DSCC/CSC/TCS 462 Assignment 2

Due Thursday, October 6, 2022 by 4:00 p.m.

Daxiang Na

This assignment will cover material from Lectures 6, 7, and 8.

1. Consider random variables X and Y. Calculate $\mathrm{Var}(3X + 2Y)$ given the following information. (Hint: At some point, you may need to use the fact that variance cannot be negative.)

   - $\mathrm{E}(3X + 2) = 8$
   - $\mathrm{E}(4X + 2Y) = 14$
   - $\mathrm{E}(2Y(X + 1)) = 28$
   - $\mathrm{E}(X^2Y^2) = 144$
   - $\mathrm{Cov}(X^2, Y^2) = 36$
   - $\mathrm{E}(X^2 + 2Y^2) = 33$

   Answer:

   1. $E(3X + 2) = 8$, then $E(X) = 2$
   2. $E(4X + 2Y) = 14$, then $E(Y) = 3$
   3. $E(2Y(X + 1)) = 28 = 2E(XY) + 2E(Y)$, then $E(XY) = 11$
   4. $E(X^2Y^2) = 144$
   5. $\mathrm{Cov}(X^2, Y^2) = 36 = E(X^2Y^2) - E(X^2)E(Y^2)$, then $E(X^2)E(Y^2) = 108$
   6. $E(X^2 + 2Y^2) = E(X^2) + 2E(Y^2) = 33$

   With all of those, we get $E(X) = 2, E(Y) = 3, E(Y^2) = 9/2$ or 12.

   We also get $Cov(X, Y) = E(XY) - E(X)E(Y) = 11 - 2 \times 3 = 5$

   Since variance has to be non-negative, $E(Y^2) = 12$, then $E(X^2) = 9$

   $\mathrm{Var}(X) = E(X^2) - E(X)^2 = 5$, $\mathrm{Var}(Y) = E(Y^2) - E(Y)^2 = 3$

   $\mathrm{Var}(3X + 2Y) = 9\,\mathrm{Var}(X) + 4\,\mathrm{Var}(Y) + 12\,\mathrm{Cov}(X, Y) = 9 \times 5 + 4 \times 3 + 12 \times 5 = 117$

2. The density function of X is given by $f_X(x) = ax^3 + bx + \frac{2}{3}$ for $x \in [0, 1]$, and $\mathrm{E}(X) = \frac{7}{15}$.

   a. Find a and b.

Answer:

Step1: $E(X) = \int (ax^3 + bx + \frac{2}{3})xdx = \int (ax^4 + bx^2 + \frac{2}{3}x)dx = \frac{a}{5}x^5 + \frac{b}{3}x^3 + \frac{1}{3}x^2$ $E(X) = \int_0^1 (ax^3 + bx + \frac{2}{3})xdx = \frac{a}{5} \cdot 1^5 + \frac{b}{3} \cdot 1^3 + \frac{1}{3} \cdot 1^2 - 0 = 7/15 \rightarrow 3a + 5b - 2 = 0$

Step2: sum of probability has to be 1, then $\int_0^1 (ax^3+bx+\frac{2}{3})dx = \frac{a}{4} \cdot 1^4 + \frac{b}{2} \cdot 1^2 + \frac{2}{3} \cdot 1 - 0 = 1$ $\rightarrow 3a + 6b - 4 = 0$

Step3: Combine 1 and 2, we get $a = -8/3, b = 2$

b. Calculate the CDF, $F(X)$.

Answer: $F(X) = \int (ax^3 + bx + \frac{2}{3})dx = -\frac{2}{3}x^4 + x^2 + \frac{2}{3}x$

c. Calculate $Pr(X > 0.75)$

```
F <- function(x) {
    return((-2/3) * x^4 + x^2 + (2/3) * x)
}
1 - F(0.75)
```

```
## [1] 0.1484375
```

Answer: $Pr(X > 0.75) = 1 - F(X = 0.75) = 0.1484375$

e. Suppose $Y = 1.5X + 2$. Calculate E(Y).

