Homework 6

Spring 2023

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday May 17 by midnight PST.
- Please answer the following questions in the order in which they are posed.
- Don't forget to knit the document frequently to make sure there are no compilation errors.
- When you are done, download the PDF file as instructed in section and submit it in Gradescope.
- Rule on collaboration: You may guide someone who is stuck by giving high level advice. However, everyone is expected to write up their answers individually, which includes deciding what and how much to explain and entirely in your own words.

Exercises

1. (Starch or sugar) The results of scoring the offspring plants of a particular plant species as either starchy or sugary and as having either a green or a white base leaf appear below.

1) starchy-green	2) starchy-white	3) sugary-green	4) sugary-white
1997	906	904	32

According to a genetic model for these traits, the probability that a plant exhibits one of these trait combinations should be $\frac{1}{4}(2+\theta_0)$ for the first combination, $\frac{1}{4}(1-\theta_0)$ for the middle two combinations and $\frac{1}{4}\theta_0$ for the last where θ_0 is a probability related to linkage closeness.

a. Determine the MLE of θ_0. Be sure to clearly show

   - Log likelihood function $\ell(\theta)$ in terms of the counts x_1, x_2, x_3, x_4 (i.e., not just for this data)
   - First derivative equation and calculate the MLE $\widehat{\theta}_0^{mle}$.

Suppose we are looking to model the scoring of the offspring of a particular plant as either starchy or sugary and as having either a green or a white base leaf. Then according to a genetic model for these traits, the probability that a plant exhibits one of these trait combinations should be $\frac{1}{4}(2+\theta_0)$ for the first combination, $\frac{1}{4}(1-\theta_0)$ for the middle two combinations and $\frac{1}{4}\theta_0$ for the last where θ_0 is a probability related to linkage closeness. If we let X_1 denote the event that a given plant is labeled as starchy-green, X_2 denote the event that a given plant is labeled as starchy-white, X_3 denote the event that a given plant is labeled as sugary-green, and X_4 denote the event that a given plant is labeled as sugary-white, then it follows that we can model the vector of random variables $\langle X_1, X_2, X_3, X_4 \rangle$ with a multinomial distribution.

Based on Theorem 14.1, the PMF of $\langle X_1, X_2, X_3, X_4\rangle$ can be written as $f(x_1,x_2,x_3,x_4) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \frac{n!}{x_1!x_2!x_3!x_4!}\pi_1^{x_1}\pi_2^{x_2}\pi_3^{x_3}\pi_4^{x_4} = \frac{n!}{x_1!x_2!x_3!x_4!}\left(\frac{1}{4}(2+\theta_0)\right)^{x_1}\left(\frac{1}{4}(1-\theta_0)\right)^{x_2}\left(\frac{1}{4}(1-\theta_0)\right)^{x_3}\left(\frac{1}{4}\theta_0\right)^{x_4}$, $0 \leq \theta_0 \leq 1$. Where $x_1+x_2+x_3 = n$ and $\pi_1+\pi_2+\pi_3 = 1$. Based on this following PMF definition, it follows that the likelihood function of θ is $L(\theta) = \frac{n!}{x_1!x_2!x_3!x_4!}\left(\frac{1}{4}(2+\theta)\right)^{x_1}\left(\frac{1}{4}(1-\theta)\right)^{x_2}\left(\frac{1}{4}(1-\theta)\right)^{x_3}\left(\frac{1}{4}\theta\right)^{x_4},\ 0 \leq \theta \leq 1$. We will now find the log-likelihood function by taking the natural log of the preceding likelihood function.

$$\begin{aligned}
\ell(\theta) &= \ln(L(\theta)) \\
&= \ln\left(\frac{n!}{x_1!x_2!x_3!x_4!}\left(\frac{1}{4}(2+\theta)\right)^{x_1}\left(\frac{1}{4}(1-\theta)\right)^{x_2}\left(\frac{1}{4}(1-\theta)\right)^{x_3}\left(\frac{1}{4}\theta\right)^{x_4}\right) \\
&= \ln\left(\frac{n!}{x_1!x_2!x_3!x_4!}\right) + \ln\left(\left(\frac{1}{4}(2+\theta)\right)^{x_1}\right) + \ln\left(\left(\frac{1}{4}(1-\theta)\right)^{x_2+x_3}\right) + \ln\left(\left(\frac{1}{4}\theta\right)^{x_4}\right) \\
&= \ln(n!) - \ln(x_1!x_2!x_3!x_4!) + x_1\ln\left(\frac{1}{4}\theta+\frac{1}{2}\right) + (x_2+x_3)\ln\left(\frac{1}{4}-\frac{1}{4}\theta\right) + x_4\ln\left(\frac{1}{4}\theta\right),\ 0 \leq \theta \leq 1
\end{aligned}$$

Thus as computed above, the log-likelihood function of θ is $\ln(n!) - \ln(x_1!x_2!x_3!x_4!) + x_1\ln\left(\frac{1}{4}\theta+\frac{1}{2}\right) + (x_2+x_3)\ln\left(\frac{1}{4}-\frac{1}{4}\theta\right) + x_4\ln\left(\frac{1}{4}\theta\right),\ 0 \leq \theta \leq 1$. Now in order to find the candidates for the MLE of θ_0, we must take the first derivative of the log-likelihood function and set it equal to zero.

