Homework 6

Spring 2023

Tanner Huck

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 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 $\widehat{\theta}_0^{mle}$.

Let x_1 be the number of plants with the trait combination starchy-green, x_2 the number of starchy-white, x_3 sugary-green, and x_4 sugary-white. The probability of a plant with starchy green is $\frac{1}{4}(2+\theta_0)$, $\frac{1}{4}(1-\theta_0)$ for starchy-white and sugary-green, and $\frac{1}{4}\theta_0$ for sugary-white, where θ_0 is a probability related to linkage closeness. We want to determine the MLE of θ_0.

First finding the likelihood function of θ using the results shown in the table above,

$$\begin{aligned} L(\theta) &= (\frac{1}{4}(2+\theta))^{1997} \times (\frac{1}{4}(1-\theta))^{906+904} \times (\frac{1}{4}\theta)^{32} \times \frac{(1997+906+904+32)!}{1997!906!904!32!} \\ &= (\frac{1}{4}(2+\theta))^{1997} \times (\frac{1}{4}(1-\theta))^{1810} \times (\frac{1}{4}\theta)^{32} \times \frac{(1997+906+904+32)!}{1997!906!904!32!} \\ &\text{for } 0 \leq \theta \leq 1 \end{aligned}$$

Then finding the log-likelihood function,

$$\begin{aligned}\ell(\theta) &= ln((\frac{1}{4}(2+\theta))^{1997} \times (\frac{1}{4}(1-\theta))^{1810} \times (\frac{1}{4}\theta)^{32} \times \frac{(1997+906+904+32)!}{1997!906!904!32!}) \\ &= 1997ln((\frac{1}{4}(2+\theta))) + 1810ln((\frac{1}{4}(1-\theta)) + 32ln((\frac{1}{4}\theta) + ln(\frac{(1997+906+904+32)!}{1997!906!904!32!})) \\ &\text{for } 0 \leq \theta \leq 1\end{aligned}$$

Then finding the derivative w.r.t. θ,

$$\begin{aligned}\frac{d}{d\theta}\ell(\theta) &= \frac{d}{d\theta}1997ln((\frac{1}{4}(2+\theta))) + 1810ln((\frac{1}{4}(1-\theta)) + 32ln((\frac{1}{4}\theta)) \\ &= \frac{1997}{2+\theta} - \frac{1810}{1-\theta} + \frac{32}{\theta}\end{aligned}$$

Then setting the derivative equal to 0 and solving for θ,

$$\begin{aligned}\frac{1997}{2+\theta} - \frac{1810}{1-\theta} + \frac{32}{\theta} &= 0 \\ (1997 \cdot (1-\theta) \cdot \theta) + (-1810 \cdot (2+\theta) \cdot (\theta)) + (32 \cdot (1-\theta) \cdot (2+\theta)) &= 0 \\ -3839\theta^2 - 1655\theta + 64 &= 0 \\ &\text{using the quadratic formula we get,} \\ \theta &= \frac{-1655}{7678} \pm \frac{\sqrt{3721809}}{7678} \\ \theta &\approx -0.4668 \text{ and } 0.0357\end{aligned}$$

Now finding the second derivative w.r.t. θ,

$$\begin{aligned}\frac{d^2}{d\theta^2}\ell(\theta) &= \frac{d}{d\theta}\frac{1997}{2+\theta} - \frac{1810}{1-\theta} + \frac{32}{\theta} \\ &= -\frac{1997}{(2+\theta)^2} - \frac{1810}{(1-\theta)^2} - \frac{32}{\theta^2}\end{aligned}$$

Since the probability cannot be negative, our critical point is about $\theta = 0.0357$. From the second derivative test, we can see that the second derivative is less than 0 for any value of θ, thus our critical point is a global maximum. Hence $\widehat{\theta}_0^{mle} \approx 0.0357$.

