Assignment 4 Key

Makena Pollon

2023-11-22

R Markdown

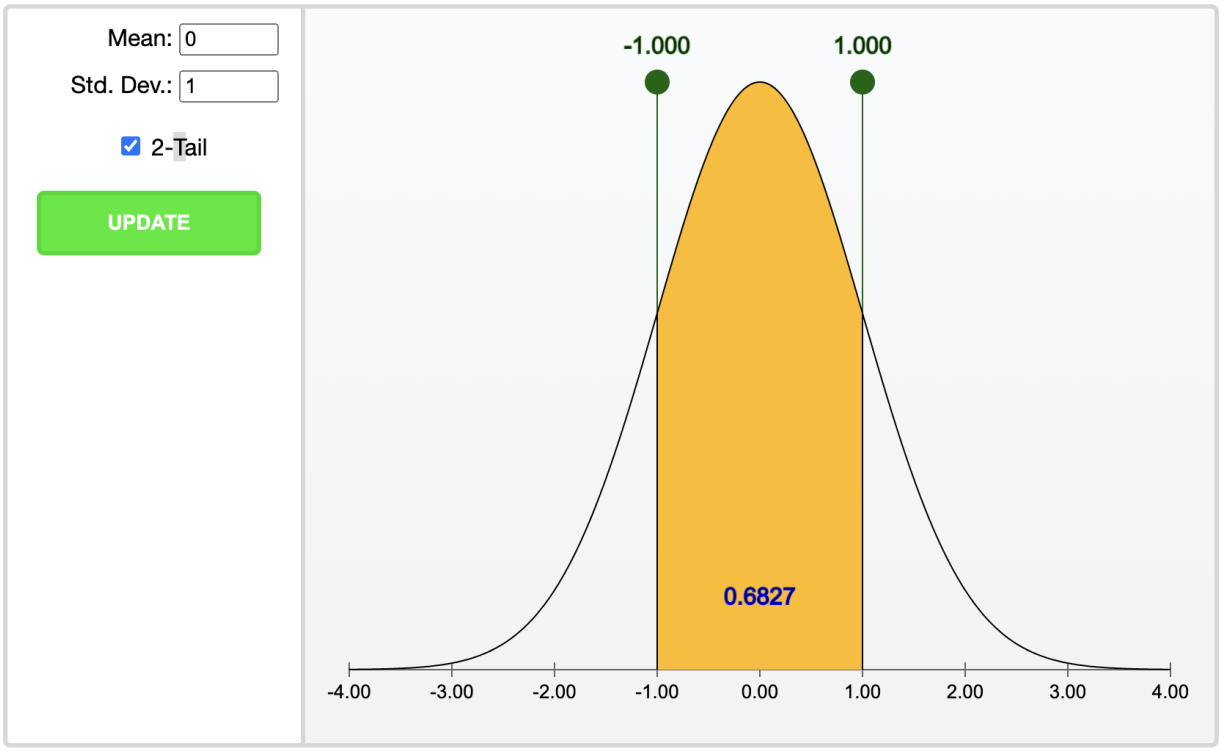
Part 1

Exercise 1

Use the applet: https://digitalfirst.bfwpub.com/stats_applet/stats_applet_7_norm.html

