w271: Homework 1 (Due: 4pm Monday Week 2)

Professor Jeffrey Yau

Due: Before the Live Session of Week 2

Instructions (Please Read it Carefully!):

- **Page limit of the pdf report: None, but be reasonable**
- Page setup:
- Do not play around with the margin, linespace, and font size;
- Use the one specified below:
  - fontsize=11pt
  - margin=1in
  - line_spacing=single
- Submission:
  - Homework needs to be completed individually; this is not a group project. Each student needs to submit his/her homework to the course github repo by the deadline; submission and revision made after the deadline will not be graded
  - Submit 2 files:
    1. A pdf file that details your answers. Include all the R codes used to produce the answers. *Please do not suppress the codes in your pdf file.*
    2. R markdown file used to produce the pdf file
  - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
    - StudentFirstNameLastName_HWNumber.fileExtension
    - For example, if the student's name is Kyle Cartman for homework 1, name your files as the follow
      - KyleCartman_HW1.Rmd
      - KyleCartman_HW1.pdf
  - Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files.
  - For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to (1) provide an explanation of why such libraries and functions are used instead and (2) reference to the library documentation. **Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.** For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc.

- For mathematical formulae, type them in your R markdown file. **Do not write them on a piece of paper, snap a photo, and either insert the image file or sumbit the image file separately. Doing so will receive a** 0 **for that whole question.**
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Question 1: True Confidence Level of Various Confidence Intervals for One Binary Random Variable

During the live session in week 1, I explained why the Wald confidence interval does not always have the stated confidence level, $1-\alpha$, where α, which is the probability of rejecting the null hypothesis when it is true, often is set to 0.05%, and I walked through the code below to explain the concept.

```
require(knitr)
```

```
## Loading required package: knitr
```

```
## Warning: package 'knitr' was built under R version 3.4.4
```

```
# Wrap long lines in R:
opts_chunk$set(tidy.opts=list(width.cutoff=80),tidy=TRUE)


pi = 0.6 # true parameter value of the probability of success
alpha = 0.05 # significane level
n = 10
w = 0:n

wald.CI.true.coverage = function(pi, alpha=0.05, n) {

  # Objective:
  #    Calculate the true confidence level of a Wald Confidence (given pi, alpha, and n)

  # Input:
  #    pi: the true parameter value
  #    alpha: significance level
  #    n: the number of trials

  # Return:
  #    wald.df: a data.frame containing
  #    (1) observed number of success, w
  #    (2) MLE of pi, pi.hat
  #    (3) Binomial probability of obtaining the number of successes from n trials, pmf
  #    (4) lower bound of the Wald confidence interval, wald.CI_lower.bound
  #    (5) upper bound of the Wald confidence interval, wald.CI_upper.bound
  #    (6) whether or not an interval contains the true parameter, covered.pi

  w = 0:n

  pi.hat = w/n
  pmf = dbinom(x=w, size=n, prob=pi)

  var.wald = pi.hat*(1-pi.hat)/n
  wald.CI_lower.bound = pi.hat - qnorm(p = 1-alpha/2)*sqrt(var.wald)
  wald.CI_upper.bound = pi.hat + qnorm(p = 1-alpha/2)*sqrt(var.wald)
```