$$\begin{aligned}
\frac{d}{d\theta}\ell(\theta) &= \frac{d}{d\theta}\left(\ln(n!) - \ln(x_1!x_2!x_3!x_4!) + x_1\ln\left(\frac{1}{4}\theta+\frac{1}{2}\right) + (x_2+x_3)\ln\left(\frac{1}{4}-\frac{1}{4}\theta\right) + x_4\ln\left(\frac{1}{4}\theta\right)\right) \\
&= \frac{\frac{1}{4}x_1}{\frac{1}{4}\theta+\frac{1}{2}} - \frac{\frac{1}{4}(x_2+x_3)}{\frac{1}{4}-\frac{1}{4}\theta} + \frac{\frac{1}{4}}{\frac{1}{4}\theta} \\
&= \frac{x_1}{\theta+2} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta},\ 0 \leq \theta \leq 1
\end{aligned}$$

Thus as computed above $\ell'(\theta) = \frac{x_1}{\theta+2} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta},\ 0 \leq \theta \leq 1$, setting this equal to zero we obtain the following expression for the MLE of θ_0.

$$\begin{aligned}
0 &= \frac{x_1}{\theta+2} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta} \\
&= x_1(1-\theta)\theta - (x_2+x_3)(\theta+2)\theta + x_4(\theta+2)(1-\theta) \\
&= x_1(\theta-\theta^2) - (x_2+x_3)(\theta^2+2\theta) + x_4(-\theta^2-\theta+2) \\
&= x_1\theta - x_1\theta^2 - x_2\theta^2 - 2x_2\theta - x_3\theta^2 - 2x_3\theta - x_4\theta^2 - x_4\theta + 2x_4 \\
&= -(x_1+x_2+x_3+x_4)\theta^2 + (x_1-2x_2-2x_3-x_4)\theta + 2x_4 \\
&= -n\theta^2 + (x_1-2x_2-2x_3-x_4)\theta + 2x_4 \quad (\text{Since } x_1+x_2+x_3+x_4 = n)
\end{aligned}$$

Hence as computed above, an expression for the MLE of θ_0 is $\widehat{\theta}_0^{mle} = \frac{-(x_1-2x_2-2x_3-x_4)\ \pm\ \sqrt{(x_1-2x_2-2x_3-x_4)^2-4(-n)(2x_4)}}{-2n}$. Plugging in our values of x_1, x_2, x_3, x_4 from above we can compute the MLE of θ_0 for our sample.

$$\begin{aligned}
\widehat{\theta}_0^{mle} &= \frac{-(x_1-2x_2-2x_3-x_4)\ \pm\ \sqrt{(x_1-2x_2-2x_3-x_4)^2-4(-n)(2x_4)}}{-2n} \\
&= \frac{-(1997-2(906+904)-32)\ \pm\ \sqrt{(1997-2(906+904)-32)^2+4(3839)(64)}}{-7678} \\
&= \frac{1655\ \pm\ \sqrt{2739025+982784}}{-7678} \\
&= \frac{1655\ \pm\ \sqrt{3721809}}{-7678} \\
&= -0.466814151681 \text{ or } 0.0357123022406
\end{aligned}$$

Since θ_0 represents a probability, it can't be less than 0, thus as computed above $\widehat{\theta}_0^{mle} = 0.0357123022406$.

b. Give an expression for the asymptotic standard error of $\widehat{\theta}_0^{mle}$ and calculate it.

By Theorem 25.1, under certain regulatory conditions (which we will discuss in part d), the asymptotic standard error of $\widehat{\theta}_0^{mle}$ takes the form $\frac{1}{\sqrt{-\ell''\left(\widehat{\theta}_0^{mle}\right)}}$. Hence, in order to give an expression for this standard error we must first find the second derivative of the log-likelihood function. This calculation is shown below.

$$\begin{aligned}\frac{d^2}{d\theta^2}\ell(\theta) &= \frac{d^2}{d\theta^2}\left(\frac{x_1}{\theta+2} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}\right) \\ &= \frac{-x_1}{(\theta+2)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2},\ 0 \leq \theta \leq 1\end{aligned}$$

As computed above the second derivative of the log-likelihood function of θ is $\ell''(\theta) = \frac{-x_1}{(\theta+2)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2},\ 0 \leq \theta \leq 1$. With that said, it follows that $-\ell''(\theta) = \frac{x_1}{(\theta+2)^2} + \frac{x_2+x_3}{(1-\theta)^2} + \frac{x_4}{\theta^2},\ 0 \leq \theta \leq 1$. Furthermore, it follows that $-\ell''(\widehat{\theta}_0^{mle}) = \frac{x_1}{\left(\widehat{\theta}_0^{mle}+2\right)^2} + \frac{x_2+x_3}{\left(1-\widehat{\theta}_0^{mle}\right)^2} + \frac{x_4}{\left(\widehat{\theta}_0^{mle}\right)^2}$. Lastly an expression for the asymptotic standard error of $\widehat{\theta}_0^{mle}$ is $\hat{SD}\left[\widehat{\theta}_0^{mle}\right] = \frac{1}{\sqrt{\frac{x_1}{\left(\widehat{\theta}_0^{mle}+2\right)^2} + \frac{x_2+x_3}{\left(1-\widehat{\theta}_0^{mle}\right)^2} + \frac{x_4}{\left(\widehat{\theta}_0^{mle}\right)^2}}}$. Below we will calculate $\hat{SD}\left[\widehat{\theta}_0^{mle}\right]$ for our given sample.

