Final Project

Jackson Dial

11/23/2021

library(Hmisc)  
library(tidyverse)  
library(tictoc)  
library(flextable)  
library(here)

## Part 1

set.seed(23)  
  
myfun1 <- function(n, theta, method, runs) {  
 #for the number of successes  
 myvec <- c()  
  
 results <- matrix(NA, nrow = runs, ncol = 3)  
  
 for (i in 1:runs) {  
 myvec[i] <- rbinom(1, n, theta)  
 results[i, ] <- binconf(myvec[i], n, .05, method = method)  
 }  
  
 colnames(results) <- c("Estimate", "Lower", "Upper")  
 resultsdf <- as.data.frame(results)  
  
 df2 <- resultsdf %>%  
 filter(Lower < theta & Upper > theta)  
  
 return(c(  
 "Coverage Probability" = length(df2$Upper) / runs,  
 "Average Width" =  
 round(sum(resultsdf$Upper - resultsdf$Lower) / runs, digits = 4  
 )))  
}

tic()  
f1 <- myfun1(n = 20, theta = .05, method = "asymptotic", runs = 10000)  
f2 <- myfun1(n = 40, theta = .05, method = "asymptotic", runs = 10000)  
f3 <- myfun1(n = 100, theta = .05, method = "asymptotic", runs = 10000)  
  
f4 <- myfun1(n = 20, theta = .05, method = "wilson", runs = 10000)  
f5 <- myfun1(n = 40, theta = .05, method = "wilson", runs = 10000)  
f6 <- myfun1(n = 100, theta = .05, method = "wilson", runs = 10000)  
  
f7 <- myfun1(n = 20, theta = .05, method = "exact", runs = 10000)  
f8 <- myfun1(n = 40, theta = .05, method = "exact", runs = 10000)  
f9 <- myfun1(n = 100, theta = .05, method = "exact", runs = 10000)  
  
f10 <- myfun1(n = 20, theta = .15, method = "asymptotic", runs = 10000)  
f11 <- myfun1(n = 40, theta = .15, method = "asymptotic", runs = 10000)  
f12 <- myfun1(n = 100, theta = .15, method = "asymptotic", runs = 10000)  
  
f13 <- myfun1(n = 20, theta = .15, method = "wilson", runs = 10000)  
f14 <- myfun1(n = 40, theta = .15, method = "wilson", runs = 10000)  
f15 <- myfun1(n = 100, theta = .15, method = "wilson", runs = 10000)  
  
f16 <- myfun1(n = 20, theta = .15, method = "exact", runs = 10000)  
f17 <- myfun1(n = 40, theta = .15, method = "exact", runs = 10000)  
f18 <- myfun1(n = 100, theta = .15, method = "exact", runs = 10000)  
toc()

## 8.54 sec elapsed

### Make Table

#coverage probability = [1], average width = [2]  
  
col1 <- rbind("Asymptotic (p = 0.05)", "N = 20", "N = 40", "N = 100",  
 "Wilson (p = 0.05)", "N = 20", "N = 40", "N = 100",  
 "Exact (p = 0.05)", "N = 20", "N = 40", "N = 100",  
 "Asymptotic (p = 0.15)", "N = 20", "N = 40", "N = 100",  
 "Wilson (p = 0.15)", "N = 20", "N = 40", "N = 100",  
 "Exact (p = 0.15)", "N = 20", "N = 40", "N = 100")  
  
col2 <- rbind("", f1[1], f2[1], f3[1],  
 "", f4[1], f5[1], f6[1],  
 "", f7[1], f8[1], f9[1],  
 "", f10[1], f11[1], f12[1],  
 "", f13[1], f14[1], f15[1],  
 "", f16[1], f17[1], f18[1])  
  
col3 <- rbind("", f1[2], f2[2], f3[2],  
 "", f4[2], f5[2], f6[2],  
 "", f7[2], f8[2], f9[2],  
 "", f10[2], f11[2], f12[2],  
 "", f13[2], f14[2], f15[2],  
 "", f16[2], f17[2], f18[2])  
  
tab1 <- cbind(col1, col2, col3)  
  
df1 <- as.data.frame(tab1) %>%  
 rename("Method/Sample Size" = V1)  
  
flex1 <- flextable(df1)  
flex1 <- width(flex1, width = 1.7)  
flex1 <- align\_text\_col(flex1, align = "center", header = TRUE)  
flex1 <- align(flex1, j = 1, align = "left")

