Lab #3: Large-sample Confidence Intervals

Carley Dziewicki

Due Friday, May 22

Goals

1. Explore the coverage properties of large sample confidence intervals.
2. Explore sensitivity of coverage probabilities to (i) sample size and (ii) the shape of the original distribution.

Exercises

For each of the following, submit the required R commands along with your answers. A script has been provided for you in this folder, `lab3_ci_simulation_script.R`. The script creates a function to plot confidence intervals and determine the number that contain the parameter of interest. The only lines you will need to modify for the simulations are between 22 and 38. Do not modify other lines. (Hint: Check out the screencast to see how to use the script if you have questions.)

```
myinterval.plot<- function (ll, ul,mu){
    y1 <- ll
    y2 <- ul
    n <- length(y1)
    plot(y1, type = "n", ylim = c(min(ll,ul), max(ll,ul)),
         xlab = "Interval Number", ylab = " ")
    condition <- (ll <= mu & ul >= mu)
    segments((1:n)[y1 < mu & y2 > mu], y1[y1 < mu & y2 > mu], (1:n)[y1 <
        mu & y2 > mu], y2[y1 < mu & y2 > mu])
    segments((1:n)[y1 > mu], y1[y1 > mu], (1:n)[y1 > mu], y2[y1 >
        mu], col = 17, lwd = 8)
    segments((1:n)[y2 < mu], y1[y2 < mu], (1:n)[y2 < mu], y2[y2 <
        mu], col = 17, lwd = 8)
    SUM <- sum(condition)
    abline(h = mu)
    paste("Number of intervals that contain",mu,"=", SUM)
}
```

1. Use the script provided for this lab to carry out a simulation experiment to investigate the coverage properties (percent ofintervals that contain the true population mean μ) of the confidence interval on page 392 of the text.

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \quad (1)$$

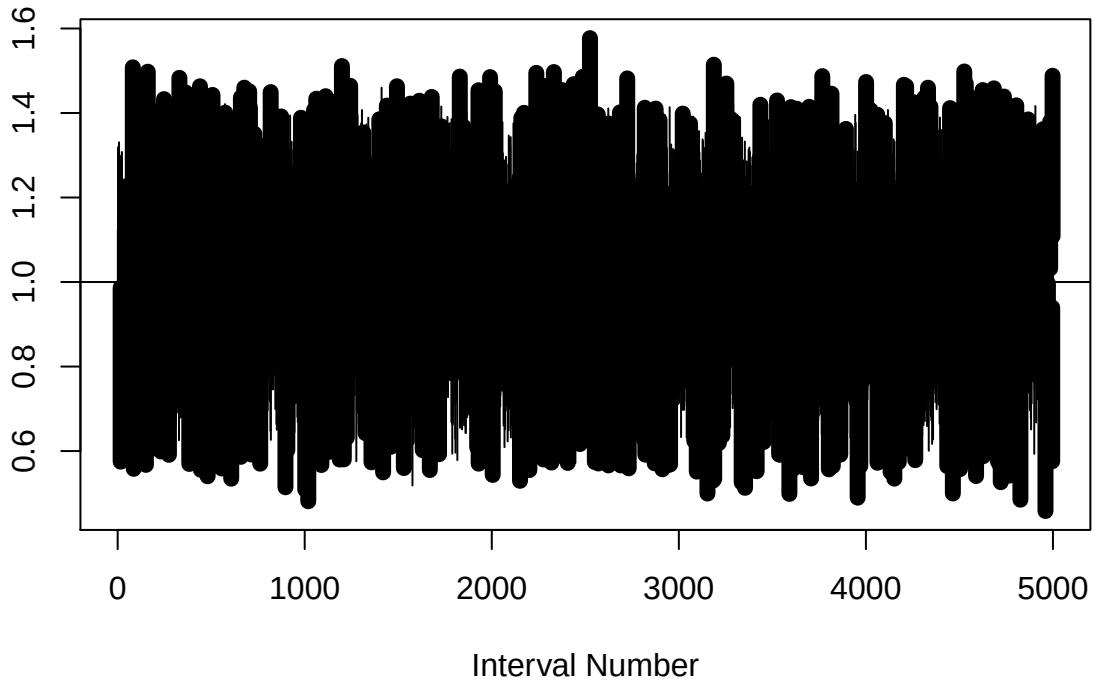
Set the number of samples M to be 5000, the mean to be 1 and the variance to be 1 and confidence level to be 95%. Carry out simulations for four different sample sizes: $n = 3$, $n = 10$, $n = 30$ and

$n = 100$. Make a summary table of your results including the proportion of the random intervals out of 5000 that contained the true mean μ in each case. [Hint: You can use this site to create a table of results: https://www.tablesgenerator.com/markdown_tables