$$\begin{aligned}\hat{SD}\left[\widehat{\theta}_0^{mle}\right] &= \frac{1}{\sqrt{\frac{x_1}{\left(\widehat{\theta}_0^{mle}+2\right)^2} + \frac{x_2+x_3}{\left(1-\widehat{\theta}_0^{mle}\right)^2} + \frac{x_4}{\left(\widehat{\theta}_0^{mle}\right)^2}}} \\ &= \frac{1}{\sqrt{\frac{1997}{(0.0357123022406+2)^2} + \frac{906+904}{(1-0.0357123022406)^2} + \frac{32}{(0.0357123022406)^2}}} \\ &= 0.00602812039349\end{aligned}$$

Thus, as calculated above, the asymptotic standard error of $\widehat{\theta}_0^{mle}$ for our given plant sample is 0.00602812039349.

c. Calculate an approximate 95% Wald confidence interval for θ_0 and report it in context.

Another consequence of Theorem 25.1, is that $\widehat{\theta}_0^{mle}$ has an approximate normal distribution when n is large. From the previous statement, it follows that $\widehat{\theta}_0^{mle} \sim Norm\left(\theta_0, \frac{1}{\sqrt{\frac{x_1}{\left(\widehat{\theta}_0^{mle}+2\right)^2} + \frac{x_2+x_3}{\left(1-\widehat{\theta}_0^{mle}\right)^2} + \frac{x_4}{\left(\widehat{\theta}_0^{mle}\right)^2}}}\right)$. From Theorem 25.1 it follows that a 95% Wald confidence interval for θ_0 takes the form $\widehat{\theta}_0^{mle} \pm z_{\alpha/2}\frac{1}{\sqrt{-\ell''(\theta)}}$. Hence in our case we can see that the 95% Wald confidence interval for θ_0 is $0.035712 \pm 1.96 \cdot 0.00602$ which leaves us with our final interval of [0.023897, 0.047527].

As shown above, the 95% Wald confidence interval of θ_0 is [0.023897, 0.047527]. This means that over repeated samples of size 3839, 95% of intervals constructed in this fashion will contain the true probability related to linkage closeness. Another possible interpretation is that, we are 95% confident that the true probability related to linkage closeness is in the range [0.023897, 0.047527].

d. Is there any reason to be concerned about the approximation in part c? Make a relevant plot and comment.

Theorem 25.1 states that given an independent sample of size n then under certain regulatory conditions, the MLE estimator is approximately normal and this fact can be used to create confidence intervals like we

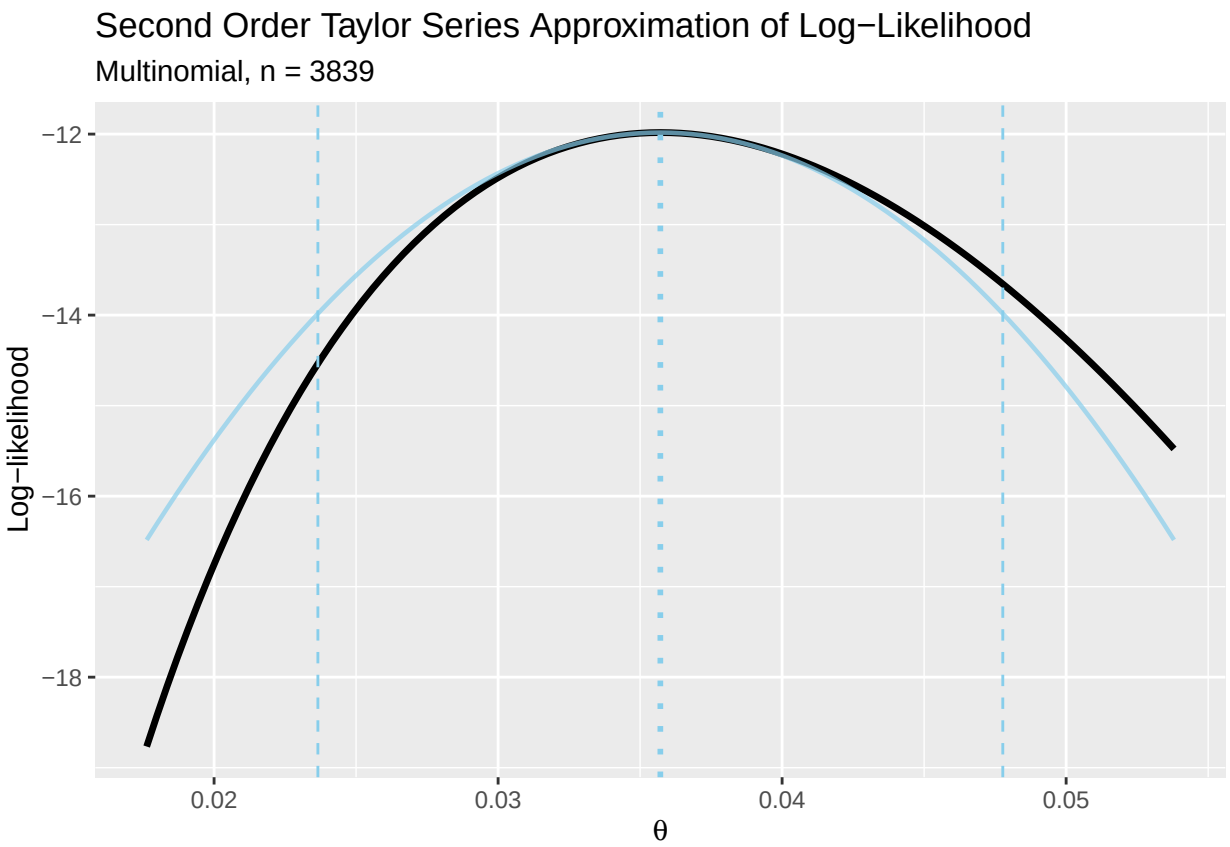
did in part c. However, this fact relies on a large sample and certain regulatory conditions. The regulatory conditions include, but are not limited to: the possible values that our random variable can take on can't depend on the parameter, and lastly, the true value of the parameter can't fall on the boundary points of that parameter.