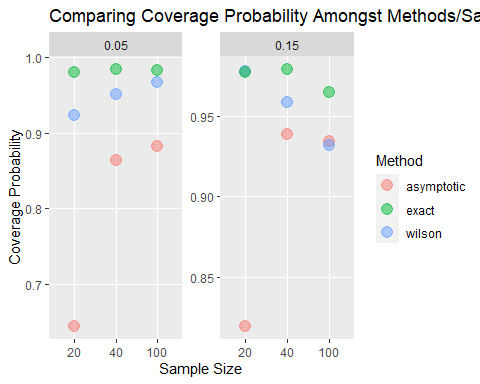
flex1

| Method/Sample Size | Coverage Probability | Average Width |
| --- | --- | --- |
| Asymptotic (p = 0.05) |  |  |
| N = 20 | 0.6447 | 0.1468 |
| N = 40 | 0.8643 | 0.12 |
| N = 100 | 0.8824 | 0.0832 |
| Wilson (p = 0.05) |  |  |
| N = 20 | 0.9237 | 0.2209 |
| N = 40 | 0.9511 | 0.1467 |
| N = 100 | 0.9666 | 0.0882 |
| Exact (p = 0.05) |  |  |
| N = 20 | 0.9807 | 0.2385 |
| N = 40 | 0.9846 | 0.1568 |
| N = 100 | 0.9833 | 0.0942 |
| Asymptotic (p = 0.15) |  |  |
| N = 20 | 0.8192 | 0.2933 |
| N = 40 | 0.9383 | 0.215 |
| N = 100 | 0.934 | 0.1385 |
| Wilson (p = 0.15) |  |  |
| N = 20 | 0.9776 | 0.2987 |
| N = 40 | 0.9582 | 0.2156 |
| N = 100 | 0.9316 | 0.1386 |
| Exact (p = 0.15) |  |  |
| N = 20 | 0.9774 | 0.334 |
| N = 40 | 0.9792 | 0.2369 |
| N = 100 | 0.9646 | 0.1476 |

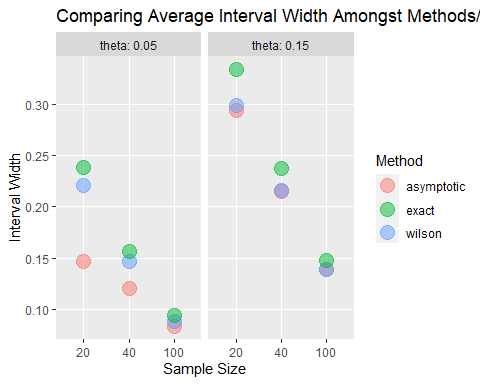
### Make Plots

type <- c(rep("asymptotic", 3), rep("wilson", 3), rep("exact", 3),  
 rep("asymptotic", 3), rep("wilson", 3), rep("exact", 3)) %>%  
 as.data.frame() %>%  
 rename(type = ".")  
  
sampsize <- rbind(20, 40, 100, 20, 40, 100, 20, 40, 100,  
 20, 40, 100, 20, 40, 100, 20, 40, 100)  
  
covprob <- rbind(  
 f1[1], f2[1], f3[1],  
 f4[1], f5[1], f6[1],  
 f7[1], f8[1], f9[1],  
 f10[1], f11[1], f12[1],  
 f13[1], f14[1], f15[1],  
 f16[1], f17[1], f18[1]  
 )  
  
avgwidth <- rbind(  
 f1[2], f2[2], f3[2],  
 f4[2], f5[2], f6[2],  
 f7[2], f8[2], f9[2],  
 f10[2], f11[2], f12[2],  
 f13[2], f14[2], f15[2],  
 f16[2], f17[2], f18[2]  
)  
  
theta <- rbind(.05, .05, .05, .05, .05, .05, .05, .05, .05,  
 .15, .15, .15, .15, .15, .15, .15, .15, .15)  
  
plotdf <- cbind(type, sampsize, theta, covprob, avgwidth) %>%  
 as.data.frame() %>%  
 rename(covprob = `Coverage Probability`,  
 avgwidth = `Average Width`)