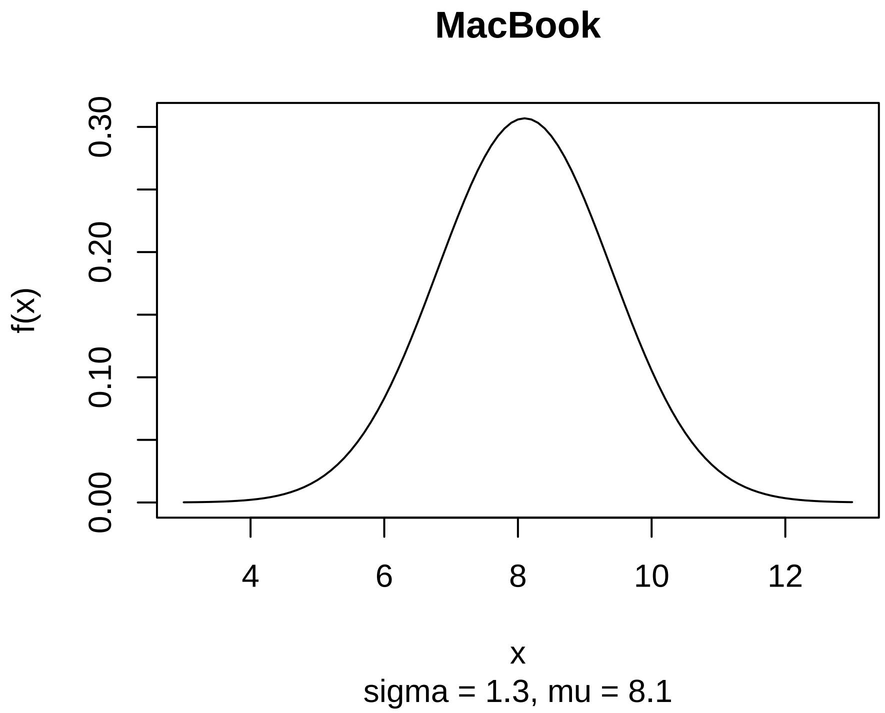
Answer: $E(Y) = 1.5E(X) + 2 = 1.5 \cdot 7/15 + 2 = 2.7$

3. The distribution of battery life of MacBook laptops is normally distributed with a mean of 8.1 hours and a standard deviation of 1.3 hours. The distribution of Dell laptops is normally distributed with a mean of 6.8 hours with a standard deviation of 0.9 hours.

   a. Calculate the probability that a randomly selected MacBook laptop battery lasts more than 9 hours.

$\mu_{MacBook} = 8.1, \sigma_{MacBook} = 1.3, \mu_{Dell} = 6.8, \sigma_{Dell} = 0.9$

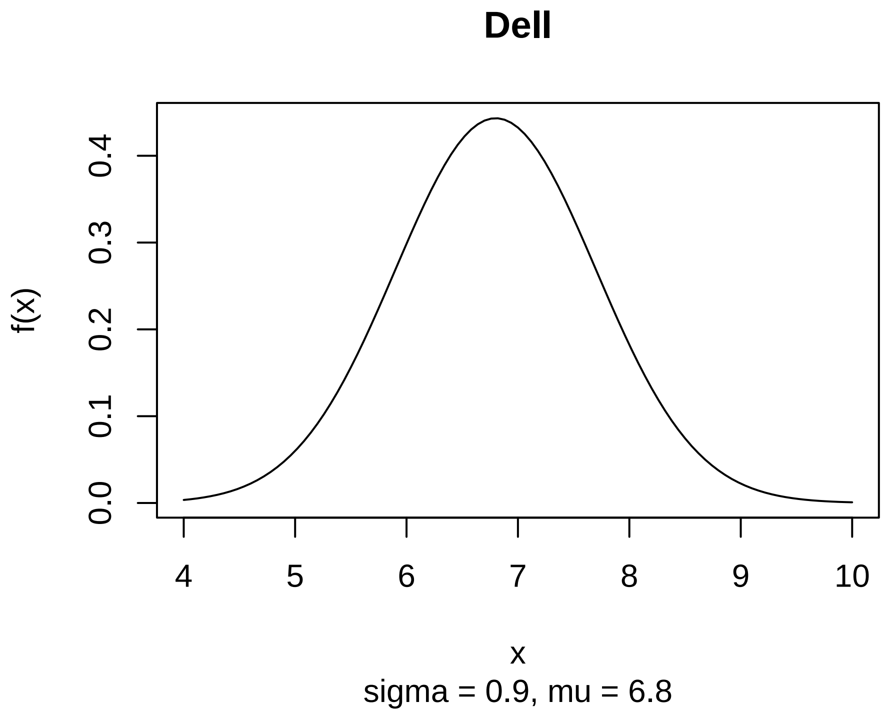
```
sigma = 1.3
mu = 8.1
f <- function(x) {
    return(1/(sqrt(2 * pi) * sigma) * exp(-0.5 * ((x - mu)/sigma)^2))
}
curve(f, from = 3, to = 13)
title(main = "MacBook", sub = "sigma = 1.3, mu = 8.1")
```