To assess the relevance of the results from part c we will plot the log-likelihood function with its corresponding second-order Taylor series approximation around the MLE and see how well the Taylor series approximates the log-likelihood function.

```
loglik_multinom <- function(theta, x) {
  pi1 <- 1/4 * (2 + theta)
  pi2 <- 1/4 * (1 - theta)
  pi3 <- 1/4 * (1 - theta)
  pi4 <- 1/4 * theta
  ifelse(theta < 0 | theta > 1,
         NA,
         dmultinom(x, prob = c(pi1, pi2, pi3, pi4), log = TRUE)
         )
}

second_order <- maxLik2(loglik = loglik_multinom,
                        start = 0.036,
                        method = "NR",
                        x = c(1997, 906, 904, 32))

plot(second_order) %>%
  gf_labs(title = "Second Order Taylor Series Approximation of Log-Likelihood",
          subtitle = "Multinomial, n = 3839",
          x = expression(theta),
          y = "Log-likelihood")
```



As can be seen above, the second order Taylor series approximation of the log-likelihood function of θ around the MLE has an okay fit despite the very large sample size. The reason why this fit isn't better than what we'd expected for such a large sample size is because the true value of θ_0 lies very close to the boundary point of 0. With that said, due to the fact that a regulatory condition may be violated there is a reason to be concerned about the approximation from part c.

2. (Discrete X) Suppose X is a discrete random variable with PMF $f(x)$ indexed by a parameter θ as shown below.

θ_0	$x=1$	$x=2$	$x=3$	$x=4$
1	1/3	1/6	1/12	5/12
2	1/2	1/4	1/6	1/12
W	0.811	0.811	1.386	0

a. Say we want to test $H_0 : \theta_0 = 1$ versus $H_1 : \theta_0 \neq 1$. Calculate the Likelihood Ratio Statistic W for each value of x and write it in the last row of the table. Briefly explain your work below. (Hint: The parameter θ_0 can only take two values, so you can find the MLE fairly easily)

If we suppose X is a discrete random variable with PMF $f(x)$ indexed by a parameter θ as shown above, and want to test $H_0 : \theta_0 = 1$ versus $H_1 : \theta_0 \neq 1$. Then to do this we will want to calculate the Likelihood Ratio Statistic W for each value of x. Below we calculate W for each value of x by deciding what the MLE of θ_0 is, which should be easy since θ_0 can only take two values.

Calculating W when x = 1:
When $x = 1$, $L(\theta) = \frac{1}{3}$ when $\theta = 1$ and $L(\theta) = \frac{1}{2}$ when $\theta = 2$. Hence $\widehat{\theta}_0^{mle} = 2$. With that said, $W = 2\ln\left[\frac{L\left(\widehat{\theta}_0^{mle}\right)}{L\left(\theta_0^{null}\right)}\right]$, which in our case $W = 2\ln\left[\frac{L(2)}{L(1)}\right] = 2\ln\left[\frac{1/2}{1/3}\right] = 0.81093$

Calculating W when x = 2:
When $x = 2$, $L(\theta) = \frac{1}{6}$ when $\theta = 1$ and $L(\theta) = \frac{1}{4}$ when $\theta = 2$. Hence $\widehat{\theta}_0^{mle} = 2$. With that said, $W = 2\ln\left[\frac{L\left(\widehat{\theta}_0^{mle}\right)}{L\left(\theta_0^{null}\right)}\right]$, which in our case $W = 2\ln\left[\frac{L(2)}{L(1)}\right] = 2\ln\left[\frac{1/4}{1/6}\right] = 0.81093$

Calculating W when x = 3:
When $x = 3$, $L(\theta) = \frac{1}{12}$ when $\theta = 1$ and $L(\theta) = \frac{1}{6}$ when $\theta = 2$. Hence $\widehat{\theta}_0^{mle} = 2$. With that said, $W = 2\ln\left[\frac{L\left(\widehat{\theta}_0^{mle}\right)}{L\left(\theta_0^{null}\right)}\right]$, which in our case $W = 2\ln\left[\frac{L(2)}{L(1)}\right] = 2\ln\left[\frac{1/6}{1/12}\right] = 1.38629$

