STATS ASSIGNMENT 4

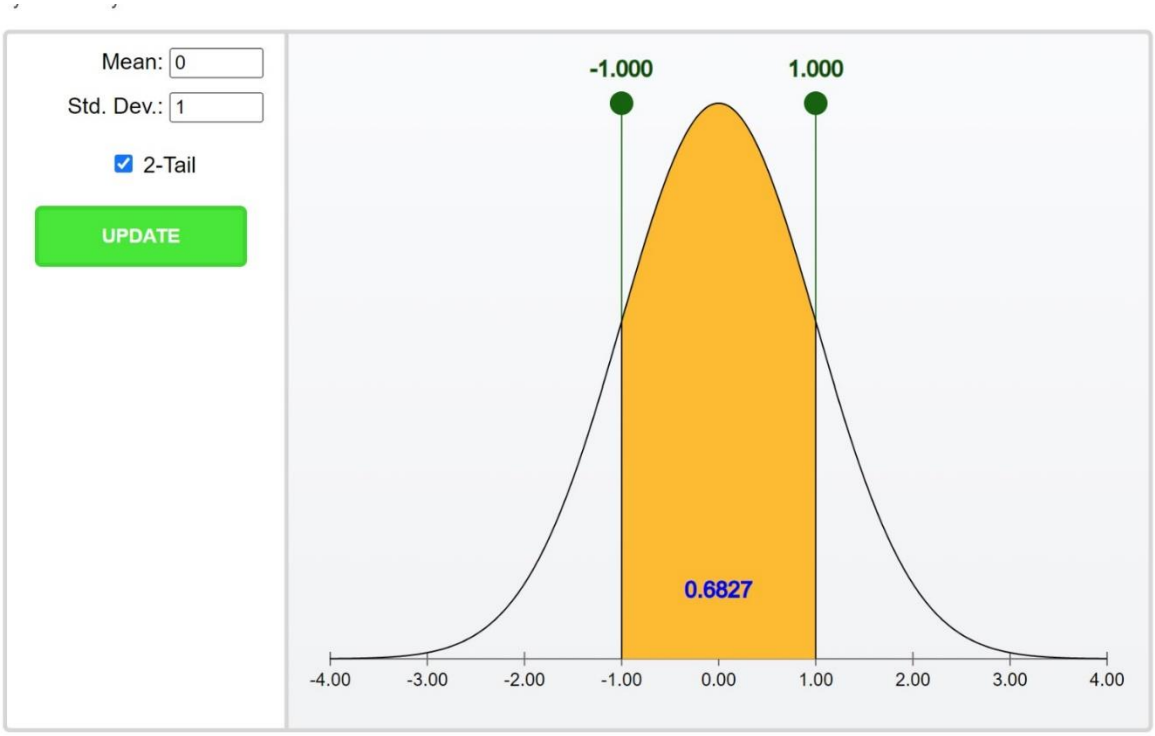
PART I

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.