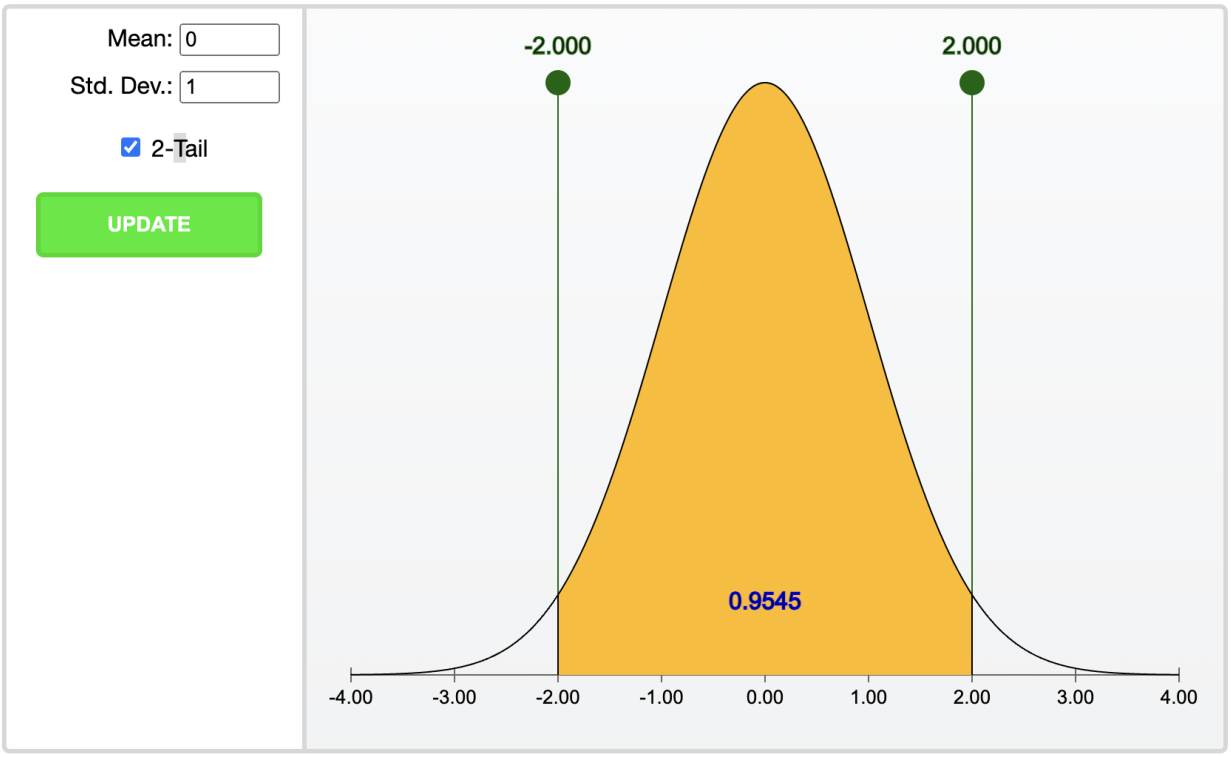
1. Set the mean to 0 and the standard deviation to 1.
2. The numbers on the horizontal axis represent the number of SD above or below the mean. So, 0 is the mean, +1 is one SD above the mean, -1 is one SD below the mean etc.

a. Place the flags 1 standard deviation on either side of the mean. What is the area between these two values? What does the empirical rule say this area is?