Calculating W when x = 4:
When $x = 4$, $L(\theta) = \frac{5}{12}$ when $\theta = 1$ and $L(\theta) = \frac{1}{12}$ when $\theta = 2$. Hence $\widehat{\theta}_0^{mle} = 1$. With that said, $W = 2\ln\left[\frac{L\left(\widehat{\theta}_0^{mle}\right)}{L\left(\theta_0^{null}\right)}\right]$, which in our case $W = 2\ln\left[\frac{L(1)}{L(1)}\right] = 2\ln\left[\frac{5/12}{5/12}\right] = 0$

b. Write the sampling distribution of W below in tabular form assuming H_0 is true. (Hint: W is a discrete random variable)

Since W is discrete and only takes on 3 different values, the sampling distribution of W under the assumption of the null hypothesis can be written in tabular form. The three possible values of W are 0, 0.811, and 1.386, with probabilities of $\frac{5}{12}, \frac{1}{2}$, and $\frac{1}{2}$, respectively. Hence the sampling distribution of W in tabular form assuming H_0 is true is:

w	0	0.811	1.386
$f(w)$	$\frac{5}{12}$	$\frac{1}{2}$	$\frac{1}{12}$

c. Suppose we observe $x = 3$. Calculate the P-value. What should we conclude at a 0.05 level of significance?

If we observe $x = 3$, then our observed W statistic value is 1.386. Looking at the sampling distribution of W we can see that $P(W = 1.386) = \frac{1}{12}$, since there are no other values as or more extreme than this we conclude that our p-value is $\frac{1}{12}$. Since $\frac{1}{12} > 0.05$, it follows that we fail to reject the null at the 5% level of significance. Thus we do not have evidence to say that the true value of θ_0 is not 1.

3. (Likelihood ratio) Suppose $X_1, X_2, \ldots, X_n$ are an *i.i.d.* sample from PDF

$$f(x) = (\theta_0 + 1)x^{\theta_0}, \quad 0 < x < 1,$$

where $\theta_0 > -1$.

a. Determine the form of the Likelihood Ratio Test statistic W for testing

$$H_0 : \theta_0 = 0, \quad H_1 : \theta_0 \neq 0$$

assuming a sample $x_1, x_2, \ldots, x_n$.

To make grading easier, please clearly indicate each of the following:

- Log likelihood function $\ell(\theta)$

- Expression for the MLE $\widehat{\theta}_0^{mle}$ (no need to verify that it is a maximum)
- Expression for the likelihood ratio statistic simplified as much as possible

To determine the form of the Likelihood Ratio Test statistic W for testing $H_0 : \theta_0 = 0$ versus $H_1 : \theta_0 \neq 0$ assuming a sample $x_1, x_2, \ldots, x_n$, we must find the likelihood function, take the natural log of the likelihood function, find the critical point of the log-likelihood function to determine the MLE of θ_0, then find the second derivative of the log-likelihood function. First off, we will compute the likelihood function as shown below.

$$\begin{aligned} L(\theta) &= f(x_1) \times f(x_2) \times \cdots \times f(x_n) \\ &= (\theta+1)x_1^{\theta} \times (\theta+1)x_2^{\theta_0} \times \cdots \times (\theta+1)x_n^{\theta} \\ &= (\theta+1)^n (x_1 \times x_2 \times \cdots \times x_n)^{\theta},\ \theta > -1 \end{aligned}$$

Thus, as computed above, the likelihood function of θ is $L(\theta) = (\theta+1)^n (x_1 \times x_2 \times \cdots \times x_n)^{\theta}$, $\theta > -1$. However, notice that taking the derivative of this function will not be easy, thus we will take the natural log of the likelihood function to turn multiplication into addition. This process is shown below.

$$\begin{aligned} \ell(\theta) &= \ln(L(\theta)) \\ &= \ln((\theta+1)^n (x_1 \times x_2 \times \cdots \times x_n)^{\theta}) \\ &= \ln((\theta+1)^n) + \ln((x_1 \times x_2 \times \cdots \times x_n)^{\theta}) \\ &= n\ln(\theta+1) + \theta \sum_{i=1}^{n} \ln(x_i),\ \theta > -1 \end{aligned}$$

As computed above, the log-likelihood function is $\ell(\theta) = n\ln(\theta+1) + \theta \sum_{i=1}^{n} \ln(x_i)$, $\theta > -1$. Now we must find the critical points of this function in order to find the candidates for the maximum likelihood estimator of θ_0. This derivative calculation is shown below.

$$\begin{aligned} \frac{d}{d\theta}\ell(\theta) &= \frac{d}{d\theta}\left(n\ln(\theta+1) + \theta \sum_{i=1}^{n} \ln(x_i) \right) \\ &= \frac{n}{\theta+1} + \sum_{i=1}^{n} \ln(x_i),\ \theta > -1 \end{aligned}$$

In order to find the critical points, we must find the values of θ such that this derivative is equal to zero.