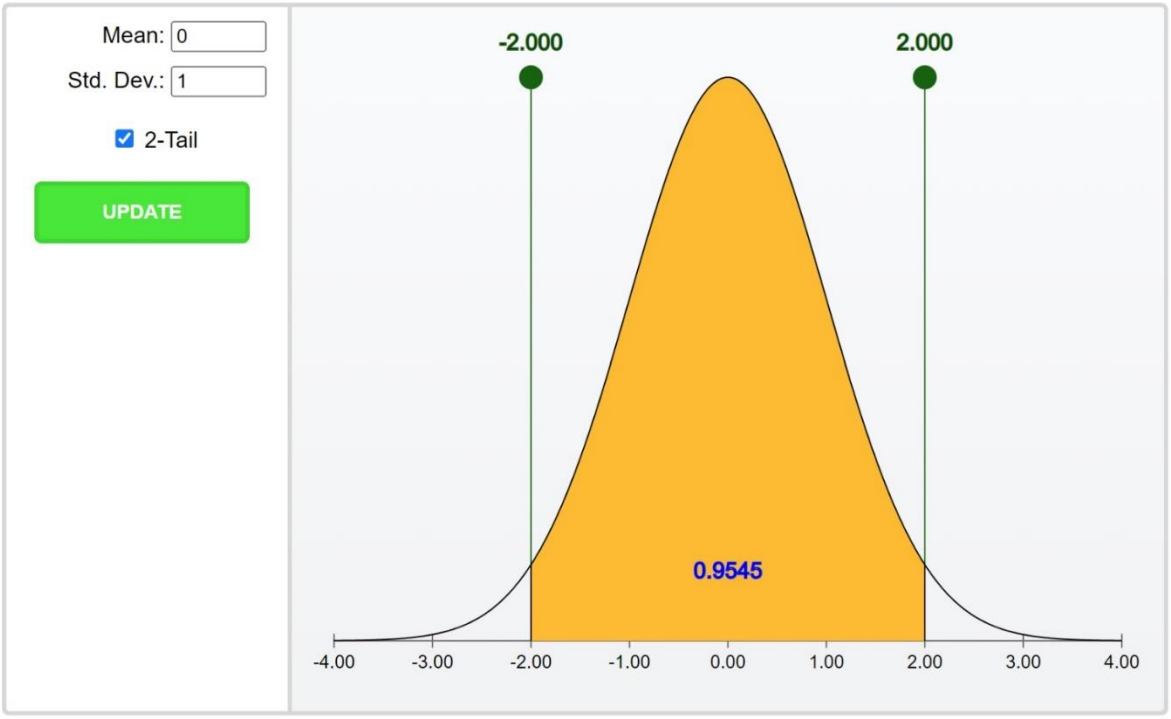
a. Place the flags 1 standard deviation on either side of the mean. What is the area between these two values? What does the empirical rule say this area is?



Area = 0.6827. 68.27% of the distribution is within 1 Standard Deviation of the mean. This is accurate as the empirical rule states that 68% of the data points will fall within one standard deviation of the mean and that can be seen here.

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.

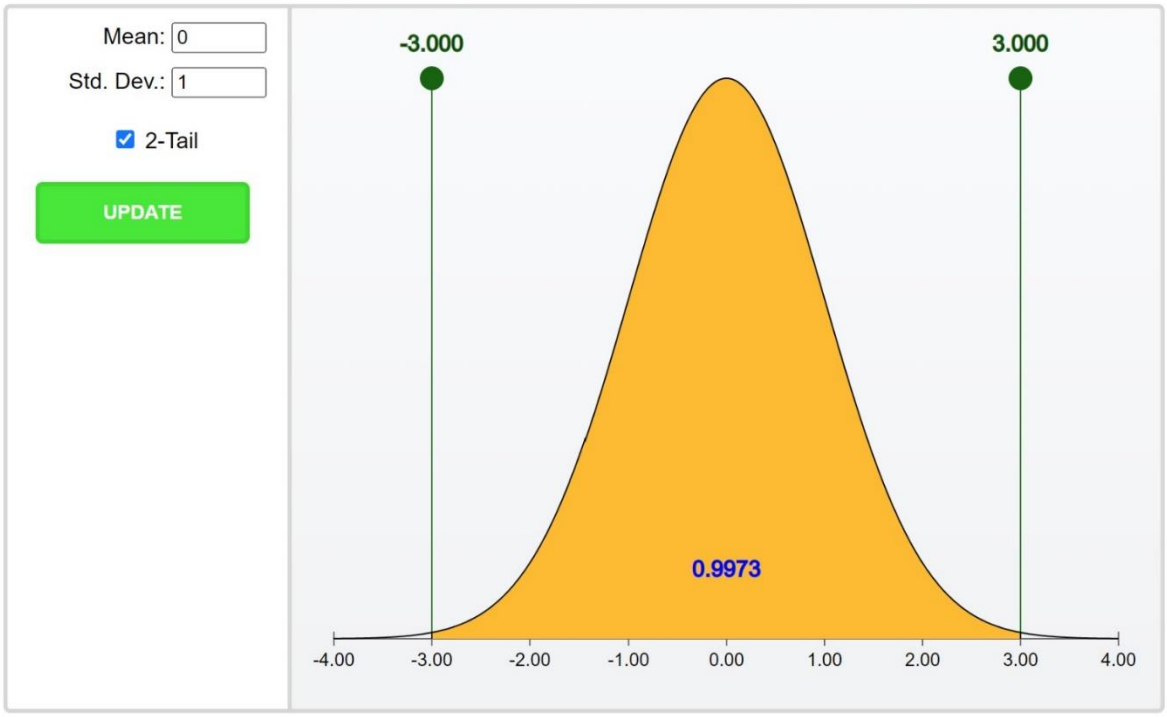
2 standard deviation:



Area: 0.9545

95.45% of the data is within 2 SDs of the mean. This is accurate and aligns with the empirical rule as it states that 95% of the data points will fall within two standard deviations of the mean.

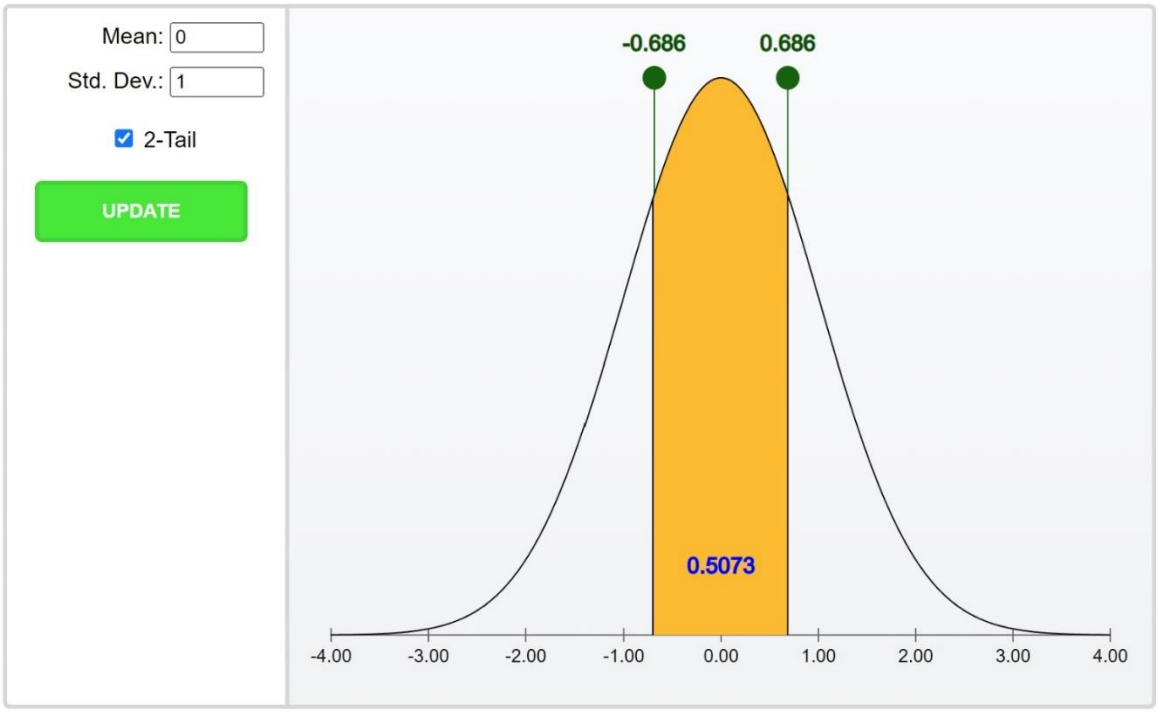
3 standard deviation:

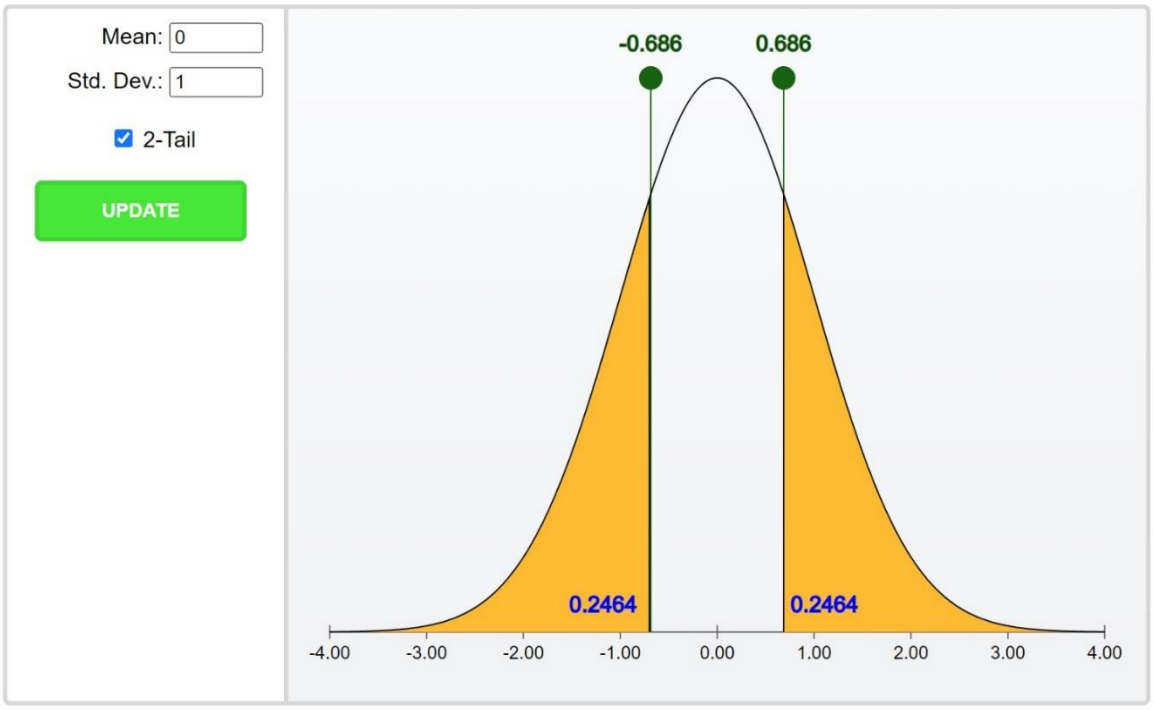


Area: 0.9973

99.73% of the data is within 3 SDs of the mean. This is accurate and aligns with the empirical rule as it states that 99.7% of the data points will fall within three standard deviations of the mean.

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).





The 1st and 3rd quartiles have 50% of the total area between them and can be ascertained by having an area value of 0.5 between the two flags on the applet.
Q1 of normal distribution is -0.686 standard deviations away from the mean as 25% of the data is below the first quartile. As normal distribution is symmetric, Q3 of normal distribution is 0.686 standard deviations above the mean as 25% of the data is above the third quartile. Q2 is 0 standard deviation away 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)?

Code:

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

Output:

```
> pnorm(65, mean=69, sd=2.8)
[1] 0.07656373
```

Hence, the proportion of males that are less than 65 inches tall = 7.66%.

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

Code:

```
1- pnorm (75,mean=69, sd=2.8 )
#OR
pnorm (75,mean=69, sd=2.8, lower.tail= FALSE )
```

Output:

```
> 1- pnorm (75,mean=69, sd=2.8 )
[1] 0.01606229
> #OR
> pnorm (75,mean=69, sd=2.8, lower.tail= FALSE )
[1] 0.01606229
```

Hence, the proportion of males that are more than 75 inches tall = 1.61%.