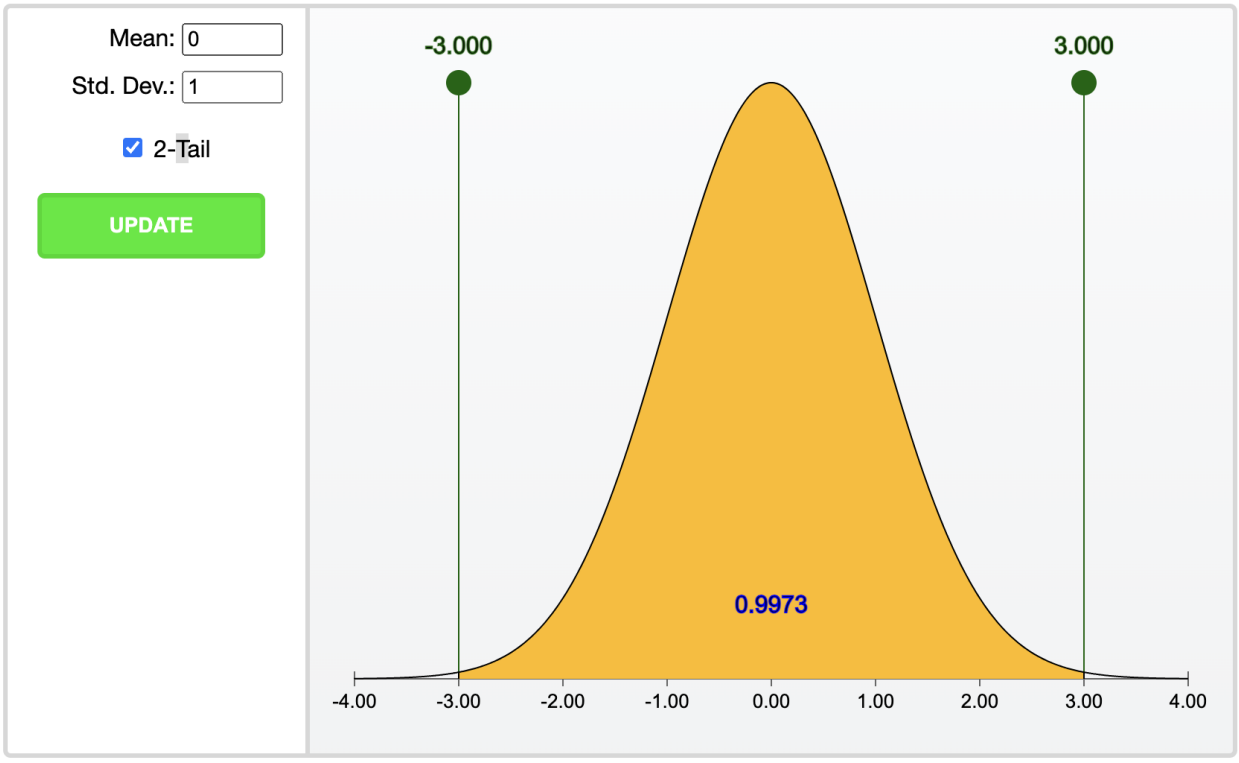


The area between these two values is around 0.6827. The empirical rule says this area is 0.68.

b. Repeat for 2 and 3 standard deviations on either side of the mean. Again, compare the empirical rule with the area given in the applet.

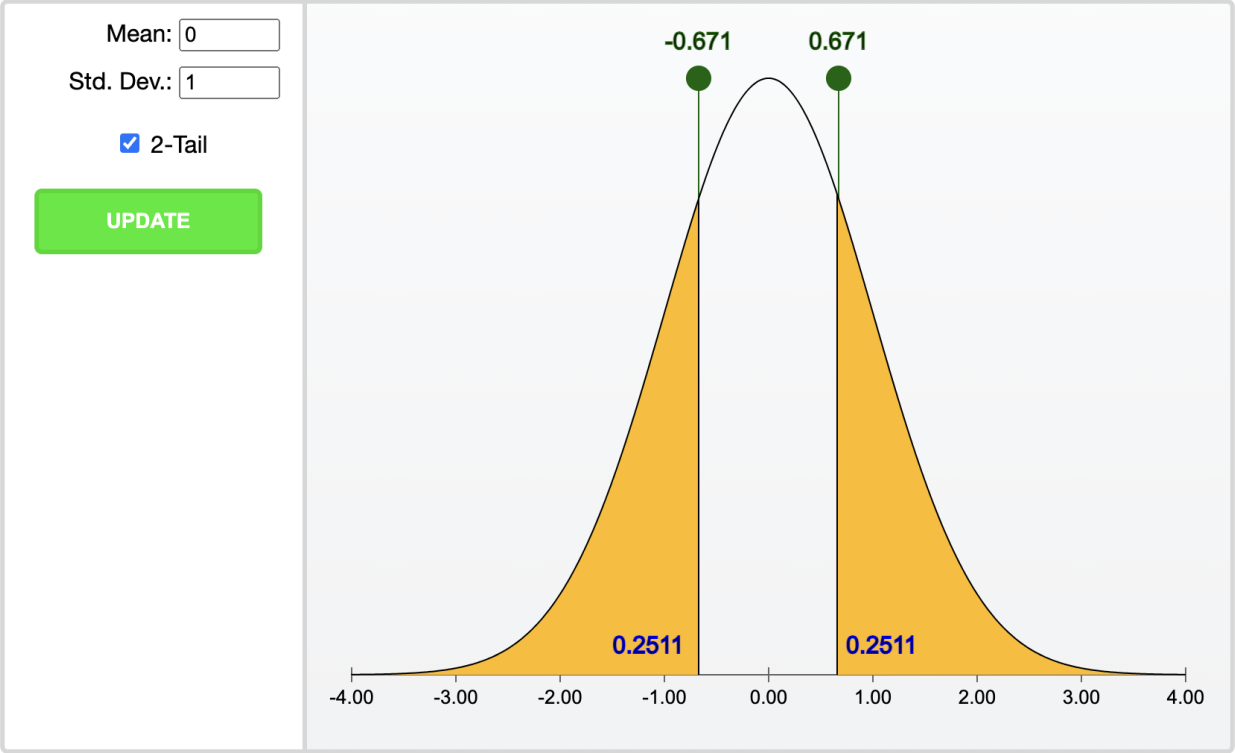


The applet says the area between the values two standard deviations from the mean is 0.9545. The empirical rule says this area is 0.95.



