Trent Bellinger

Stats 10 Lab 4

Exercise 1

1a

n <- 365
p <- 0.40

1b

binom_mean <- n*p
binom_mean

[1] 146

The mean of heavy rain days in 2019 is 146.

binom_var <- n*p*(1-p)
binom_var

[1] 87.6

binom_std <- sqrt(n*p*(1-p)) # or binom_std <9 sqrt(binom_var)
binom_std

[1] 9.359487

The standard deviation of heavy rain days in 2019 is about 9.36.

1c

dbinom(145, size=n, prob=p)

[1] 0.04239996

The probability that the park will experience exactly 145 days of heavy rain is about 0.042.

1d

pbinom(175, size=n, prob=p) - pbinom(124, size=n, prob=p)

[1] 0.9888137

The probability that the park will see between 125 and 175 days of heavy rain is about 0.989.

1e

mu <- 200
sigma <- 20

1 - pnorm(230, mean=mu, sd=sigma)

[1] 0.0668072

or

pnorm(230, mean=mu, sd=sigma, lower.tail=FALSE)

[1] 0.0668072

The probability that the park will experience more that 230 days of heavy rain is about 0.067.

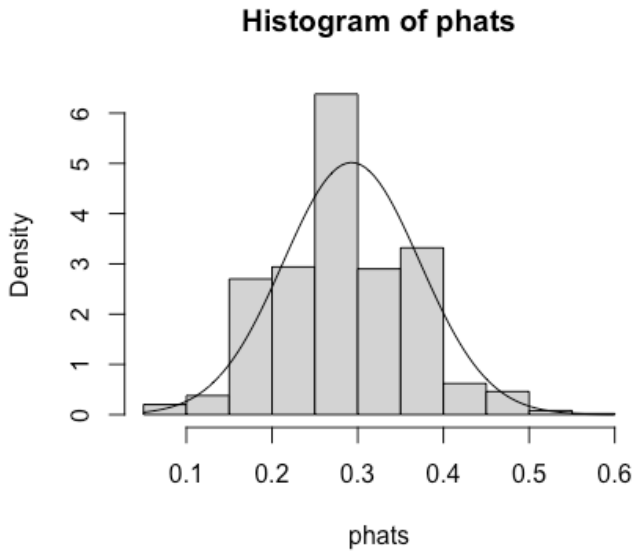
Exercise 2

pawnee <- read.csv('pawnee.csv')

2a

We first create objects for common quantities we will use for this exercise.
n <- 30 # The sample size
N <- 541 # The population size
M <- 1000 # Number of samples/repetitions
Create vectors to store the simulated proportions from each repetition.
phats <- numeric(M) # for sample proportions
Set the seed for reproduceability
set.seed(123)
Always set the seed OUTSIDE the for loop.
Now we start the loop. Let i cycle over the numbers 1 to 1000 (i.e., iterate 1000 times).
for(i in seq_len(M)){
  # The i-th iteration of the for loop represents a single repetition.
  # Take a simple random sample of size n from the population of size N.
  index <- sample(N, size = n)
  # Save the random sample in the sample_i vector.
  sample_i <- pawnee[index,]
  # Compute the proportion of the i-th sample of households with a new health issue.
  phats[i] <- mean(sample_i$New_hlth_issue == "Y")
}

hist(phats, prob=TRUE)
curve(dnorm(x, mean(phats), sd(phats)), add = TRUE)



2b

mean(phats)

[1] 0.2928

sd(phats)

[1] 0.07951963

The mean of the simulated sample proportion was about 29% (or 0.29) with a standard deviation of about 7.95% (or 0.0795).

2c

Yes I believe the simulated distribution of sample proportions is approximately normal. There is a sufficient agreement with the mode and moderate agreement elsewhere. More importantly our simulated distribution is unimodal and symmetric.

2d

p = mean(phats)
p_sd = sqrt(p * (1 - p) / n)

p

[1] 0.2928

p_sd

[1] 0.08307991

We would predict the mean to be about 29% (or 0.29) and the standard deviation to be about 8.3% (or 0.083).
These values are very close to our answers in part b. In particular our theory standard deviation is within 0.004 of our empirical standard deviation and our theory mean is equal to our empirical mean.