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

Code:

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

Output:

```
>  pnorm (72, mean=69,sd= 2.8 ) - pnorm( 66, mean= 69, sd= 2.8)
[1] 0.7160232
```

Hence, the proportion of males that are between 66 and 72 inches tall = 71.6%.

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.

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

Code:

```
qnorm(0.005, mean= 69, sd= 2.8)
```

Output:

```
>  qnorm(0.005, mean= 69, sd= 2.8)
[1] 61.78768
```

Hence, the male must be utmost 61.78768 inches or shorter tall to be among the shortest 0.5% of males.

b. How tall must a male be in order to be among the tallest 0.25% of males?

Code:

```
qnorm(1- 0.0025, mean= 69, sd= 2.8)
```

Output:

```
>  qnorm(1- 0.0025, mean= 69, sd= 2.8)
[1] 76.85969
```

Hence, the male must be at least 76.85969 inches or more tall to be among the tallest 0.25% of males.

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 with following instructions:

• If you use the histogram() function from the mosaic package, add the argument: fit = "normal".

• If you use the hist() function from base R, add the argument: prob = TRUE, then run the command: curve(dnorm(x, mean(phats), sd(phats)), add = TRUE).

Code:

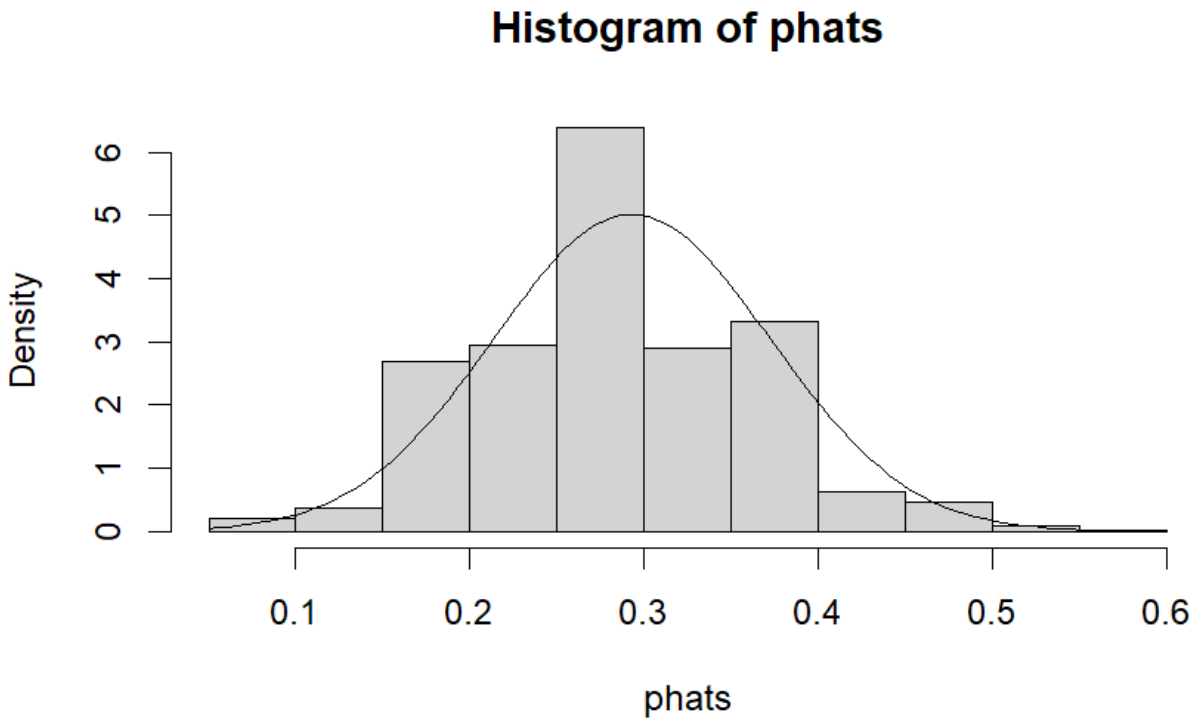
```
#4a

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

n<- 30 #the sample size
N <- 541 #the population size
M<- 1000 #number of samples/ repetitions

phats<- numeric(M) #for sample proportions

#set seed for reproduceability
set.seed(123)

#always set seed outside of the loop
#now we start the loop. let i cycle over the numbers 1 and 1000

for (i in seq_len(M)) {
  index <- sample(N, size=n)
  #save the random sample in the sample_i vector

  sample_i <- pawnee[index,]

  #compute the proportion of the ith sample of the 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)
```