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

Now we want to find the asymptotic standard error of $\widehat{\theta}_0^{mle}$. First we need to assume that each offspring plant is independent, meaning any plants starchiness and color will not affect the starch or color of any other plant. We also need to assume that our plants are indexed by the true unknown θ_0 with a large value of n and regulatory conditions are met. Then the estimated standard error is,

$$\frac{1}{\sqrt{-\ell(\widehat{\theta}_0^{mle})}}$$

Thus plugging in what we found for the second derivative in part a, we have

$$\begin{aligned}
\frac{1}{\sqrt{-\ell''(\widehat{\theta}_0^{mle})}} &= \frac{1}{\sqrt{-(-\frac{1997}{(2+\theta)^2} - \frac{1810}{(1-\theta)^2} - \frac{32}{\theta^2})}} \\
&= \frac{1}{\sqrt{\frac{1997}{(2+\theta)^2} + \frac{1810}{(1-\theta)^2} + \frac{32}{\theta^2}}} \\
&\text{substituting in our value of } \theta \approx 0.0357 \\
&= \frac{1}{\sqrt{\frac{1997}{(2+0.0357)^2} + \frac{1810}{(1-0.0357)^2} + \frac{32}{0.0357^2}}} \\
&\approx 0.00603
\end{aligned}$$

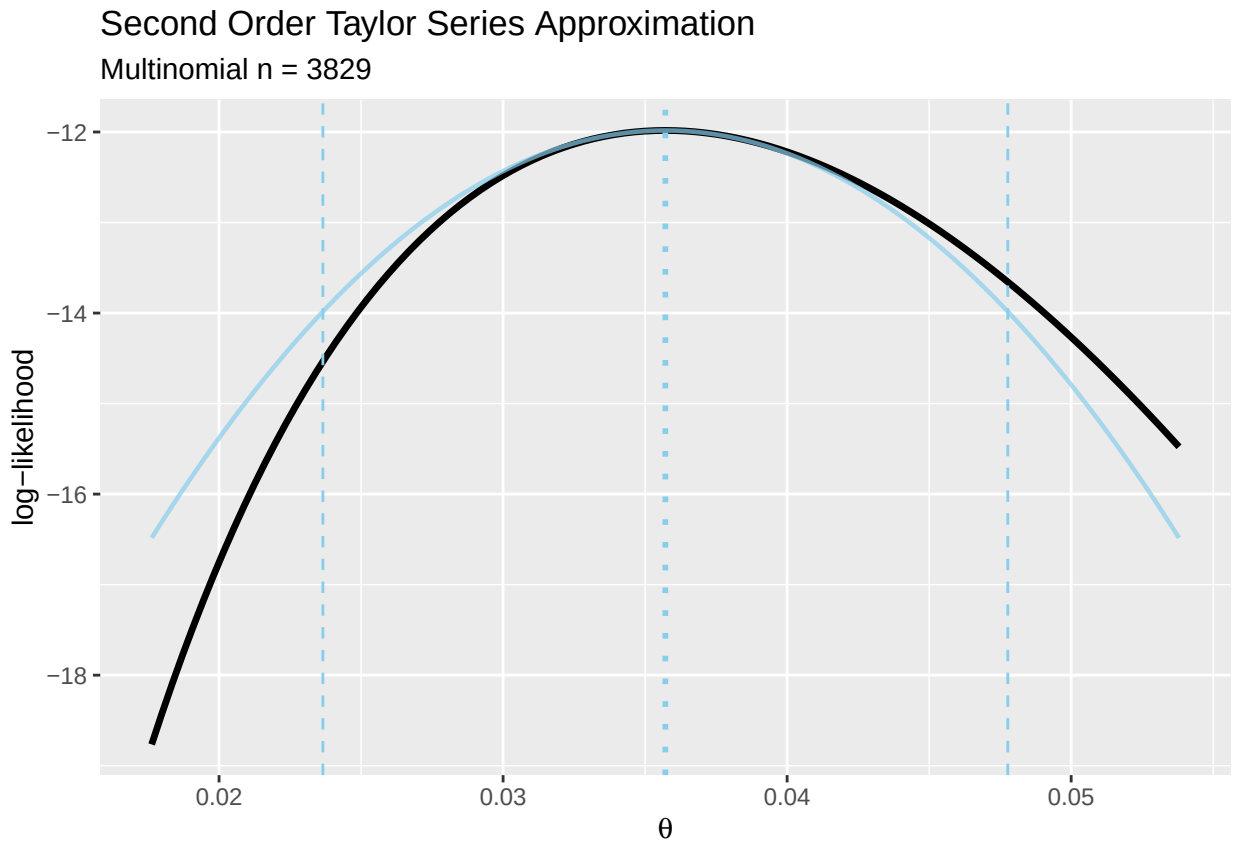
Hence the asymptotic standard error of $\widehat{\theta}_0^{mle}$ is about 0.00603.