$$\begin{aligned} 0 &= \frac{n}{\theta+1} + \sum_{i=1}^{n} \ln(x_i) \\ \frac{-n}{\theta+1} &= \sum_{i=1}^{n} \ln(x_i)) \\ -n &= (\theta+1) \sum_{i=1}^{n} \ln(x_i) \\ -n &= \theta \sum_{i=1}^{n} \ln(x_i) + \sum_{i=1}^{n} \ln(x_i) \\ -n - \sum_{i=1}^{n} \ln(x_i) &= \theta \sum_{i=1}^{n} \ln(x_i) \\ \theta &= \frac{-n - \sum_{i=1}^{n} \ln(x_i)}{\sum_{i=1}^{n} \ln(x_i)} \end{aligned}$$

As computed above, the critical point $\theta = \dfrac{-n-\sum_{i=1}^{n}\ln(x_i)}{\sum_{i=1}^{n}\ln(x_i)}$ is a candidate for the maximum likelihood estimator for θ_0 In Homework 4.1.b we proved this was the MLE, so we will say $\widehat{\theta}_0^{mle} = \dfrac{-n-\sum_{i=1}^{n}\ln(x_i)}{\sum_{i=1}^{n}\ln(x_i)}$.

Now we can find an expression for W of the form $2\left(\ell\left(\widehat{\theta}_0^{mle}\right) - \ell\left(\theta_0^{null}\right)\right)$. This expression is derived below.

$$\begin{aligned} W &= 2\left(\ell\left(\widehat{\theta}_0^{mle}\right) - \ell\left(\theta_0^{null}\right)\right) \\ &= 2\left[n\ln\left(\widehat{\theta}_0^{mle}+1\right) + \widehat{\theta}_0^{mle}\sum_{i=1}^{n}\ln(x_i) - n\ln\left(\theta_0^{null}+1\right) - \theta_0^{null}\sum_{i=1}^{n}\ln(x_i)\right] \end{aligned}$$

However, from the problem description we know that $\theta_0^{null} = 0$ and $\widehat{\theta}_0^{mle} = \dfrac{-n-\sum_{i=1}^{n}\ln(x_i)}{\sum_{i=1}^{n}\ln(x_i)}$. Hence we can simplify our expression for W even further.

$$\begin{aligned} W &= 2\left[n\ln\left(\frac{-n-\sum_{i=1}^{n}\ln(x_i)}{\sum_{i=1}^{n}\ln(x_i)}+1\right) + \frac{-n-\sum_{i=1}^{n}\ln(x_i)}{\sum_{i=1}^{n}\ln(x_i)}\sum_{i=1}^{n}\ln(x_i) - n\ln(0+1) - 0\cdot\sum_{i=1}^{n}\ln(x_i)\right] \\ &= 2\left(n\ln\left(\frac{-n}{\sum_{i=1}^{n}\ln(x_i)}-1+1\right) - n - \sum_{i=1}^{n}\ln(x_i)\right) \end{aligned}$$

Thus as computed above, a simplified expression for the Likelihood Ratio Test statistic W is $W = 2\left(n\ln\left(\frac{-n}{\sum_{i=1}^{n}\ln(x_i)}-1+1\right) - n - \sum_{i=1}^{n}\ln(x_i)\right)$.

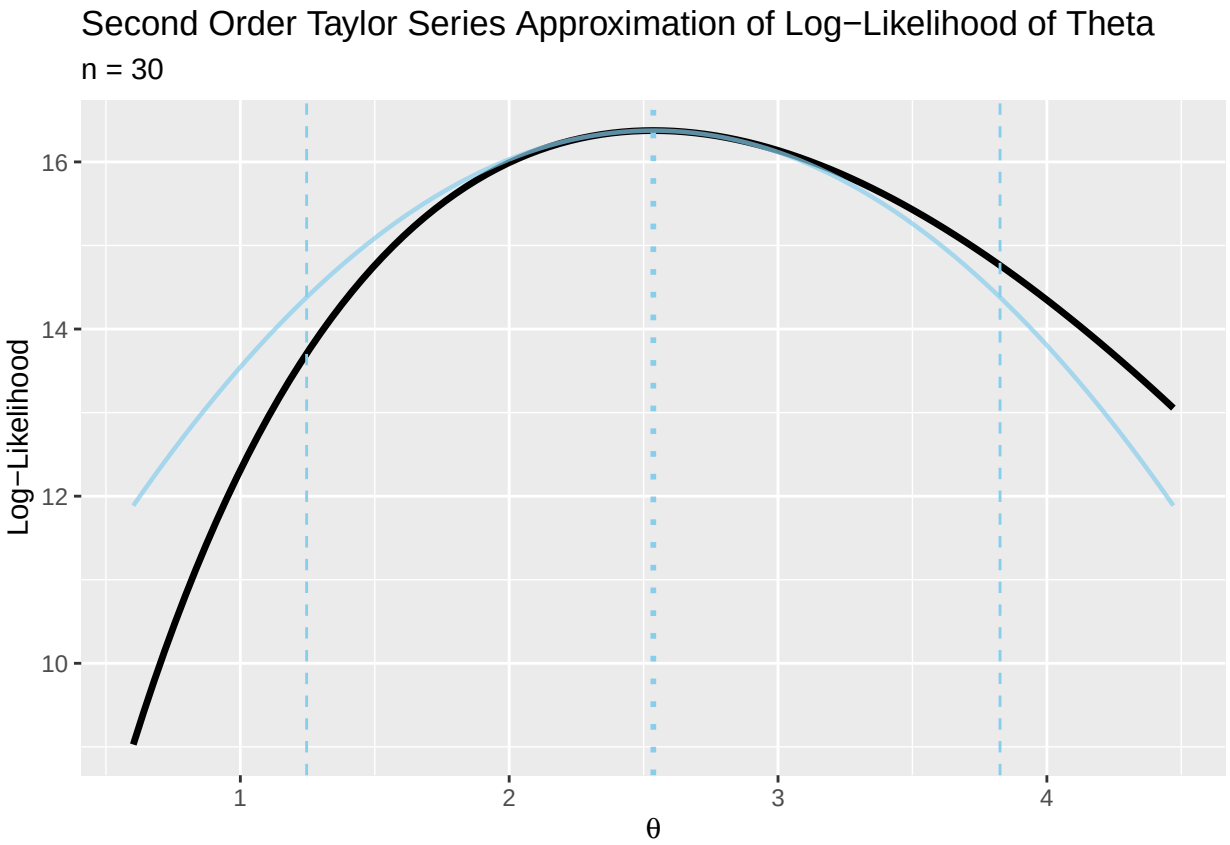
b. The 30 values below are a random sample from this distribution for some true (unknown) value θ_0. Calculate the likelihood ratio statistic for this data. (Write all your code for the remaining parts in the code chunk labeled `lik-ratio` but show the code in the Appendix. Report answers (rounded to 4 digits) using inline code.