Answer: $P_{MacBook}(X > 9) = 1 - pnorm((9 - 8.1)/1.3) = 0.2443721$

b. Calculate the probability that a randomly selected Dell laptop battery lasts between 6 and 8 hours.

```
sigma = 0.9
mu = 6.8
f <- function(x) {
    return(1/(sqrt(2 * pi) * sigma) * exp(-0.5 * ((x - mu)/sigma)^2))
}
curve(f, from = 4, to = 10)
title(main = "Dell", sub = "sigma = 0.9, mu = 6.8")
```



Answer: $P_{Dell}(6 < X < 8) = pnorm((8-6.8)/0.9) - pnorm((6-6.8)/0.9) = 0.7217574$

c. How long must a MacBook laptop battery last to be in the top 3%?

Answer: $x = qnorm(1 - 0.03) * 1.3 + 8.1 = 10.54503$

d. How long must a Dell laptop battery last to be at the 30th percentile?

Answer: $x = qnorm(0.3) * 0.9 + 6.8 = 6.32804$

e. Calculate the probability that a randomly selected MacBook laptop lasts longer than the 25th percentile of Dell laptops.

Answer: the 25th percentile of Dell laptops = qnorm(0.25)*0.9+6.8 = 6.192959

$P_{MacBook}(X > 6.192959) = 1 - pnorm((6.192959 - 8.1)/1.3) = 0.9288058$

f. A randomly selected laptop has a battery life of at least 8.5 hours. Calculate the probability of this laptop being a MacBook and the probability of it being a Dell.

Answer:

$P_{MacBook}(X > 8.5) = 1 - pnorm((8.5 - 8.1)/1.3) = 0.3791582$

$P_{Dell}(X > 8.5) = 1 - pnorm((8.5 - 6.8)/0.9) = 0.02945336$

4. Payton applies for 12 jobs, each of which he has a 70% chance of getting a job offer for. Assume that job offers are independent of each other.

a. How many job offers is Payton expected to receive?

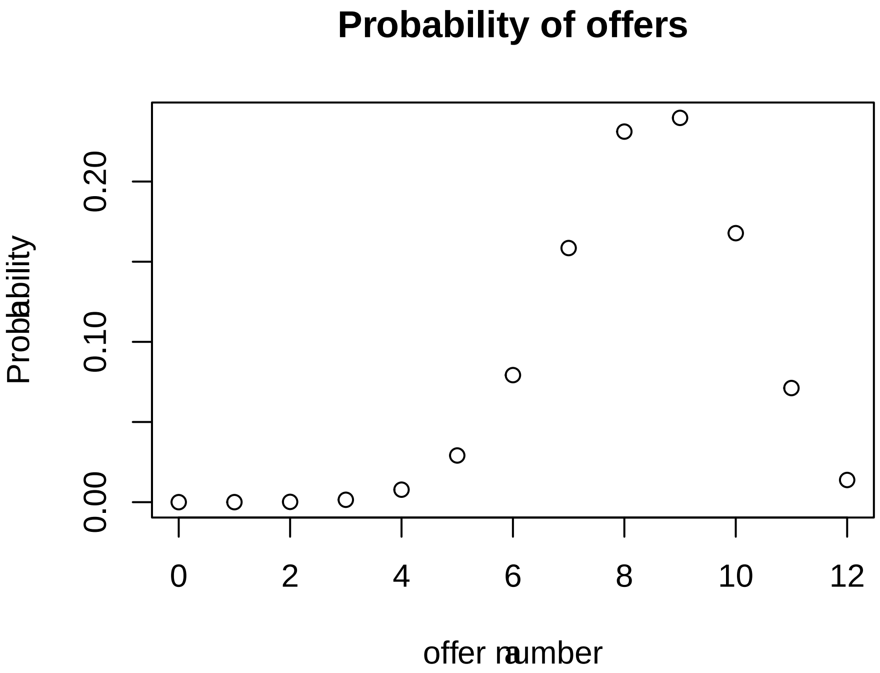
Answer: $E(X) = np = 12 \times 0.7 = 8.4$

b. Calculate the probability that Payton receives job offers from all 12 places.

