEs for these data using the Log-Normal distribution.

```
    library(tidyverse)

#Data and Data Cleaning
data <- read_csv("agacis.csv")

## Rows:  115 Columns:  14
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr (13):  Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, Annual
## dbl  (1):  Year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

data.long <- data |>
  dplyr::select(-Annual) |>                  # Remove annual column
  pivot_longer(cols = c(Jan, Feb, Mar, Apr,   # pivot the column data into one col
                      May, Jun, Jul, Aug,
                      Sep, Oct, Nov, Dec),
               values_to = "Precipitation",   # store the values in Precipitation
               names_to = "Month") |>         # store the months in Month
  mutate(Precipitation = case_when(Precipitation == "M" ~ NA_character_,
                                    TRUE                ~ Precipitation))|>
  mutate(Precipitation = as.numeric(Precipitation))
    #MLE for lognormal
n=1380 #1380 observations
lognorm.mu.hat.mle   <- mean(log(data.long$Precipitation), na.rm = T)
lognorm.sigma.hat.mle <-sqrt(sum((log(data.long$Precipitation) - lognorm.mu.hat.mle)^2, na.rm =
```

(c) Compute the likelihood ratio to compare the Weibull and the Gamma distribution. Which has a better fit according to the likelhiood ratio?

$$Q = \frac{L(\{\hat{a}, \hat{\sigma}\}|\mathbf{x})}{L(\{\hat{\alpha}, \hat{\beta}\}|\mathbf{x})} = e^{\left[\mathcal{L}(\{\hat{a}, \hat{\sigma}\}|\mathbf{x}) - \mathcal{L}(\{\hat{\alpha}, \hat{\beta}\}|\mathbf{x})\right]}$$

```
#Data and Data Cleaning
data <- read_csv("agacis.csv")

## Rows:  115 Columns:  14
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr (13):  Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, Annual
## dbl  (1):  Year
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
data.long <- data |>
  dplyr::select(-Annual) |>                    # Remove annual column
  pivot_longer(cols = c(Jan, Feb, Mar, Apr,    # pivot the column data into one col
                        May, Jun, Jul, Aug,
                        Sep, Oct, Nov, Dec),
               values_to = "Precipitation",    # store the values in Precipitation
               names_to = "Month") |>          # store the months in Month
  mutate(Precipitation = case_when(Precipitation == "M" ~ NA_character_,
                                   TRUE                 ~ Precipitation))|>
  mutate(Precipitation = as.numeric(Precipitation))

#MLE for Gamma
llgamma <- function(data, par, neg=F){
  alpha <- exp(par[1])
  beta <- exp(par[2])

  loglik <- sum(log(dgamma(x=data, shape=alpha, rate=beta)), na.rm=T)

  return(ifelse(neg, -loglik, loglik))
}
mle_gamma <- optim(par = c(1,1),
                   fn = llgamma,
                   data=data.long$Precipitation,
                   neg=T)
# Log-Likelihood for Weibull
loglik_weibull <- -2166.496

# Log-Likelihood for Gamma
loglik_gamma <- -mle_gamma$value  # Negative of the value from optim()

# Likelihood Ratio for Weibull vs Gamma
likelihood_ratio_gamma <- exp(loglik_weibull - loglik_gamma)
likelihood_ratio_gamma
```

Since the likelihood ratio is 2.16133×10^{-7}, it suggests that the Gamma Distribution fits the data better than the Weibull Distribution.