The applet says the area between the values two standard deviations from the mean is 0.9973. The empirical rule says this area is 0.997.

c. Using the applet, how many standard deviations above and below the mean do the quartiles of any normal distribution lie? Use the closest available values (the applet can't hit every value exactly).



Proportionally, 0.25 of our values lie below Q1 and .25 of our values lie above Q3. The closest our applet can get to these proportions is 0.2511 on either side. From here, we see this roughly equates to Q1 and Q3 being around 0.671 standard deviations from the mean.

Exercise 2

Adult male height (X) follows (approximately) a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches. Use R to find the answers for the following questions.

a. What proportion of males are less than 65 inches tall? In other words, what is P(X < 65)?

```
pnorm(65, mean = 69, sd = 2.8)
```

```
## [1] 0.07656373
```

b. What proportion of males are more than 75 inches tall? In other words, what is P(X > 75)?

```
pnorm(75, mean = 69, sd = 2.8, lower.tail = FALSE)
```

```
## [1] 0.01606229
```

c. What proportion of males are between 66 and 72 inches tall? In other words, what is P(66 < X < 72)?

```
pnorm(72, mean = 69, sd = 2.8) - pnorm(66, mean = 69, sd = 2.8)
```

```
## [1] 0.7160232
```

Exercise 3

Suppose adult male height follows a normal distribution with a mean of 69 inches and a standard deviation of 2.8 inches. Use R to find the answers for the following questions.

\textbf{a. How tall must a male be in order to be among the shortest 0.5% of males?}

```
qnorm(.005, mean = 69, sd = 2.8)
```

```
## [1] 61.78768
```

```
qnorm(.0025, mean = 69, sd = 2.8, lower.tail = FALSE)
```

```
## [1] 76.85969
```

Exercise 4

a. Run the entire chunk of code in the lab 4 section 3 to run a "for loop" that creates a vector of sample proportions. Using the results, create a relative frequency histogram of the sampling distribution of sample proportions. Superimpose a normal curve to your histogram.

Import data:

```
pawnee <- read.csv("pawnee.csv")
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
```

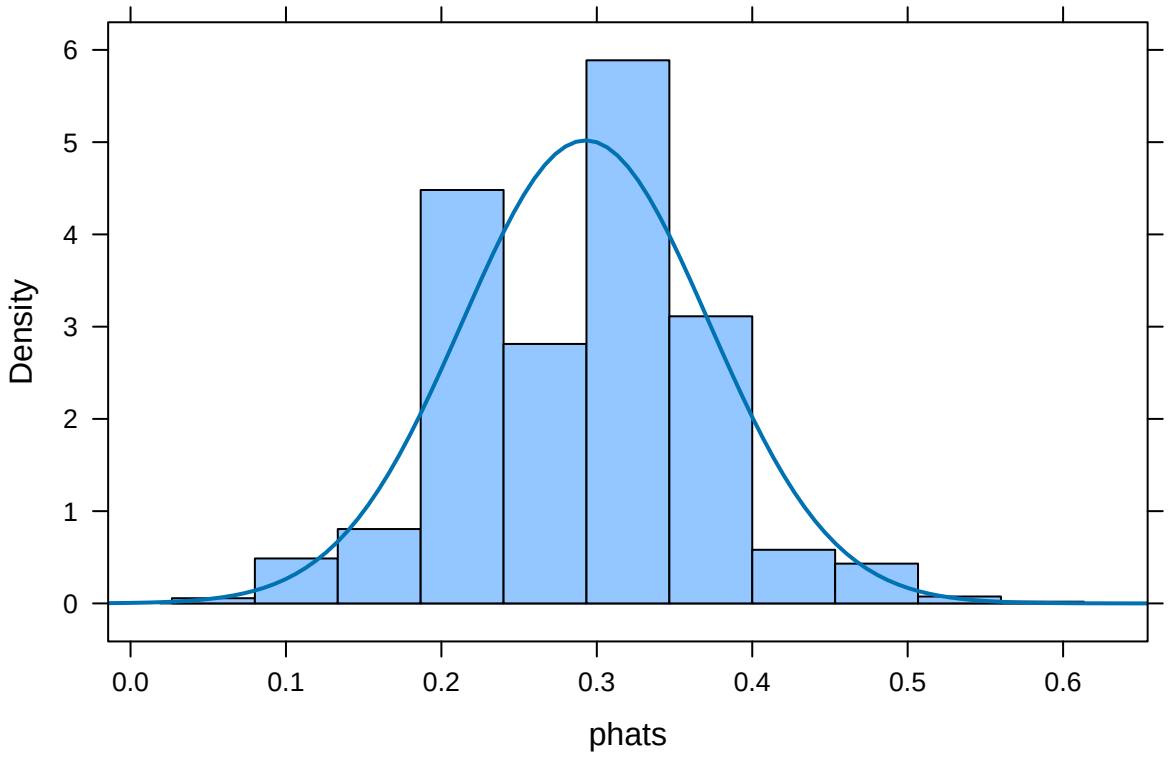
Loop from Lab 4 Manual:

```
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)
```