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

The 95% Wald confidence interval for θ_0 is given by $\widehat{\theta}_0^{mle} \pm z_{\alpha/2}\widehat{SE}$. Plugging in our found values for $\widehat{\theta}_0^{mle}$ and $\widehat{SE}$ we get about $0.0357 \pm 1.96 \times 0.00603$ or about [0.0239, 0.0475]. This tells us that we are 95% confident that the true probability of θ_0, the probability related to linkage closeness, is in the range [0.0239, 0.0475].

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

```
multi_plot
```



From this graph we can see the second order Taylor Series approximation for the log likelihood of θ. We can see that the fit is fairly good. It does a good job of approximating around θ values between 0.03 and 0.04,

but not as good for smaller and larger *theta* values. Since we have a very large $n = 3829$ and we still do not have a great approximation near the tails we may be concerned about the approximation in 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)

Recall that the likelihood ratio test statistic is given by $W = 2ln[\frac{L(\widehat{\theta_0}^{mle})}{L(\widehat{\theta_0}^{null})}]$. Thus we can calculate this statistic for each value of x in the table above. Note that under the null hypothesis, $\theta_0^{null} = 1$. Thus the parameter θ_0 can only take two values. Hence,

For $x = 1$, $L(\theta) = \frac{1}{2}$ because from the table we can see that likelihood is either 1/3 or 1/2 and 1/2 is larger. Thus for $x = 1$, $W = 2 \cdot ln[\frac{L(2)}{L(1)}] = 2 \cdot ln[\frac{1/2}{1/3}] \approx 0.811$.

Then following similar reasoning for the remainder of the x values.

$$x = 2 \rightarrow W = 2 \cdot ln[\frac{L(2)}{L(1)}] = 2 \cdot ln[\frac{1/4}{1/6}] \approx 0.811$$

$$x = 3 \rightarrow W = 2 \cdot ln[\frac{L(2)}{L(1)}] = 2 \cdot ln[\frac{1/6}{1/12}] \approx 1.386$$

$$x = 4 \rightarrow W = 2 \cdot ln[\frac{L(1)}{L(1)}] = 2 \cdot ln[\frac{5/12}{5/12}] = 0$$

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

When the H_0 is true, $\theta_0 = 1$. Thus creating a table of the sampling distribution of W, from what we calculated in part a,

w	1.386	0.811	0
f(w)	1/12	1/2	5/12

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

Given that we observe $x = 3$, we know that $W \approx 1.386$, from part a. Then from part b we know that the probability of obtaining this value for W is $\frac{1}{12}$. Thus under the null hypothesis that $\theta_0 = 1$, the p-value of observing $x = 3$ is $\frac{1}{12}$. Since $\frac{1}{12} > 0.05$ we have do not have sufficient evidence to reject the null hypothesis. Hence we do not have enough evidence to conclude that $\theta_0 \neq 1$.

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

Given that $X_1, \ldots, X_n$ are an iid sample from the PDF $f(x) = (\theta_0 + 1)x^{\theta_0}$, we can first find the likelihood function of θ for $0 < x < 1$ and $\theta > -1$.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} (\theta + 1) x_i^{\theta} \\ &= (\theta + 1)^n \prod_{i=1}^{n} x_i^{\theta} \end{aligned}$$

Then we can calculate the log-likelihood function as,

$$\begin{aligned} \ell(\theta) &= ln((\theta + 1)^n \prod_{i=1}^{n} x_i^{\theta}) \\ &= nln(\theta + 1) + \theta \sum_{i=1}^{n} ln(x_i) \end{aligned}$$

Then finding the derivative w.r.t. θ,

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

Setting the derivative equal to 0 and solving for θ,

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

Hence $\widehat{\theta}_0^{mle} = -\frac{n}{\sum_{i=1}^{n} ln(x_i)} - 1$.