Answer: $P(X = 12) = choose(12, 12) * 0.7^{12} * 0.3^0 = 0.01384129$

c. Calculate the probability that Payton receives between 5 and 7 (inclusive, i.e., 5, 6, or 7) job offers.

```r
f <- function(x) {
    return(choose(12, x) * (0.7^x) * (0.3^(12 - x)))
}
a <- c(0:12)
b <- f(a)
plot(a, b)
title(main = "Probability of offers", xlab = "offer number", ylab = "Probability")
```



```r
F <- function(x) {
    s <- 0
    for (i in x) {
        p <- choose(12, i) * (0.7^i) * (0.3^(12 - i))
        s <- p + s
    }
    return(s)
}
F(c(5:7))
```

```
## [1] 0.2668552
```

Answer: $P(x = 5, 6 or 7) = 0.2668552$

d. Calculate the probability that Payton receives strictly more than 9 job offers.

```
F(c(10:12))
```

```
## [1] 0.2528153
```

Answer: $P(x > 9) = 0.2528153$

e. Calculate the probability that Payton receives strictly fewer than 3 job offers.

```
F(c(0:2))
```

```
## [1] 0.0002063763
```

Answer: $P(x < 3) = 0.0002063763$

f. Calculate the variance of the number of job offer Payton is expected to receive.

Answer: $Var(X) = np(1 - p) = 2.52$

5. Suppose a company has three email accounts, where the number of emails received at each account follows a Poisson distribution. Account A is expected to receive 4.2 emails per hour, account B is expected to receive 5.9 emails per hour, and account C is expected to received 2.4 emails per hour. Assume the three accounts are independent of each other.

a. Calculate the variance of emails received for each of the three accounts.

Answer:

$\mu_A = \sigma_A^2 = \lambda_A = 4.2$

$\mu_B = \sigma_B^2 = \lambda_B = 5.9$

$\mu_C = \sigma_C^2 = \lambda_C = 2.4$

b. Calculate the probability that account A receives at least 8 emails in an hour.

Answer: $P_A(X \geq 8) = 1 - ppois(8-1, 4.2) = 0.06394334$

c. Calculate the probability that account B receives exactly 4 emails in an hour.

Answer: $P_B(X = 4) = dpois(4, 5.9) = 0.1383118$

d. Calculate the probability that account C receives at most 3 emails in an hour.

Answer: $P_C(X \leq 3) = ppois(3, 2.4) = 0.7787229$

e. Calculate the probability that account B receives between 2 and 4 emails in an hour.

Answer: $P_B(2 \leq X \leq 4) = ppois(4, 5.9) - ppois(1, 5.9) = 0.2797626$

f. Calculate the probability that the company receives more than 10 emails total in an hour. (Hint: the sum of Poisson random variables is also Poisson distributed. Determine λ by doing $\mathrm{E}(A + B + C)$.)

Answer:

$\lambda = \mathrm{E}(A + B + C) = 4.2 + 5.9 + 2.4 = 12.5$

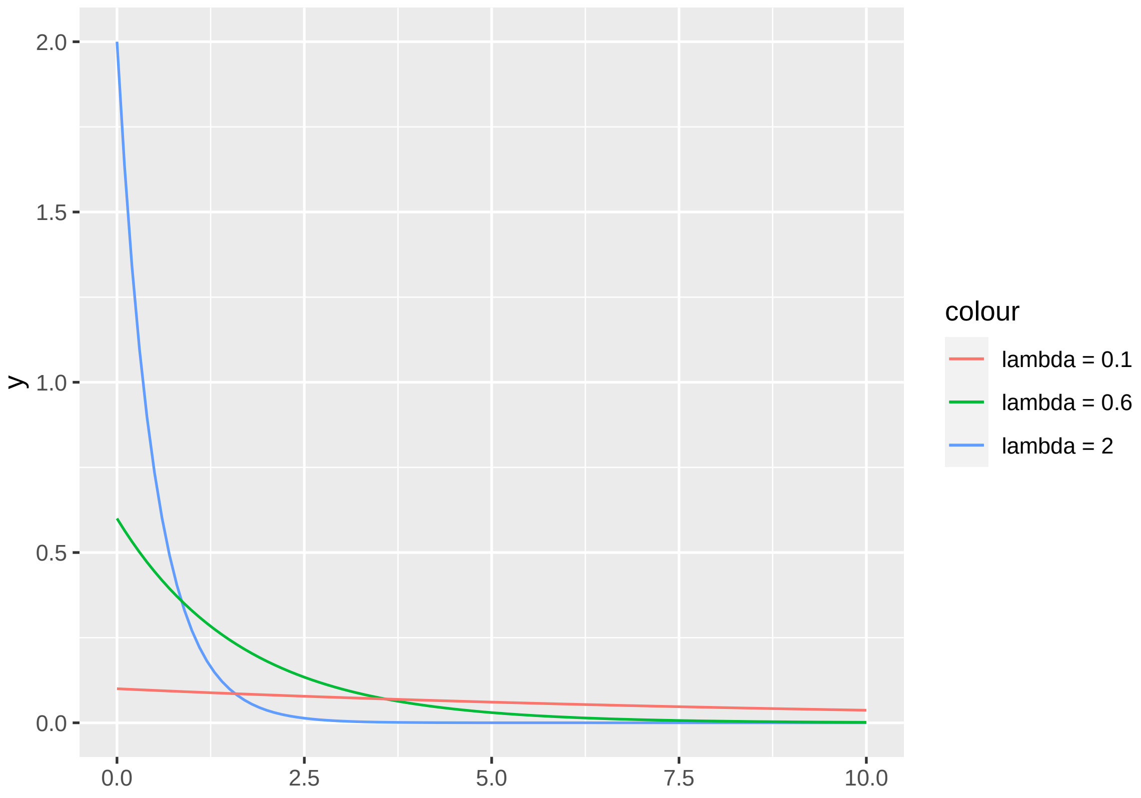
$P(X > 10) = 1 - ppois(10, 12.5) = 0.7029253$