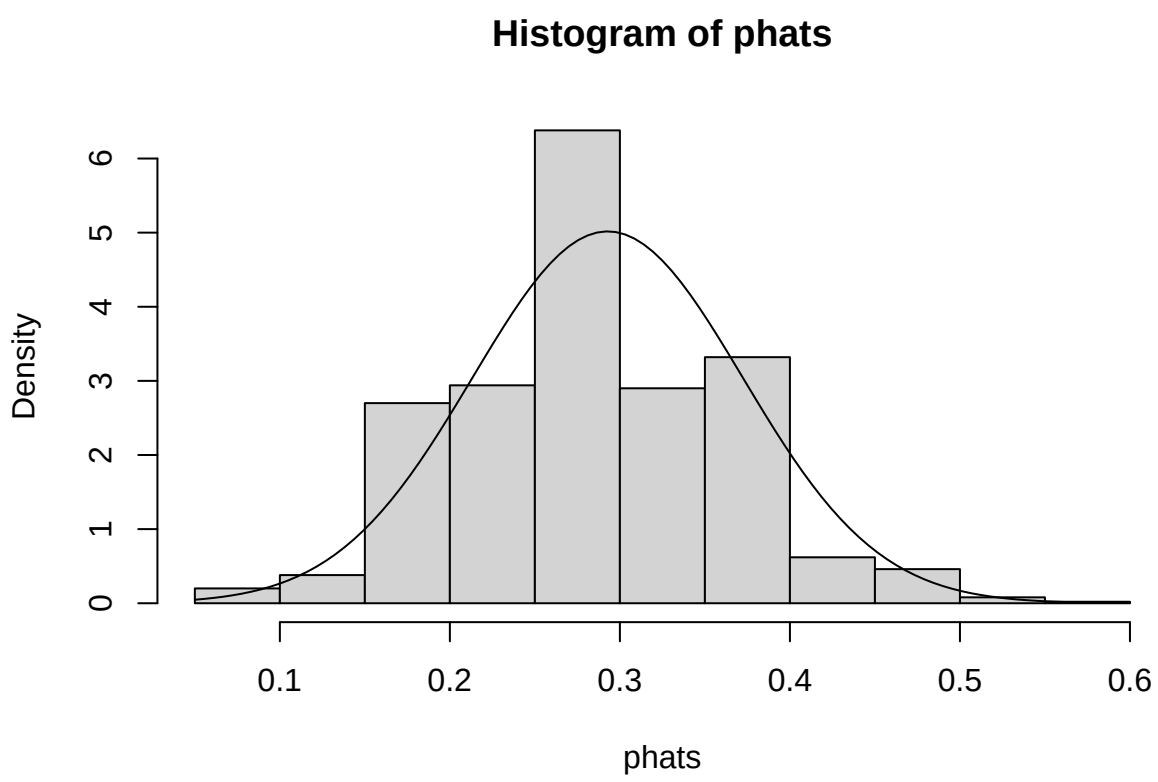
```r
# Always set the seed OUTSIDE the for loop.
# Now we start the loop. Let i cycle over the numbers 1 to 1000 (i.e., iterate 1000 times).
for(i in seq_len(M)){
  # The i-th iteration of the for loop represents a single repetition.
  # Take a simple random sample of size n from the population of size 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")
}
```

Either of these graphs for credit:

```r
histogram(phats, fit = "normal")
```




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



b. What is the mean and standard deviation of the simulated sample proportions?

```
mean(phats)
```

```
## [1] 0.2928
```

```
sd(phats)
```

```
## [1] 0.07951963
```

c. Do you think the simulated distribution of sample proportions is approximately normal? Explain why or why not.