Output:

```
>  pawnee <- read.csv("pawnee.csv")
>
>  n<- 30 #the sample size
>  N <- 541 #the population size
>  M<- 1000 #number of samples/ repetitions
>
>  phats<- numeric(M) #for sample proportions
>
>  #set seed for reproduceability
```

```
>  set.seed(123)
>
>  #always set seed outside of the loop
>  #now we start the loop. let i cycle over the numbers 1 and 1000
>
>  for (i in seq_len(M)) {
+    index <- sample(N, size=n)
+    #save the random sample in the sample_i vector
+
+    sample_i <- pawnee[index,]
+
+    #compute the proportion of the ith sample of the households with a ne
w health issue
+    phats[i]<- mean(sample_i$New_hlth_issue== "Y")
+  }
>
>  hist(phats, prob= TRUE)
>  curve(dnorm(x, mean(phats), sd(phats)), add= TRUE)
```

Data	
pawnee	541 obs. of 6 variables
sample_i	30 obs. of 6 variables
Values	
i	1000L
index	int [1:30] 64 141 378 421 197 57 19 343 489…
M	1000
n	30
N	541
phats	num [1:1000] 0.167 0.467 0.533 0.267 0.233 …



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

Code:

```
mean (phats)
sd(phats)
```

Output:

```
> mean (phats)
[1] 0.2928
>  sd(phats)
[1] 0.07951963
```

Hence, the mean of the simulated sample proportions = 0.2928 and the standard deviation of the simulated sample proportions = 0.07951963.

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

Yes, we expect the simulated distribution of the sample proportions to be approximately normal as it is unimodal and symmetric. It also satisfies the three conditions of the Central Limit Theorem which are random and independent observations, large sample and big population.

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?

Code:

```
#4d

#population proportion
p=mean(pawnee$New_hlth_issue == "Y")
p

#population sd
pop_sd = sqrt(p*(1-p))
pop_sd

#sampling distribution SD
sampling_sd = pop_sd / sqrt(n)
sampling_sd

#checking how close we were
p - mean(phats)
sampling_sd - sd(phats)
```

Output:

```
#4d
>
>  #population proportion
>  p=mean(pawnee$New_hlth_issue == "Y")
> p
[1] 0.2920518
>
>  #population sd
>  pop_sd = sqrt(p*(1-p))
>  pop_sd
[1] 0.454706
>
>  #sampling distribution SD
>  sampling_sd = pop_sd / sqrt(n)
[1] 0.08301757
>
>  #checking how close we were
>  p - mean(phats)
[1] -0.000748244
>  sampling_sd - sd(phats)
[1] 0.00349794
```

Hence,

The predicted mean is 0.2920518 which is very close to the values obtained in part b differing by 0.000748244.

The predicted standard deviation is 0.8301757 which is also very close to the values obtained in part b differing only by 0.00349794.

ASSIGNMENT 4 PART 2

27 November 2023 21:34

Part II

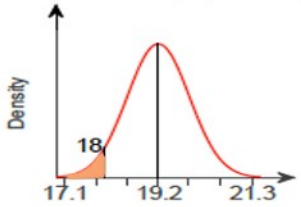
You may choose to type or write your answers electronically or scan your handwritten solutions. Please ensure that you show all steps and explanations to receive full credit, unless otherwise instructed.

In this question we are given the task to find the probability associated with z-score using the z-table.

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.