Given that the null hypothesis is that $\theta_0 = 0$, we now want to find the likelihood ratio statistic. Starting

from the definition of the likelihood ratio statistic,

$$W = 2ln[\frac{L(\widehat{\theta_0}^{mle})}{L(\widehat{\theta_0}^{null})}] = 2[\ell(\widehat{\theta_0}^{mle}) - \ell(\widehat{\theta_0}^{null})]$$

substituting in our values we found, and under the null hypothesis

$$= 2[\ell(-\frac{n}{\sum_{i=1}^{n} ln(x_i)} - 1) - \ell(0)]$$

plugging these into the log-liklihood we defined above

$$= 2[nln(-\frac{n}{\sum_{i=1}^{n} ln(x_i)} - 1 + 1) + (-\frac{n}{\sum_{i=1}^{n} ln(x_i)} - 1)\sum_{i=1}^{n} ln(x_i) - nln(0+1) + 0\sum_{i=1}^{n} ln(x_i)]$$

$$= 2[nln(-\frac{n}{\sum_{i=1}^{n} ln(x_i)}) + (-\frac{n}{\sum_{i=1}^{n} ln(x_i)} - 1)\sum_{i=1}^{n} ln(x_i)]$$

$$= 2[nln(-\frac{n}{\sum_{i=1}^{n} ln(x_i)}) + (-\frac{n\sum_{i=1}^{n} ln(x_i)}{\sum_{i=1}^{n} ln(x_i)} - \sum_{i=1}^{n} ln(x_i))]$$

$$= 2[nln(-\frac{n}{\sum_{i=1}^{n} ln(x_i)}) - n - \sum_{i=1}^{n} ln(x_i))]$$

$$= 2nln(-\frac{n}{\sum_{i=1}^{n} ln(x_i)}) - 2n - 2\sum_{i=1}^{n} ln(x_i))$$

Thus the Log liklihood function is $nln(\theta+1)+\theta\sum_{i=1}^{n} ln(x_i)$, the expression for the MLE $\widehat{\theta}_0^{mle}$ is $-\frac{n}{\sum_{i=1}^{n} ln(x_i)}-1$ and the likelihood ratio statistic is $2nln(-\frac{n}{\sum_{i=1}^{n} ln(x_i)}) - 2n - 2\sum_{i=1}^{n} ln(x_i))$.

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.

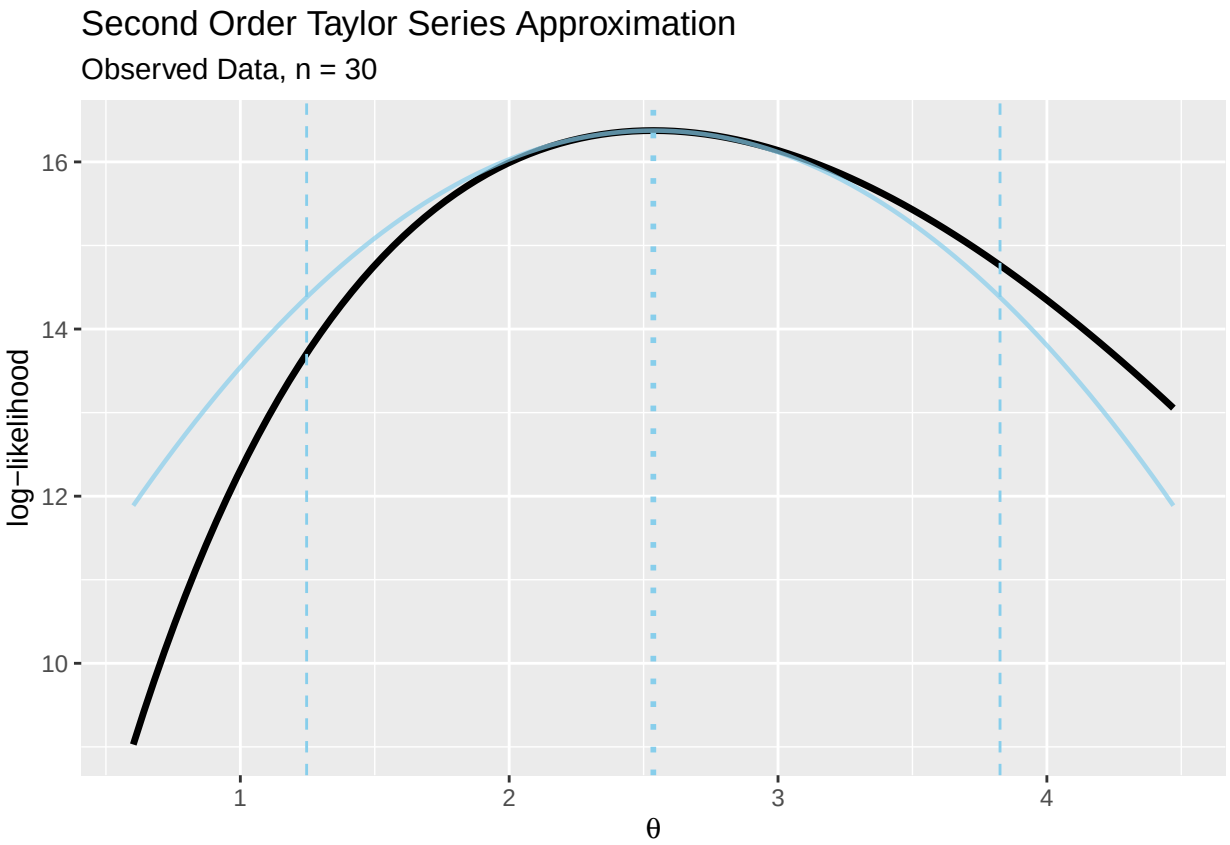
Using the values above, we can calculate the likelihood ratio statistic to be about 32.7502.