Yes the distribution of sampling proportions is approximately normal because it is relatively unimodal and symmetric.

d. Using the theory-based method (i.e., normal approximation by invoking the Central Limit Theorem), what would you predict the mean and standard deviation of the sampling distribution of sample proportions to be? How close are these predictions to your answers from Part b?

By CLT, our mean and standard deviation would be $\hat{p}$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, respectively. We use the population proportion for our calculations because our *hatp* is unknown in this case.

```r
p = mean(pawnee$New_hlth_issue == "Y")
sdev = sqrt((p*(1-p))/n)
print(paste0("Theory-based mean: ", round(p,6), "; Theory-based sd: ", round(sdev,6)))
```

```
## [1] "Theory-based mean: 0.292052; Theory-based sd: 0.083018"
```

Compared to the predictions in Part b), the mean is very close but the standard deviation is much larger in the simulation.

Part 2

Exercise 1

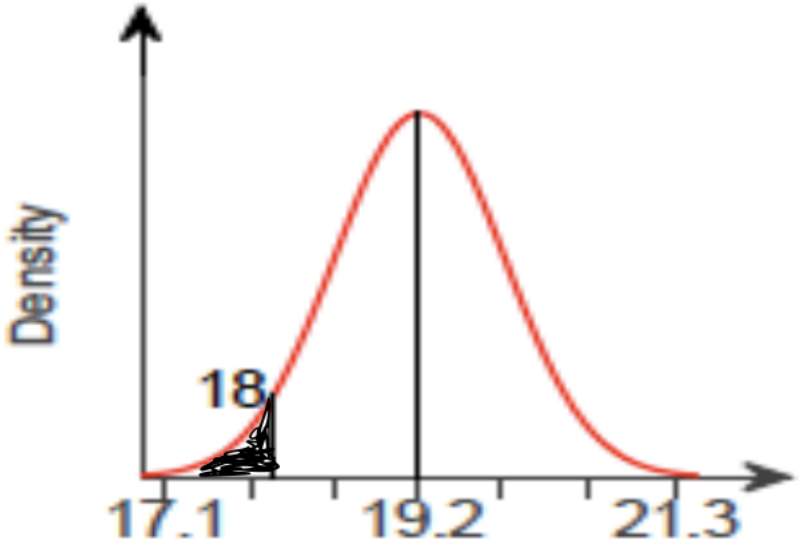
According to a statistical journal, the average length of a newborn baby is 19.2 inches with a standard deviation of 0.7 inches. The distribution of lengths is approximately normal. Use your knowledge about normal distribution to answer questions below. (Round to four decimal places as needed.)

a.What is the probability that a newborn baby will have a length of 18 inches or less? Shade the area of the graph that represents the probability and find the corresponding value.

Standardizing data:

$$z = \frac{x - \mu}{\sigma} = \frac{18 - 19.2}{0.7} \approx -1.71$$

Using Z-Table to find $P(Z \leq -1.71)$ gives us 0.0436. Therefore, the probability that newborn baby will have a length of 18 inches or less is around 4.36%.

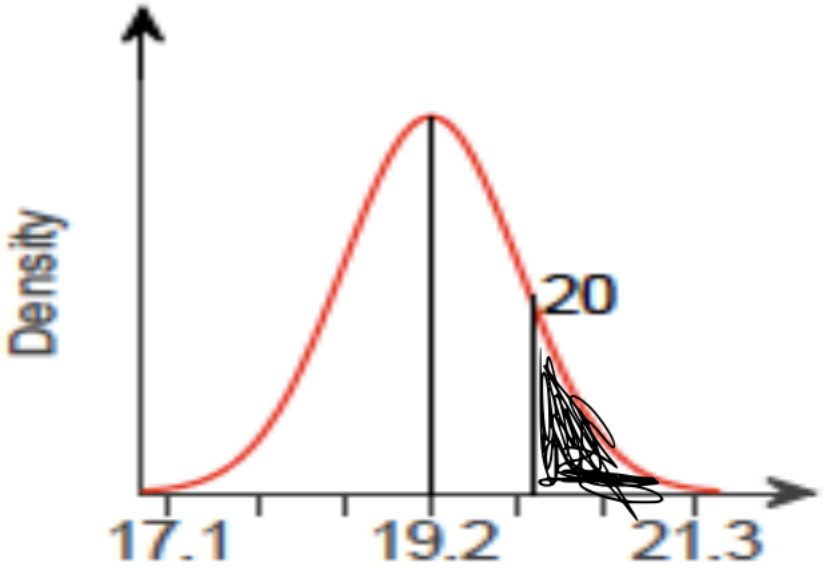


b. What percentage of newborn babies will be longer than 20 inches? Shade the area of the graph that represents the probability and find the corresponding value.