As computed from the 30 given data points, the MLE of θ_0 is 2.5362. Thus our observed W value is 32.7502.

c. Calculate the P-value for the Likelihood Ratio Test statistic using the approximate chi square distribution. Is there reason to be concerned about using the chi-squared distribution? Compare the log likelihood function with its quadratic approximation.

Using the approximate chi square distribution of the Likelihood Ratio Test statistic, the p-value for our observed value of W is 1.0479×10^{-8}, which is essentially 0.

Below we will compare the log likelihood function with its quadratic approximation to see if there is reason to be concerned about using the chi-squared distribution for W



As can be seen above, the second order Taylor series approximation of $\ell(\theta)$ around the MLE of θ_0 is an okay approximation. Since the approximation gets weaker and weaker as θ gets both smaller and larger, it is reasonable to be skeptical about using the chi squared distribution for W. Although it is important to note that this approximation is by no means bad for only having a sample size of $n = 30$.

d. An alternative to using the chi squared distribution to estimate the p-value is to calculate an empirical p-value by generating a large number B of samples from the null hypothesis. Follow the steps below to calculate an empirical p-value.

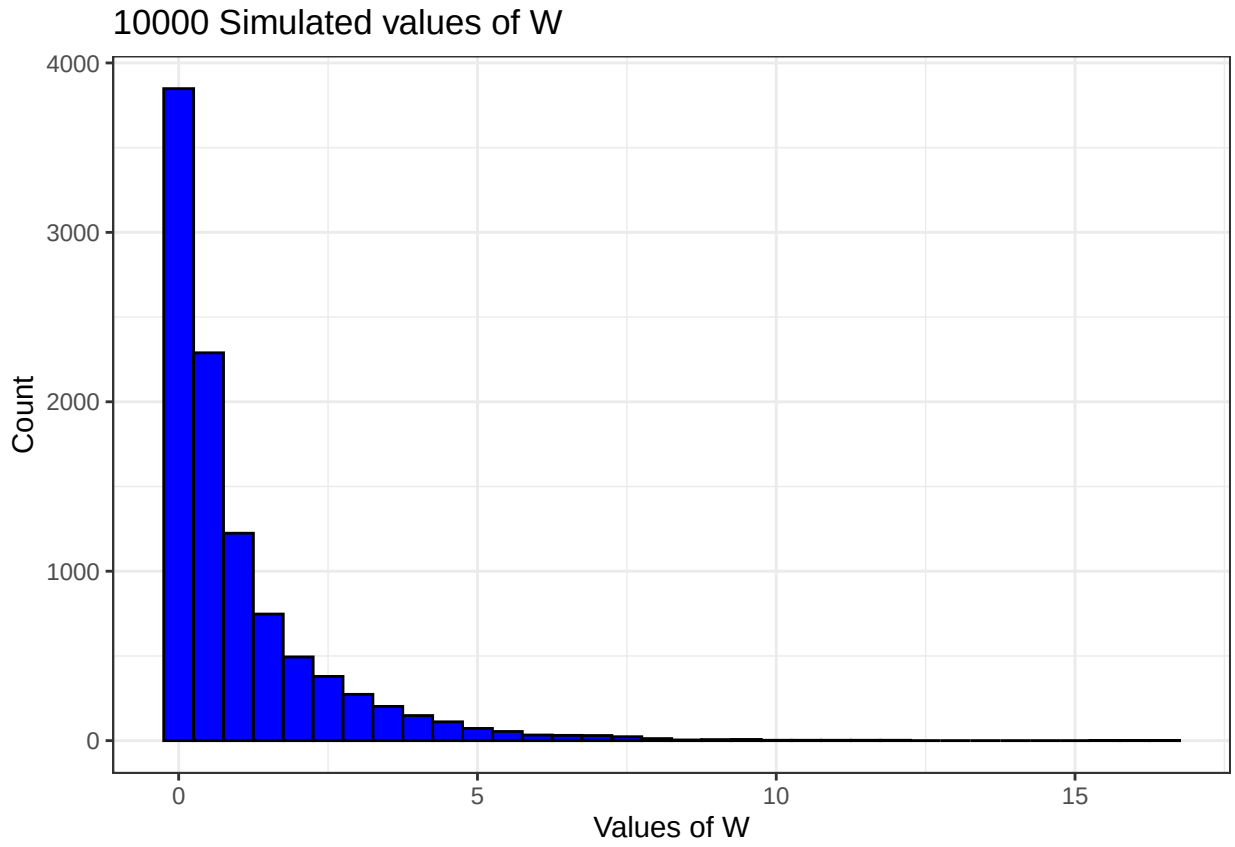
- Step 0: Set the random number seed to 414.
- Step 1: Generate $x_1^*, x_2^*, \ldots, x_{30}^* \overset{i.i.d.}{\sim} Unif(0, 1)$. (why?)
- Step 2: For the generated sample, calculate the value of the MLE $\widehat{\theta}_0^*$ and the likelihood ratio test statistic w^*.
- Step 3: Repeat steps 1 and 2 a large number $B = 10000$ times. (Don't forget to cache the code chunk `lik-ratio`)
- Step 4: Count the fraction of times that w^* from the generated samples exceeds the w we observed. Report the empirical P-value and also make a histogram of the values of w^*. (Don't forget those labels and title)