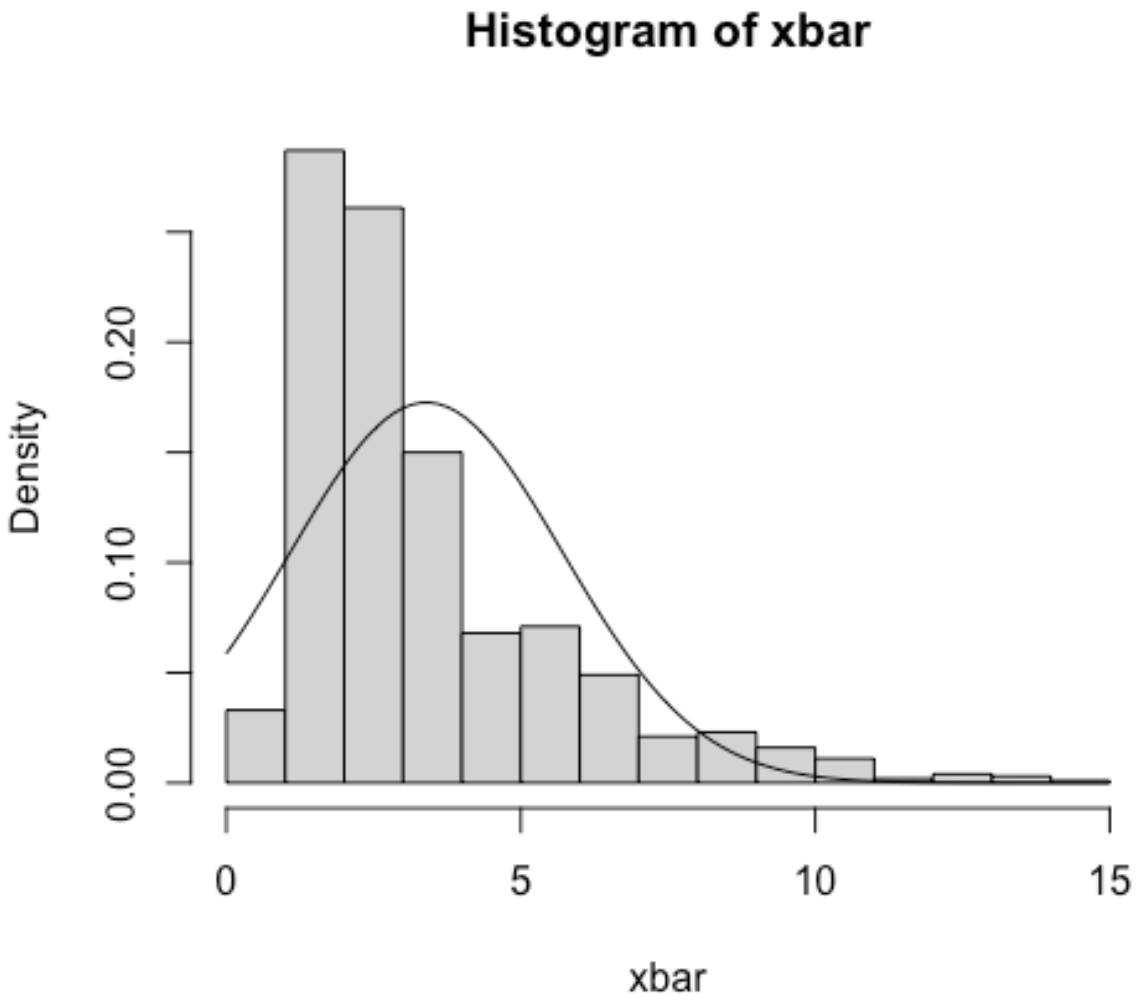
Exercise 3

3a

```
# We first create objects for common quantities we will use for this exercise.
n <- 30 # The sample size
N <- 541 # The population size
M <- 1000 # Number of samples/repetitions
# Create vectors to store the simulated proportions from each repetition.
xbar <- numeric(M) # for sample means
# Set the seed for reproduceability
set.seed(123)
# Always set the seed OUTSIDE the for loop.
# Now we start the loop. Let i cycle over the numbers 1 to 1000 (i.e., iterate 1000 times).
for(i in seq_len(M)){