Standardizing data:

$$z = \frac{x - \mu}{\sigma} = \frac{20 - 19.2}{0.7} \approx 1.14$$

Using Z-Table to find $1 - P(Z \leq 1.14)$ gives us $1 - .8729 = 0.1271$. Therefore, the probability that a newborn baby will have a length longer than 20 inces is around 12.71%.

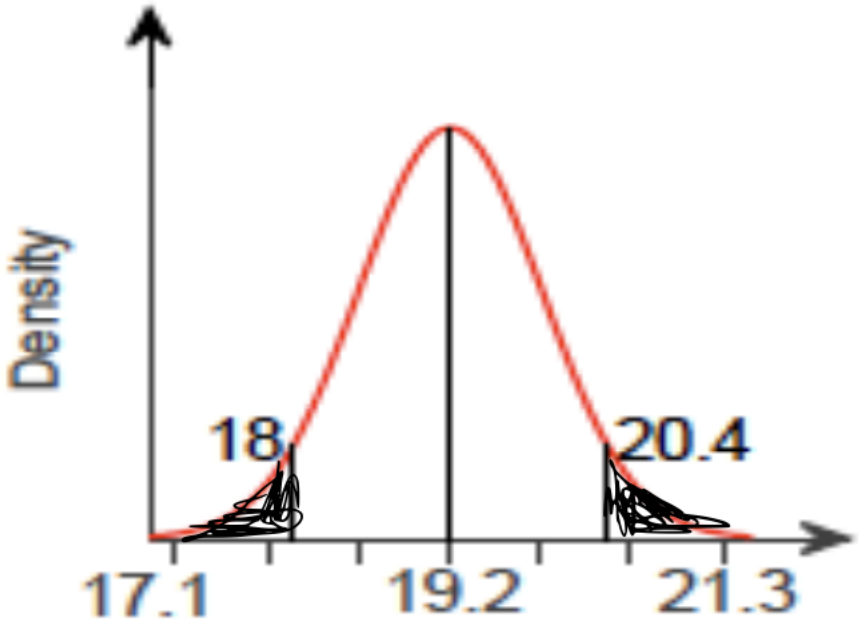


c. Baby clothes are sold in a newborn size that fits infants who are between 18 and 20.4 inches long. What percentage of newborn babies will NOT fit into the "newborn" size either because they are too long or too short?

We have the standardized statistic for the value of 18 from part a). Standardizing 20.4:

$$z = \frac{x - \mu}{\sigma} = \frac{20.4 - 19.2}{0.7} \approx 1.71$$

Subtracting $P(Z \leq -1.71)$ from $P(Z \leq 1.71)$ gives us the probability of $P(18 < X < 20.4)$ which is $0.9564 - 0.0436 = 0.9128$. $P(X < 18 \text{ or } X > 20.4)$ is the compliment of $P(18 < X < 20.4)$ so we have $1 - 0.9128 = 0.0872$. Therefore, there is an 8.72% chance of a baby being too big or too small for the baby clothes.


Density
18
20.4
17.1
19.2
21.3

Exercise 2

A school gives an entry exam for admission. Suppose the score of this exam follows a normal distribution N(400, 60). This year, the school decides to admit students who score in the top 30%. Suppose a student scored 428 on the test. Will the student be admitted? Explain your reasoning.

Answer:

Standardizing data:

$$z = \frac{x - \mu}{\sigma} = \frac{428 - 400}{60} \approx 0.47$$

This gives us a z-score probability of $P(Z \leq .47) = .6808$. Therefore, the student's score is only in the top $(1 - .6808) * 100 = 32.92\%$ of students and thus, not in the top 30% so they would not be admitted.

Exercise 3

According to a newspaper, 58% of high school seniors have a driver's license. Suppose we take a random sample of 100 high school seniors and find the proportion who have a driver's license.

a. What value should we expect for our sample proportion?