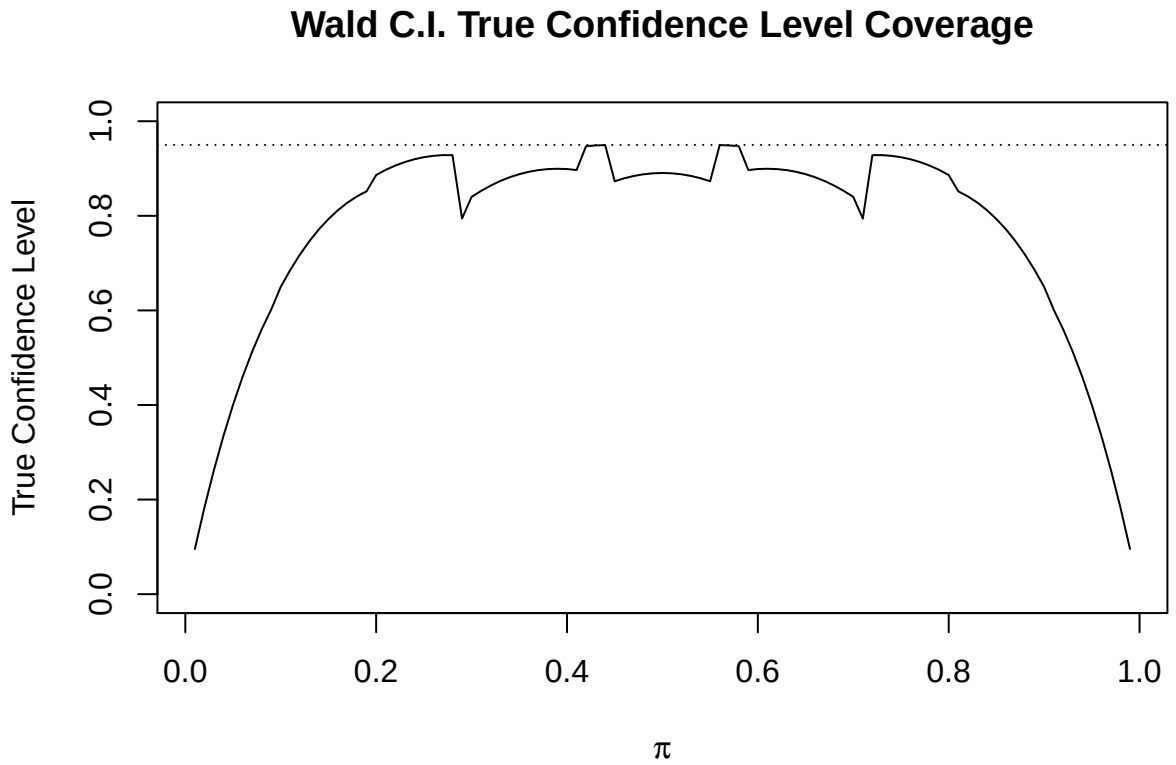
```
  covered.pi = ifelse(test = pi>wald.CI_lower.bound,
                      yes = ifelse(test = pi<wald.CI_upper.bound, yes=1, no=0), no=0)

  wald.CI.true.coverage = sum(covered.pi*pmf)

  wald.df = data.frame(w, pi.hat,
                       round(data.frame(pmf, wald.CI_lower.bound,wald.CI_upper.bound),4),
                       covered.pi)

  return(wald.df)
}

# Call the function with user-provided arguments (pi, alpha, n) to
# generate the data.frame that contains
# (1) the observed number of success, w
# (2) MLE of pi, pi.hat
# (3) Binomial probability of obtaining the number of successes from n trials, pmf
# (4) the lower bound of the Wald confidence interval, wald.CI_lower.bound
# (5) the upper bound of the Wald confidence interval, wald.CI_upper.bound
# (6) whether or not an interval contains the true parameter, covered.pi

wald.df = wald.CI.true.coverage(pi=0.6, alpha=0.05, n=10)

# Obtain the true confidence level from the Wald Confidence,
# given pi, alpha, and n
wald.CI.true.coverage.level = sum(wald.df$covered.pi*wald.df$pmf)

# Generalize the above computation to a sequence of pi's

# Generate an example sequence of pi (feel free to make the increment smaller)
pi.seq = seq(0.01, 0.99, by=0.01)

# Create a matrix to store (1) pi and (2) the true confidence level of
# the Wald Confidence Interval corresponding to the specific pi
wald.CI.true.matrix = matrix(data=NA,nrow=length(pi.seq),ncol=2)

# Loop through the sequence of pi's to obtain the true confidence level of
# the Wald Confidence Interval corresponding to the specific pi
counter=1
for (pi in pi.seq) {
    wald.df2 = wald.CI.true.coverage(pi=pi, alpha=0.05, n=10)
    #print(paste('True Coverage is', sum(wald.df2$covered.pi*wald.df2$pmf)))
    wald.CI.true.matrix[counter,] = c(pi,sum(wald.df2$covered.pi*wald.df2$pmf))
    counter = counter+1
}
str(wald.CI.true.matrix)
```

```
##  num [1:99, 1:2] 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 ...
```

```
wald.CI.true.matrix[1:5,]
```

```
##      [,1]   [,2]
## [1,] 0.01 0.0956
## [2,] 0.02 0.1828
## [3,] 0.03 0.2624
## [4,] 0.04 0.3347
## [5,] 0.05 0.4002
```

```
# Plot the true coverage level (for given n and alpha)
plot(x=wald.CI.true.matrix[,1],
     y=wald.CI.true.matrix[,2],
     ylim=c(0,1),
     main = "Wald C.I. True Confidence Level Coverage", xlab=expression(pi),
     ylab="True Confidence Level",
     type="l")
abline(h=1-alpha, lty="dotted")
```