#plot for coverage probability  
ggplot(plotdf, aes(x = as.factor(sampsize), y = covprob)) +  
 geom\_point(aes(col = as.factor(type)), size = 4, alpha = .5) +  
 facet\_wrap(~as.factor(theta), scales = "free") +  
 theme(panel.grid.minor = element\_blank()) +  
 labs(x = "Sample Size",  
 y = "Coverage Probability",  
 title = "Comparing Coverage Probability Amongst Methods/Sample Sizes",  
 col = "Method")



#Plot for interval width  
ggplot(plotdf, aes(x = as.factor(sampsize), y = avgwidth)) +  
 geom\_point(aes(col = as.factor(type)), size = 5, alpha = .5) +  
 facet\_wrap(~theta, labeller = label\_both) +  
 theme(panel.grid.minor = element\_blank()) +  
 labs(x = "Sample Size",  
 y = "Interval Width",  
 title = "Comparing Average Interval Width Amongst Methods/Sample Sizes",  
 col = "Method")



## Part 2

Based on the results of the simulation, I would choose the Wilson method. It is clear in the plot of coverage probability that exact has the highest of all 3 methods, but the Wilson method is always close, but it also always has a smaller average interval width, which is favorable.

## Part 3

dat <- read\_csv(here("Data", "mortality\_data.csv"))  
  
summary(dat)

## id death sex   
## Min. : 1.00 Min. :0.00000 Length:62   
## 1st Qu.:16.25 1st Qu.:0.00000 Class :character   
## Median :31.50 Median :0.00000 Mode :character   
## Mean :31.50 Mean :0.08065   
## 3rd Qu.:46.75 3rd Qu.:0.00000   
## Max. :62.00 Max. :1.00000

levels(as.factor(dat$sex))

## [1] "-999" "F" "M"

dat\_clean <- dat %>%  
 filter(sex %in% c("M", "F"))  
sum(is.na(dat\_clean))

## [1] 0

#one obs has been removed from the original df due to an invalid data value

### CI

x <- dat\_clean %>%  
 filter(death == 1) %>%  
 nrow()  
x

## [1] 5

n <- dat\_clean %>%  
 nrow()  
n

## [1] 61

x / n

## [1] 0.08196721

wci <- round(binconf(x, n, alpha = .05, method = "wilson"), 4)  
wci

## PointEst Lower Upper  
## 0.082 0.0355 0.1779

### Men CI

datm <- dat\_clean %>%  
 filter(sex == "M")  
  
xm <- datm %>%  
 filter(death == 1) %>%  
 nrow()  
nm <- datm %>%  
 nrow()  
xm / nm

## [1] 0.1153846

wcim <- round(binconf(xm, nm, .05, "wilson"), 4)  
wcim

## PointEst Lower Upper  
## 0.1154 0.04 0.2898

### Women CI

datf <- dat\_clean %>%  
 filter(sex == "F")  
  
xf <- datf %>%  
 filter(death == 1) %>%  
 nrow()  
nf <- datf %>%  
 nrow()  
xf / nf

## [1] 0.05714286

wcif <- round(binconf(xf, nf, .05, "wilson"), 4)  
wcif

## PointEst Lower Upper  
## 0.0571 0.0158 0.1861

### Table

col1 <- rbind("Wilson CI with Full DF",  
 wci[1],  
 "Wilson CI Male Only",  
 wcim[1],  
 "Wilson CI Female Only",  
 wcif[1])  
  
col2 <- rbind(" ",  
 wci[2],  
 " ",  
 wcim[2],  
 " ",  
 wcif[2])  
  
col3 <- rbind(" ",  
 wci[3],  
 " ",  
 wcim[3],  
 " ",  
 wcif[3])  
  
cidf <- as.data.frame(cbind(col1, col2, col3)) %>%  
 rename("Point Estimate" = V1, "Lower Bound" = V2, "Upper Bound" = V3)  
  
myflex <- flextable(cidf)  
myflex <- width(myflex, j = c(1, 2, 3), width = c(1.9, 1.3, 1.3))  
myflex <- align\_text\_col(myflex, align = "center", header = TRUE)  
myflex <- align(myflex, j = 1, align = "left")  
  
myflex

| Point Estimate | Lower Bound | Upper Bound |
| --- | --- | --- |
| Wilson CI with Full DF |  |  |
| 0.082 | 0.0355 | 0.1779 |
| Wilson CI Male Only |  |  |
| 0.1154 | 0.04 | 0.2898 |
| Wilson CI Female Only |  |  |
| 0.0571 | 0.0158 | 0.1861 |