Sample proportion is equal to the amount of occurences over the total sample size. Given that 58% of high school seniors have a driver's license, we'd expect 58/100 of them or .58 of them to have their driver's license.

b. What is the standard error of the sample statistic? (Type an integer or decimal rounded to three decimal places as needed.)

$$SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{.58(1 - .58)}{100}} \approx .049$$

c. Use your answers to parts (a) and (b) to complete this sentence:

We expect 58% of the students in the sample to have their driver's license, give or take 4.9%.

d. Suppose we increased the sample size from 100 to 700. What effect would this have on the standard error? Recalculate the standard error to see if your prediction was correct. (Type an integer or decimal rounded to three decimal places as needed.)

I predict that increasing the sample size would reduce the standard error.

$$SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{.58(1 - .58)}{700}} \approx .019$$

Exercise 4

According to a survey, 58% of young Americans aged 18 to 29 say the primary way they watch television is through streaming services on the Internet. Suppose a random sample of 300 Americans from this age group is selected.

a. What percentage of the sample would we expect to watch television primarily through streaming services?

We would expect 58% of the sample to watch television primarily through streaming services.

b. Verify that the conditions for the Central Limit Theorem are met. And find the sampling distribution of the sample proportion.

Central Limit Theorem Conditions:

- 1. Random and Independent; Satisfied, stated in problem statement.
- 2. Large Sample; Satisfied; $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ for $n = 300$ and $\hat{p} = .58$.
- 3. Big Population; Satisfied; Population of American young adults is definitely larger than $10 * n$.

Since the conditions of CLT are satisfied, we know the sampling distribution follows:

$$\hat{p} \sim N(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = N(.58, \sqrt{\frac{.58 * (1-.58)}{300}}) \approx N(.58, .0285)$$

c. Would it be surprising to find that 181 people in the sample watched television primarily through streaming services? Why or why not?

Observed proportion is $181/300 \approx .603$. Calculating our Z-Score:

$$z = \frac{x-\mu}{\sigma} = \frac{.603 - .58}{.0285} \approx 0.819$$

This Z-Score is not greater than 2, thus our observed proportion would not be unusual. Therefore, it is not surprising to see that 181 people in the sample watched television in that way.

d. What is the probability of more than 65streaming services? (Type an integer or decimal rounded to three decimal places as needed.)

$$P(\hat{p} \geq .65) = P(z \geq \frac{.65 - .58}{.0285}) \approx P(z \geq 2.46) = 1 - P(z \leq 2.46) = 1 - .9931 = .0069$$

Therefore, there is a 0.69% chance that more than 65% of the sample watches TV primarily through streaming services.

Exercise 5

A survey of 800 randomly selected adults in a certain country found that 82% believed that protecting the rights of those with unpopular views is a very important component of a strong democracy.

a. Verify the Central Limit Theorem conditions.

Central Limit Theorem Conditions:

- 1. Random and Independent; Satisfied, stated in problem statement.
- 2. Large Sample; Satisfied; $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ for $n = 800$ and $\hat{p} = .82$.
- 3. Big Population; Satisfied; Can assume the population of a country is larger than $10 * n$.

b. Find a 95% confidence interval for the proportion of adults in the country who believe that protecting the rights of those with unpopular views is a very important component of a strong democracy.

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{1}$$

$$.82 \pm 1.96 \times \sqrt{\frac{.82(1-.82)}{800}} \tag{2}$$

$$.82 \pm 1.96 \times 0.0136 \tag{3}$$

$$.82 \pm .027 \tag{4}$$

Therefore, our 95% confidence interval is $[0.793, 0.847]$.

c. Would a 90% confidence interval based on this sample be wider or narrower than the 95% interval? Give a reason for your answer.

A 90% confidence interval would be narrower than the 95% confidence interval because the margin of error in the 90% CI would be smaller. This can be shown by calculating the same CI in part b) but using a 90% positive z-value i.e.

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5}$$

$$.82 \pm 1.645 \times \sqrt{\frac{.82(1-.82)}{800}} \tag{6}$$

$$.82 \pm 1.645 \times 0.0136 \tag{7}$$

$$.82 \pm .022 \tag{8}$$

Here we'd see the 90% CI is $[0.798, 0.842]$, which has a smaller range than the 95/ CI above.