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).

```r
# Load and clean data about precipitation in Madison County
dat.precip <- read_csv(file = "agacis.csv")

dat.precip.long <- dat.precip |>
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
#switch 'M' to NA values and convert numbers to integers from Strings
mutate(Precipitation = case_when(Precipitation == "M" ~ NA_character_,
                                 TRUE                 ~ Precipitation))|>
mutate(Precipitation = as.numeric(Precipitation))
```

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

```r
#Function to compute Maximum Likelihood
llgamma <- function(par, data, neg = F){
  alpha <- par[1] #get alpha and beta
  beta <- par[2]

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

  return(ifelse(neg, - loglik, loglik))
}

#Compute MLE
MLE.gamma <- optim(par = c(1,1),
               fn = llgamma,
               data=dat.precip.long$Precipitation,
               neg=T)

#extract alpha and beta
alpha.MLE <- MLE.gamma$par[1]
beta.MLE <- MLE.gamma$par[2]

#print the values
alpha.MLE

## [1] 4.174581

beta.MLE

## [1] 1.189099
```

We computed the MLEs for these data using a Gamma distribution to obtain $\alpha = 4.1745814$ and $\beta = 1.1890993$.

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

```r
#Function to compute Maximum Likelihood
lllognorm <- function(par, data, neg = F){
  mu <- par[1] #get mu and sigma
  sigma <- par[2]

  #compute log likelihood
  loglik <- sum(log(dlnorm(x = data, meanlog = mu, sdlog = sigma)), na.rm = T)

  return(ifelse(neg, - loglik, loglik))
}

#Compute MLE
MLE.lognorm <- optim(par = c(1,1),
               fn = lllognorm,
               data=dat.precip.long$Precipitation,
               neg=T)

#extract alpha and beta
mu.MLE <- MLE.lognorm$par[1]
sigma.MLE <- MLE.lognorm$par[2]

#print the values
mu.MLE

## [1] 1.131261

sigma.MLE

## [1] 0.5333417
```

We computed the MLEs for these data using a Log-Normal distribution to obtain $\mu = 1.1312609$ and $\sigma = 0.5333417$.

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

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

```r
#compute log-likelihood for Gamma distribution
Gamma.loglik <- llgamma(par = c(alpha.MLE, beta.MLE), data = dat.precip.long$Precipitation, neg = T)

#compute the likelihood ratio
Weibull.loglik <- 2166.496
q.gamma.weibull <- exp(Weibull.loglik-Gamma.loglik)

#print Q
q.gamma.weibull

## [1] 4626807
```

The likelihood ratio of the Weibull and the Gamma distribution is 4.6268066×10^6. Because $Q > 1$ (Q is significantly larger than 1), the numerator provides a better fit for the data than the denominator. Weibull distribution is in the numerator, so it is a better fit according to the likelihood ratio.

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

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

```r
#compute log-likelihood for Log-Normal distribution
Lognorm.loglik <- lllognorm(c(mu.MLE, sigma.MLE), data = dat.precip.long$Precipitation, neg = T)

#compute the likelihood ratio
q.lognorm.weibull <- exp(Weibull.loglik-Lognorm.loglik)

#print Q
q.lognorm.weibull

## [1] 4.218221e-17
```

The likelihood ratio of the Weibull and the Log-Normal distribution is $4.2182211 \times 10^{-17}$. Because $Q < 1$, the denominator provides a better fit for the data than the numerator. Log-Normal distribution is in the denominator, so it is a better fit according to the likelihood ratio.

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

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

```
#compute the likelihood ratio
q.gamma.lognorm <- exp(Gamma.loglik-Lognorm.loglik)

#print Q
q.gamma.lognorm

## [1] 9.116917e-24
```

The likelihood ratio of the Gamma and the Log-Normal distribution is $9.1169169 \times 10^{-24}$. Because $Q < 1$, the denominator provides a better fit for the data than the numerator. Log-Normal distribution is in the denominator, so it is a better fit according to the likelihood ratio.

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.