  # The i-th iteration of the for loop represents a single repetition.
  # Take a simple random sample of size n from the population of size N.
  index <- sample(N, size = n)
  # Save the random sample in the sample_i vector.
  sample_i <- pawnee[index, ]
  # Compute the proportion of the i-th sample of households with a new health issue.
  xbar[i] <- mean(sample_i$Arsenic)
}
```

3b

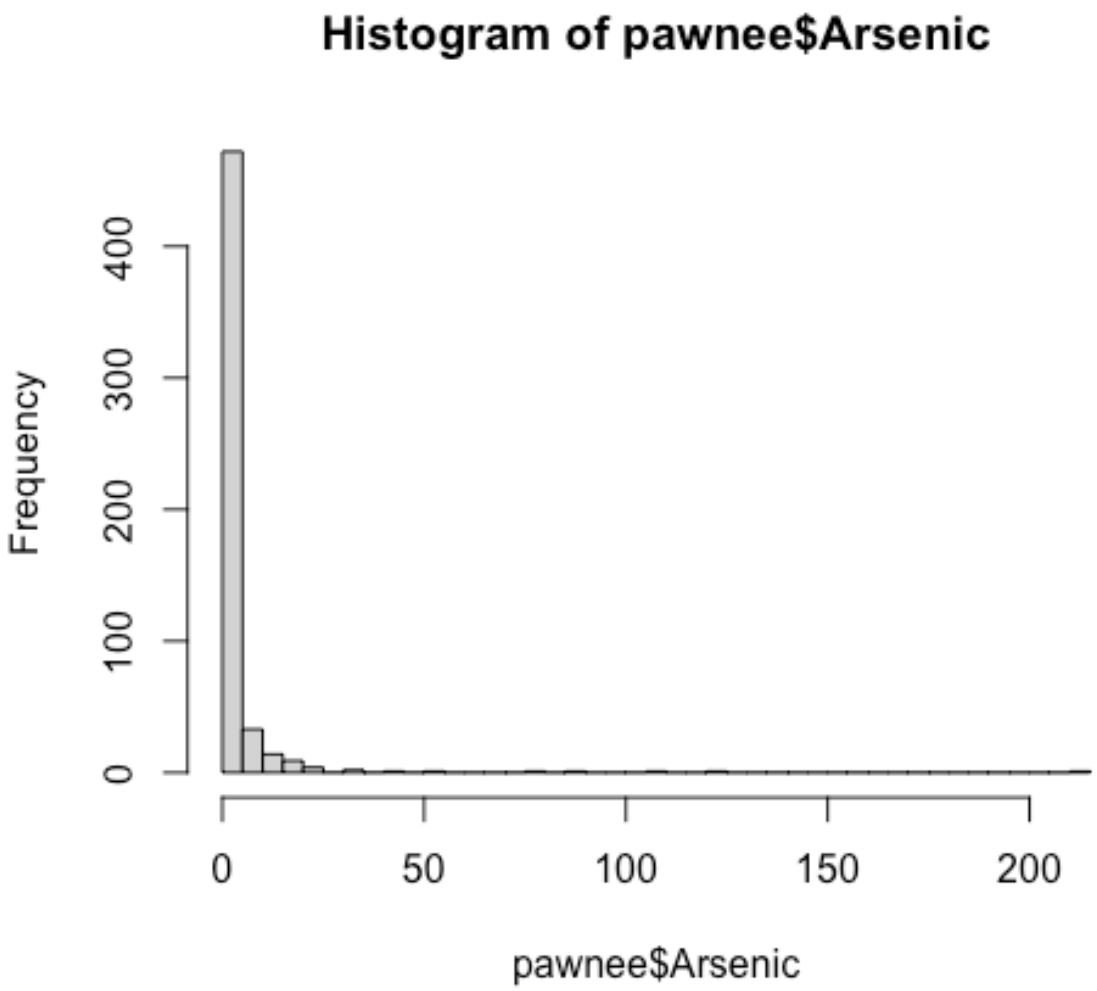
hist(xbar, prob=TRUE)
curve(dnorm(x, mean(xbar), sd(xbar)), add = TRUE)