Question 1a: Use the code above and (1) redo the following exercise for $n = 50, n = 100, n = 500$, (2) plot the graphs, and (3) describe what you have observed from the results. Use the same $pi.seq$ as I used in the code above.

Question 1b: (1) Modify the code above for the Wilson Interval. (2) Do the exercise for $n = 10, n = 50, n = 100, n = 500$. (3) Plot the graphs. (4) Describe what you have observed from the results and compare the Wald and Wilson intervals based on your results. Use the same $pi.seq$ as in the code above.

Note: The discussion of the Wilson confidence interval is in the book page 11 and 12.

Question 2: Confidence Interval Interpretation

Is it okay to say that the "estimated" confidence interval has $(1-\alpha)100\%$ probability of containing the ture parameter, named θ?

For instance, suppose we have a sample of data, and we use that sample to estimate a parameter, θ, of a statistical model and the confidence interval of the estimate. Suppose the resulting estimated 95% confidence interval is $[-2, 2]$. From a frequentist perspective, can we say that this estimated confidence interval contains the true parameter, θ, 95% of the time?

Please answer (1) Yes or No, and (2) give the reasoning of your answer provided in (1).

Question 3: Odds Ratios

When studying the multiple binary random varibles, we often use the notion of odds. The "odds" is simply the probability of a success divided by the probability of a failure: $\frac{\pi}{1-\pi}$

Suppose $\pi = 0.1$

Question 3a: What are the corresponding odds?

Question 3b: Interpret it in the following two types of statements

- **1. The odds of success are X. (Fill in X)**
- **2. The probability of failure is X times the probability of success. (Fill in X)**

The notion of odds ratio becomes relevant when there are more than one groups and we to compare their odds.

The odds ratio is the ratio of two odds. Mathematically, it is

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

where

- π_i denotes the probability of success of Group i, $i \in \{1, 2\}$
- $odds_i$ represents the odds of a success of group i, $i \in \{1, 2\}$

Question 3c: Suppose the $OR = 3$. Write down the odds of success of group 1 in relation to the odds of success of group 2.

Question 4: Binary Logistic Regression

Do **Exercise 8 a, b, c, and d (on page 131 of Bilder and Loughin's textbook)**. Please write down each of the questions. The dataset for this question is stored in the file *"placekick.BW.csv"*. The dataset is provided to you. In general, all the R codes and datasets used in Bilder and Loughin's book are provided on the book's website: chrisbilder.com

For **question 8b**, change it to the following: Re-estimate the model in part (a) using "Sun" as the base level category for $Weather$.