(d) Compute the likelihood ratio to compare the Weibull and the Log-Normal distribution. Which has a better fit according to the likelihood ratio?

$$Q = \frac{L(\{\hat{a}, \hat{\sigma}\}|\mathbf{x})}{L(\{\hat{\mu}, \hat{\sigma}\}|\mathbf{x})} = e^{[\mathcal{L}(\{\hat{a},\hat{\sigma}\}|\mathbf{x}) - \mathcal{L}(\{\hat{\mu},\hat{\sigma}\}|\mathbf{x})]}$$

```r
#Data and Data Cleaning
data <- read_csv("agacis.csv")
```

```
## Rows: 115 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (13): Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, Annual
## dbl  (1): Year
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

data.long <- data |>
  dplyr::select(-Annual) |>                   # Remove annual column
  pivot_longer(cols = c(Jan, Feb, Mar, Apr,   # pivot the column data into one col
                        May, Jun, Jul, Aug,
                        Sep, Oct, Nov, Dec),
               values_to = "Precipitation",   # store the values in Precipitation
               names_to = "Month") |>         # store the months in Month
  mutate(Precipitation = case_when(Precipitation == "M" ~ NA_character_,
                                   TRUE                 ~ Precipitation))|>
  mutate(Precipitation = as.numeric(Precipitation))

# Log-Likelihood for Weibull
loglik_weibull <- -2166.496

# Log-Likelihood for Lognormal
loglik_lognormal <- sum(-log(data.long$Precipitation) - 0.5 * log(2 * pi * lognorm.sigma.hat.ml


# Likelihood Ratio for Weibull vs Lognormal
likelihood_ratio_lognormal <- exp(loglik_weibull - loglik_lognormal)
likelihood_ratio_lognormal
```

Since the likelihood ratio is very large, 8.112098×10^{20}, it suggests that the Weibull distribution fits the data a lot better than the Lognormal distribution.

(e) Compute the likelihood ratio to compare the Gamma and the Log-Normal distribution. Which has a better fit according to the likelhiood ratio?

$$Q = \frac{L(\{\hat{\alpha}, \hat{\beta}\}|\mathbf{x})}{L(\{\hat{\mu}, \hat{\sigma}\}|\mathbf{x})} = e^{\left[\mathcal{L}(\{\hat{\alpha}, \hat{\beta}\}|\mathbf{x}) - \mathcal{L}(\{\hat{\mu}, \hat{\sigma}\}|\mathbf{x})\right]}$$

```
#Data and Data Cleaning
data <- read_csv("agacis.csv")

## Rows:  115 Columns:  14
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr (13):  Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec, Annual
## dbl  (1):  Year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

data.long <- data |>
  dplyr::select(-Annual) |>                   # Remove annual column
  pivot_longer(cols = c(Jan, Feb, Mar, Apr,   # pivot the column data into one col
                        May, Jun, Jul, Aug,
                        Sep, Oct, Nov, Dec),
               values_to = "Precipitation",   # store the values in Precipitation
               names_to = "Month") |>         # store the months in Month
```

```r
  mutate(Precipitation = case_when(Precipitation == "M" ~ NA_character_,
                                   TRUE                 ~ Precipitation))|>
  mutate(Precipitation = as.numeric(Precipitation))

# Log-Likelihood for Gamma
loglik_gamma <- -mle_gamma$value  # Negative of the value from optim()

# Log-Likelihood for Lognormal
loglik_lognormal <- sum(-log(data.long$Precipitation) - 0.5 * log(2 * pi * lognorm.sigma.hat.ml

# Likelihood Ratio for Gamma vs Lognormal
likelihood_ratio_gamma_lognormal <- exp(loglik_gamma - loglik_lognormal)
likelihood_ratio_gamma_lognormal

## [1] 3.75329e+27
```

Since the likelihood ratio is very large, 3.75329×10^{27}, the Gamma distribution fits the data better for this model.

2. Optional Coding Challenge. Choose the "best" distribution and refit the model by season.
   (a) Fit the Distribution for Winter (December-February).
   (b) Fit the Distribution for Spring (March-May).
   (c) Fit the Distribution for Summer (June-August).
   (d) Fit the Distribution for Fall (September-November).
   (e) Plot the four distributions in one plot using `cyan3` for Winter, `chartreuse3` for Spring, `red3` for Summer, and `chocolate3` for Fall. Note any similarities/differences you observe across the seasons.