Answer:

```
M <- 5000 # Number of samples
n <- 100 # Sample size
mu <- 1
sigma <- 1
alpha <- .05
clevel <- 1-alpha
#Create the confidence intervals
xbar <- numeric(M)
ll <- numeric(M)
ul <- numeric(M)
for (i in 1:M) {
# Next line specifies that we're sampling from a normal
# distribution with mean = mu and sd = sigma
xvec <- rnorm(n,mean = mu,sd = sigma)
# xbar values saved as a vector to allow plotting of sampling distribution
xbar[i] <- mean(xvec)
s <- sd(xvec)
ll[i] <- xbar[i] + qnorm(alpha/2)*s/sqrt(n)
ul[i] <- xbar[i] + qnorm(1-alpha/2)*s/sqrt(n)
}

myinterval.plot(ll,ul,mu)
```




```
## [1] "Number of intervals that contain 1 = 4731"
```

*I used the same code for each n value, but did not include each in the document to save room. Above is for the case n=100.

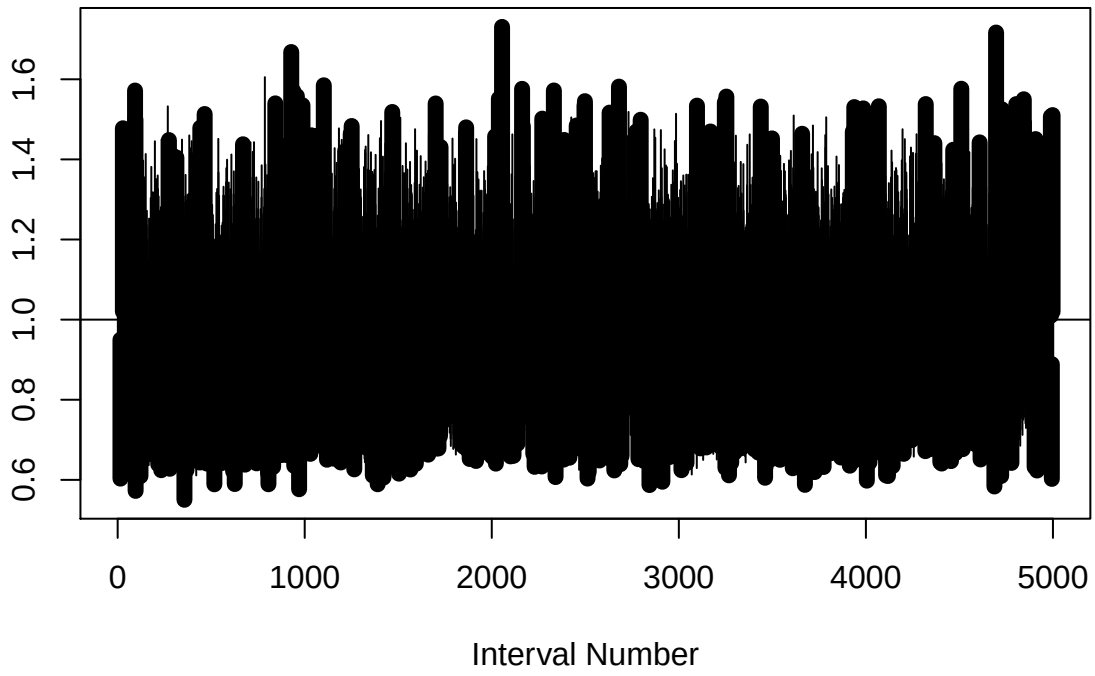
n	prop
3	.8128
10	.9164
30	.9402
100	.9462

2. Repeat the previous simulation experiment but this time modify the script so that the sample is drawn from an Exponential distribution with rate parameter equal to 1.

Answer:

```
M <- 5000 # Number of samples
n <- 100 # Sample size
rate <- 1
alpha <- .05
clevel <- 1-alpha
#Create the confidence intervals
xbar <- numeric(M)
ll <- numeric(M)
ul <- numeric(M)
for (i in 1:M) {
# Next line specifies that we're sampling from a normal
# distribution with mean = mu and sd = sigma
xvec <- rexp(n,rate = rate)
# xbar values saved as a vector to allow plotting of sampling distribution
xbar[i] <- mean(xvec)
s <- sd(xvec)
ll[i] <- xbar[i] + qnorm(alpha/2)*s/sqrt(n)
ul[i] <- xbar[i] + qnorm(1-alpha/2)*s/sqrt(n)
}

myinterval.plot(ll,ul,mu)
```



```
## [1] "Number of intervals that contain 1 = 4686"
```