6. Suppose that we are interested in the length of time before the next lightning strike. There are three types of lightning we are interested in: cloud-to-ground (G), cloud-to-air (A), and cloud-to-cloud (C). For all types of lightning, the length of time before the next strike is distributed according to an exponential distribution, but the exponential distribution has a different parameter for each type of lightning. In particular, $\lambda_G = 2$, $\lambda_A = 0.6$, and $\lambda_C = 0.1$.

   a. On a single plot, visualize the PDFs over the range $x \in [0, 10]$ for each of these exponential distributions. It may be helpful to use the function "dexp" in R.

```
library(ggplot2)
base <- ggplot() + xlim(0, 10)

base + geom_function(aes(colour = "lambda = 2"), fun = dexp, args = list(rate = 2))
    geom_function(aes(colour = "lambda = 0.6"), fun = dexp, args = list(rate = 0.6)
    geom_function(aes(colour = "lambda = 0.1"), fun = dexp, args = list(rate = 0.1)
```



b. What are $\mathrm{E}(G)$, $\mathrm{E}(A)$, and $\mathrm{E}(C)$, as well as $\mathrm{Var}(G)$, $\mathrm{Var}(A)$, and $\mathrm{Var}(C)$?

$\mathrm{E}(G) = 1/2 = 0.5$, $\mathrm{Var}(G) = 1/4 = 0.25$

$\mathrm{E}(A) = 1/0.6 = 1.666667$, $\mathrm{Var}(A) = (1/0.6)^2 = 2.777778$

$\mathrm{E}(G) = 1/0.1 = 10$, $\mathrm{Var}(G) = (1/0.1)^2 = 100$

c. Suppose that we repeatedly sample collections of $n = 100$ observations from the distribution of cloud-to-ground (G) lightning strike timings. What is the mean and variance of this sample distribution?

Answer: if the repeating time is large enough, then according to CLT,

$X \sim N(\mathrm{E}(G), \sqrt{\mathrm{Var}(G)}/\sqrt{n})$

The mean of this sample distribution: $\mu_X = \mathrm{E}(G) = 0.5$

The variance of this sample distribution: $\sigma_X^2 = \mathrm{Var}(G)/n = 0.25/100 = 0.0025$

d. Now, let us examine the empirical sampling distribution of cloud-to-ground (G) lightning. For each value of $n = \{10, 100, 1000\}$, sample n points from the exponential distribution from G a grand total of $m = 5000$ times, and record the mean of each sample. You should end up with three different sets of 5000 sample means. For each of these sets, report the sample mean and sample standard deviation. Comment on how the observed values line up with what you would expect in theory. It may be useful to use the R function “rexp()“.

```r
set.seed(42)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
f <- function(x) {
    N <- 5000
    df <- as.data.frame(replicate(N, rexp(x, 2))) %>%
        summarise_if(is.numeric, mean)
    a <- as.numeric(as.vector(df[1, ]))
    return(a)
}
n_10 <- f(10)
n_100 <- f(100)
n_1000 <- f(1000)
```

```r
mu_10 = mean(n_10)
mu_100 = mean(n_100)
mu_1000 = mean(n_1000)

sigma_10 = sd(n_10)
sigma_100 = sd(n_100)
sigma_1000 = sd(n_1000)

mu_10
```

```
## [1] 0.4990976
```

```r
mu_100
```

```
## [1] 0.4998623
```

```
mu_1000
```

```
## [1] 0.5001065
```

```
sigma_10
```

```
## [1] 0.1582641
```

```
sigma_100
```

```
## [1] 0.0505395
```

```
sigma_1000
```

```
## [1] 0.01577291
```