3c

We do not believe the simulated distribution for sample arsenic means is approximately normal. First, the distribution is right skewed. Second, there seems to be some distance between the mode of our simulated distribution and the theoretical normal distribution. The theoretical normal distribution has significant chance of having negative values. This is not possible with our simulated distribution.

hist(pawnee$Arsenic, breaks=35)



The underlying distribution of arsenic is heavily skewed, so it should take more samples (larger n and N) for the Central Limit Theorem approximation to hold. This is why our result is different.

Exercise 4

pawnee <- read.csv('pawnee.csv')

4a

head(pawnee)

```
  ID Latitude Longitude Arsenic Sulfur New_hlth_issue
1  1 41.09414 -85.60974       0    0        N
2  2 41.09054 -85.70344       0    130      N
3  3 41.08601 -85.71996       4    170      N
4  4 41.08100 -85.75415       0    0        Y
5  5 41.07435 -85.70043       0    0        N
6  6 41.07399 -85.71788       0    0        N
```

dim(pawnee)

```
[1] 541   6
```

There are 541 rows and 6 columns.

4b

set.seed(1337)
sample_indices <- sample(541, size=30)
pawnee_sample <- pawnee[sample_indices,]

sample_indices

```
 [1] 147  49 210 356 425 239 126 355 350   7 524  69 502 516 334 467  73 172 127 163 419 271
43 395 257 325 368 428  97
[30] 248
```

head(pawnee_sample)

```
         ID Latitude   Longitude Arsenic Sulfur New_hlth_issue
147 147 41.03971 -85.72783           2    100            N
49      49 41.06113 -85.65553           0      0            Y
210 210 41.03178 -85.64253           0      0            N
356 356 41.01178 -85.66516           0      0            N
425 425 41.00096 -85.72899           0      0            N
239 239 41.02772 -85.72901           0      0            N
```

4c

mean(pawnee_sample$Arsenic)

[1] 0.85

mean(pawnee_sample$New_hlth_issue == 'Y')

[1] 0.2

The mean arsenic level from our sample is 0.85 ppm. The proportion of households experiencing a major health issue in the sample is about 20%.

4d

The symbol from lecture for the mean arsenic level is xbar or bar(x). The symbol from lecture for the sample proportion is phat or hat(p).

4e

The data is independent and random.

541 * mean(phats)

[1] 158.4048

541 * (1 - mean(phats))

[1] 382.5952

The data satisfies the large sample condition for CLT.

541 > 30*10

[1] TRUE

The data satisfied the big population condition for the CLT.

phat <- mean(pawnee_sample$New_hlth_issue == 'Y')
phat_sd <- sqrt(phat*(1-phat)/n)

90% Confidence Interval

phat + phat_sd * qnorm(c(0.05, 0.95))

[1] 0.07987688 0.32012312

The 90% confidence interval is (0.0799, 0.3201).

95% Confidence Interval

phat + phat_sd * qnorm(c(0.025, 0.975))

[1] 0.05686447 0.34313553

The 95% confidence interval is (0.0569, 0.3431)

99% Confidence Interval

phat + phat_sd * qnorm(c(0.005, 0.995))

[1] 0.01188802 0.38811198

The 99% confidence interval is (0.0119, 0.3881)

4f

The 100% confidence interval for the population proportions is just all possible probabilities. This is the interval [0, 1].

4g

mean(pawnee$New_hlth_issue == 'Y')

[1] 0.2920518

The true population mean of 0.292 falls within each of the confidence intervals from 4e.

4h

hist(pawnee$Arsenic, xlab='Arsenic [ppm]', main = 'Arsenic Histogram', breaks=45)