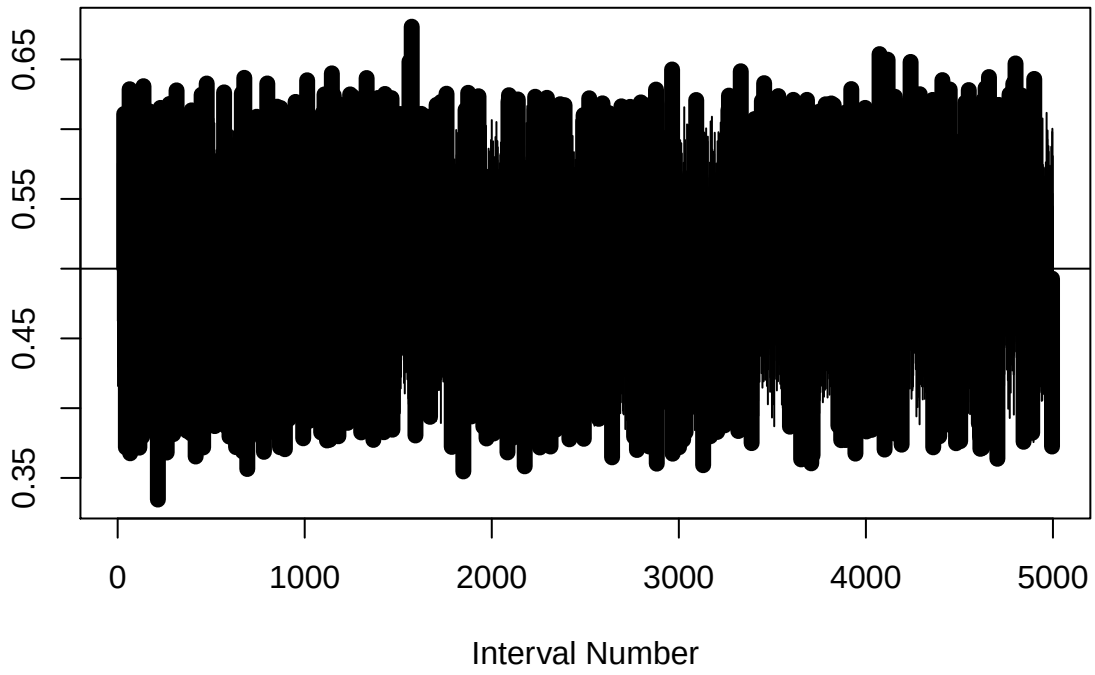
n	prop
3	.7518
10	.8646
30	.9208
100	.9392

3. Repeat the previous simulation experiment again but this time modify the script so that the sample is drawn from a Uniform $[0, 1]$ population. Note: Be sure to notice that the mean and standard deviation are no longer 1. Use the sample mean and standard deviation (from the Uniform distribution with $A = 0$ and $B = 1$) to compute the intervals as in equation (1) above. To calculate the proportion that contain the true population mean, be sure to use the mean (expected value) of the Uniform $[0, 1]$ distribution (not the sample mean).

Answer:

```
M <- 5000 # Number of samples
n <- 100 # Sample size
A <- 0
B <- 1
alpha <- .05
clevel <- 1-alpha
#Create the confidence intervals
xbar <- numeric(M)
ll <- numeric(M)
ul <- numeric(M)
for (i in 1:M) {
# Next line specifies that we're sampling from a normal
# distribution with mean = mu and sd = sigma
xvec <- runif(n,min=A, max=B)
# xbar values saved as a vector to allow plotting of sampling distribution
xbar[i] <- mean(xvec)
s <- sd(xvec)
ll[i] <- xbar[i] + qnorm(alpha/2)*s/sqrt(n)
ul[i] <- xbar[i] + qnorm(1-alpha/2)*s/sqrt(n)
}

myinterval.plot(ll,ul,.5)
```




```
## [1] "Number of intervals that contain 0.5 = 4709"
```

n	prop
3	.7822
10	.9176
30	.9430
100	.9446

4. Summarize your findings from these simulation experiments. Describe how the coverage properties of this confidence interval change as n increases and comment on similarities and differences across the distributions.

Answer:
From the simulations it can be seen that as the sample size (n) increases the proportion of "successful" intervals (or those that capture the true mean) also increases. The proportion of successful intervals is not only increasing but it seems to be converging on the confidence level that we specified at .95. This is similar across all the distributions, that as n increases the proportion approached .95. As for differences across the distributions we can see that exponential distribution required a larger sample size to get more accurate, meaning with n=3 and n=10 the prop was still pretty far from .95, whereas the uniform and normal distributions approached .95 "faster" or were more accurate even at n=10. Overall the normal distibution has the highest proportions comparitivly across the distirbutions and sample sizes.