Answer:

For $n = 10$ sample size, $\mu = 0.4990976$, $\sigma = 0.1582641$

For $n = 100$ sample size, $\mu = 0.4998623$, $\sigma = 0.0505395$

For $n = 1000$ sample size, $\mu = 0.5001065$, $\sigma = 0.01577291$

$\sqrt{\text{Var}(G)}/\sqrt{10} = 0.1581139$

$\sqrt{\text{Var}(G)}/\sqrt{100} = 0.05$

$\sqrt{\text{Var}(G)}/\sqrt{1000} = 0.01581139$

Comments on results: According to CLT, the sampling mean follows normal distribution: $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, as long as the n is large enough. From those results, we can obviously see that the means of sampling are very close to the population mean, and the standard deviations are amazingly close to $\frac{\sigma}{\sqrt{n}}$. Those results show that CLT works very well.

7. Suppose that we would like to get an idea of how much coffee is consumed by the entire University of Rochester each day. We take a sample of 100 days and find that the average amount of coffee consumed by the University of Rochester per day is 580 gallons.

   a. Assume that coffee consumption comes from a normal distribution with $\sigma = 90$. Find a two-sided 95% confidence interval for the average amount of coffee consumed by the University of Rochester each day.

```
x <- 580
n <- 100
sigma <- 90
upper <- x + qnorm(0.975) * sigma/sqrt(n)
lower <- x + qnorm(0.025) * sigma/sqrt(n)
upper
```

```
## [1] 597.6397
```

```
lower
```

```
## [1] 562.3603
```

Answer: The two-sided 95% confidence interval is (562.3603, 597.6397).

b. Assuming the same information as part a, suppose that we now only want a upper-bound confidence interval. Calculate a one-sided 95% upper-bound confidence interval for the average amount of coffee consumed by the University of Rochester each day.

```
upper_2 <- x + qnorm(0.95) * sigma/sqrt(n)
upper_2
```

```
## [1] 594.8037
```

Answer: The one-sided 95% upper-bound confidence interval is 594.8037 gallons.

c. Now, suppose that we do not know the variance of the true distribution of coffee consumption. However, in our sample, we see that $s = 80$. Find a two-sided 95% confidence interval for the average amount of coffee consumed by the University of Rochester each day.

```
s <- 80
df <- 100 - 1
upper_s <- x + qt(0.975, df) * s/sqrt(n)
lower_s <- x + qt(0.025, df) * s/sqrt(n)
upper_s
```

```
## [1] 595.8737
```

```
lower_s
```

```
## [1] 564.1263
```

Answer: The two-sided 95% confidence interval for the average amount of coffee consumed by the University of Rochester each day is (564.1263, 595.8737)

d. Assuming the same information as part c, suppose that we now only want a upper-bound confidence interval. Calculate a one-sided 95% upper-bound confidence interval for the average amount of coffee consumed by the University of Rochester each day.

```
x + qt(0.95, df) * s/sqrt(n)
```

```
## [1] 593.2831
```

Answer: The one-sided 95% upper-bound confidence interval for the average amount of coffee consumed by the University of Rochester each day is 593.2831 gallons.

e. Assuming the same information as part a (i.e., known population variance), calculate the number of samples needed in order to get a two-sided 95% confidence interval for the average amount of coffee consumed by the University of Rochester each day of length 16.

Answer:

Margin of error $m = 16/2 = 8$

Given that $n = \left\lceil \frac{z^2_{\alpha/2} \cdot \sigma^2}{m^2} \right\rceil$

$n = ceiling((qnorm(0.975)^2) * 90^2/8^2) = 487$

So to get the two-sided 95% confidence interval of length 16, we need 487 samples.

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?

I didn't count the hours I spent but it took me two days in total. I think the questions were designed very well, helped me to practice the knowledge in class and get deeper understanding.

- Who, if anyone, did you work with on this assignment?

Myself, but I did discuss it with Dr. Kahng.

- What questions do you have relating to any of the material we have covered so far in class?

None for now.