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 our likelihood ratio statistic, we can calculate the p-value using the approximate chi square distribution. From our R calculations we obtain a p-value of about 1.0479322×10^{-8}.

Then making a graph to compare the log likelihood function with its quadratic approximation.

```
x_plot
```



From this plot of the second order Taylor series approximation of the likelihood function, we can see that it is an ok approximation of the MLE for θ_0. Just like the previous graph, the approximation is not as good for the tails. However, for this new graph we only have sample of $n = 30$, thus because we do not have as large of a sample, we can more easily overlook this and say that the fit does a decent job of approximating the MLE.

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.

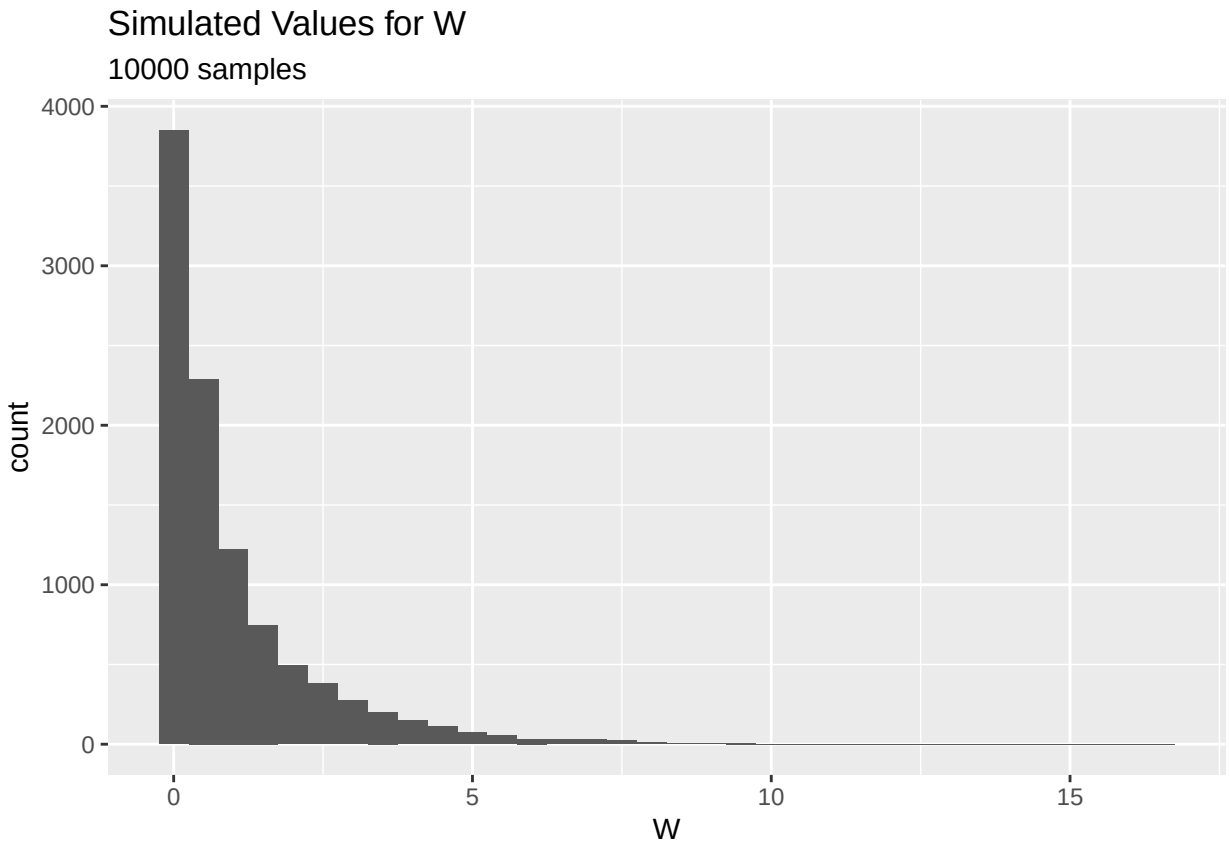
- Step 0: Set the random number seed to 414.
- Step 1: Generate $x_1^*, x_2^*, \ldots, x_{30}^* \overset{i.i.d.}{\sim} Unif(0, 1)$. (why?)
- Step 2: For the generated sample, calculate the value of the MLE $\widehat{\theta}_0^*$ and the likelihood ratio test statistic w^*.
- Step 3: Repeat steps 1 and 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)