As computed above, there were no times in which w^* from the generated samples exceeded the w we observed. Hence out empirical p-value is thus 0. Below we will plot a histogram of our simulated w^* values.

However as introduced in class/office hours, since we observed a p-value of 0, there is one work around to this that will allow us to obtain a non-zero p-value. Since we already know that we have one W values that is at least as big as the W we observed, which is the W we observed, we can think of our simulated p-value

as the number of simulated W values that are greater than our observed W value plus 1 over the number of simulated values plus 1. Thus we could say out p-value is actually $\frac{1}{10001} = 0.0001$.

```
hist
```



As can be seen from the above histogram of the 10000 simulated W values, the distribution is heavily left skewed with the majority of the density at smaller values of W. This is exactly what we would expect from a chi square distribution with 1 degree of freedom.

Appendix

Code for problem 1

```
# Part b:

# Read in the data:
x<-c(0.64,0.92,0.73,0.96,0.98,0.33,0.80,0.96,0.81,0.76,0.98,0.75,
     0.87,0.82,0.44,0.96,0.61,0.32,0.67,0.98,0.96,0.88,0.85,1.00,
     0.86,0.88,0.80,0.83,0.64,0.5)

# Compute sample size:
n <- length(x)

# Find the MLE of the above sample:
theta_mle <- (-n - sum(log(x))) / sum(log(x))

# Find the W statistic from the above sample:
```

```r
w <- 2*n*log(-n / sum(log(x))) - 2*n - 2*sum(log(x))


# Part c:

# Find the p-value of the above W statistic:
p_val <- pchisq(q = w, df = 1, lower.tail = F)

loglik_theta <- function(theta, sample) {
  ifelse(theta <= -1,
         NA,
         length(sample) * log(theta + 1) + theta * sum(log(sample))
         )
}

theta_second <- maxLik2(loglik = loglik_theta,
                        start = theta_mle,
                        method = "NR",
                        sample = x)

plot <- plot(theta_second) %>%
  gf_labs(title = "Second Order Taylor Series Approximation of Log-Likelihood of Theta",
          subtitle = "n = 30",
          x = expression(theta),
          y = "Log-Likelihood")

# Part d:
set.seed(414)

# Step 3: Repeat Steps 1 and 2 10000 times:
B = 10000

w_sim <- lapply(1:B, FUN = function(i) {
    # Step 1: Generate 30 samples from Unif(0, 1)
    sample = runif(n = 30, min = 0, max = 1)

    # Step 2: Generate the value of the MLE:
    theta = (-length(sample) - sum(log(sample))) / (sum(log(sample)))

    # Step 2: Generate the value of W:
    w_stat = 2*(loglik_theta(theta, sample) - loglik_theta(0, sample))

    # Make a data frame:
    data.frame(w_obs = w_stat)
})

# Combine the list of dataframes:
sim_w <- do.call(rbind, w_sim)

# Step 4 i: Count the fraction of times that the w from the generated samples
# exceeds the w we observed:
sim_p_val <- sum(sim_w$w_obs >= w) / B
```

```r
# Step 4 ii: Make a histogram of the values of w:
hist <- ggplot(data = sim_w) +
  geom_histogram(mapping = aes(x = w_obs), binwidth = 0.5,
                 color = "black", fill = "blue") +
  labs(title = "10000 Simulated values of W",
       x = "Values of W",
       y = "Count") +
  theme_bw()
```