Following the above steps,

```
empirical_p_val
```

```
## [1] 0
```

```
hist
```



We obtain a empirical p-value of 0. This tells us that there are 0 cases where the generated sample was larger than the observed w. From the histogram, we can see that the simulated W values are skewed to the right.

Appendix

Code for problem 1

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

ml.multi <- maxLik2(loglik = loglik_fun,
                        start = 0.001,
                        method = "NR",
                        x = c(1997, 906, 904, 32))

multi_plot <- plot(ml.multi) %>% gf_labs(title = "Second Order Taylor Series Approximation",
                            subtitle = "Multinomial n = 3829",
```

```
                          x=expression(theta))
```

Code for problem 3

```
n <- length(x)
ratio_val = 2 * n * log(-n / sum(log(x))) - (2 * n) - (2 * sum(log(x)))
ratio_val
```

```
## [1] 32.75024
```

```
p_val <- pchisq(ratio_val, 1, lower.tail=F)
p_val
```

```
## [1] 1.047932e-08
```

```
loglik_fun <- function(theta, x) {
  ifelse(theta <= -1,
         NA,
         length(x) * log(theta + 1) + theta * sum(log(x))
  )
}

mle <- -n / sum(log(x)) - 1

ml <- maxLik2(loglik = loglik_fun,
                       start = mle,
                       method = "NR",
                       x = x)

x_plot <- plot(ml) %>% gf_labs(title = "Second Order Taylor Series Approximation",
                           subtitle = "Observed Data, n = 30",
                           x=expression(theta))
```

```
# step 0. seed
set.seed(414)

# step 1-3. generate data
B = 10000

sim <- lapply(1:B, FUN=function(x){
  samp = runif(30, 0, 1)
  theta = -length(samp)/sum(log(samp)) - 1
  W_statistic = 2 * (loglik_fun(theta, samp) - loglik_fun(0, samp))

  data.frame(W = W_statistic)
})

sim_data <- do.call(rbind, sim)

# step 4. fraction of w* exceeding w obs. + histogram
empirical_p_val <- sum(sim_data$W >= ratio_val) / B
empirical_p_val
```

```
## [1] 0
```

```
hist <- ggplot(data=sim_data) +
  geom_histogram(mapping=aes(x=W), binwidth=0.5) +
  labs(title="Simulated Values for W",
       subtitle="10000 samples")
hist
```