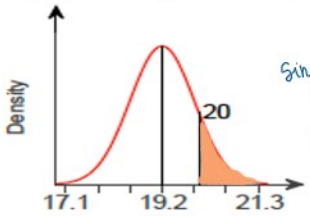
$z\text{ score} = \frac{x - \text{mean}(\mu)}{\text{sd}(\sigma)} = \frac{18 - 19.2}{0.7} = -1.714286$

$= -1.71$

(using the z score table)

the probability that a newborn baby will have length of 18 inches or less $= 0.0436 = 4.36\%$

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.



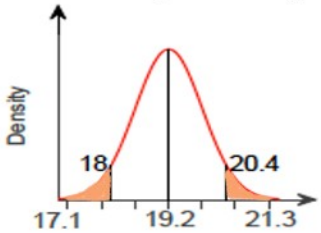
$z\text{ score} = \frac{20 - 19.2}{0.7} = 1.142857 = 1.14$

Since we want the percentage of new born babies longer than 20 inches we look at the right tail. So it is a normal distribution:

the percentage of newborn babies longer than 20 inches $= 1 - 0.8729$ (using z score table)

$= 0.1271 = 12.71\%$

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?



$z\text{ score} = \frac{18 - 19.2}{0.7} = -1.71$

the probability that a newborn baby will have length of 18 inches or less $= 4.36\%$ (from part a.)

$z\text{ score} = \frac{20.4 - 19.2}{0.7} = 1.714286 = 1.71$

the probability that a new born baby will have length longer than 20.4 inches $= 1 - 0.9564$

$= 0.436 = 4.36\%$

Hence the percentage of babies that will not fit into the "newborn" size because they are either too long or too short $= 0.872 = 8.72\%$

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.

mean = 400 sd = 60

If the student scores 428 on the test : $z\text{ score} = \frac{428 - 400}{60}$

$= 0.467 = 0.47$

% of students that scored higher than 428 = $1 - 0.6808$ (using z score tables)

$= 0.3192$

$= 31.92\%$

to be in the top 30%, the score of the student should be :

z score = 0.52 using z score table

$0.52 = \frac{x - 400}{60}$ $= 31.2 + 400$

$x = 431.2$

∴ As the school was only admitting students that were in the top 30% a student who scored 428 on the test has approximately 32% of the students who scored higher, hence the student fails to be in the top 30% and won't get admitted in the school.

The student should have scored atleast 431.2 to be in the top 30%.

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?

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

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 % (If your answer was a decimal, convert to percentage here)

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 % (If your answer was a decimal, convert to percentage here).

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.)

$p = 0.58 \qquad n = 100$

a) sample proportion $= \hat{p} = \frac{\text{no. of high school students who have a drivers licence}}{\text{no. of high school students in the sample}} = \frac{58}{100}$

$= 0.58 = 58\%$ ∴ we expect our sample proportion to be 0.58.

b) Standard error $= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.58(1-0.58)}{100}}$

$= \sqrt{0.002436}$

$= 0.049356 = 0.049$

c) 58%, 4.9 %

d) If we increase the sample size the standard error decreases

new SE $= \sqrt{\frac{0.58(1-0.58)}{700}} = \sqrt{0.000348} = 0.018655$

$= 0.0187$

$= 1.87\%$

Hence, after recalculating the new SE we can see that if we increase the sample size, the standard error significantly decreases.

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?

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

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

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

a) $\hat{p} = p = 0.58$

$SE = \sqrt{\frac{0.58\,(1-0.58}{300}} = 0.028$

we expect 58% of the sample to watch television primarily through streaming services give or take 2.8%

b) Conditions for central limit theorem:

1 → random and independent: sample should be selected randomly we must assume that the survey is well designed and the sampling is done correctly. SATISFIED.

2 → Large sample: $n = 300 \quad p = 0.58$

$np = 174 \geq 10 \qquad n(1-p) = 126 \geq 10$

Hence, this condition is also satisfied

3 → Big population: $N \geq 10n$ here, $n = 300$ so $N \geq 3000$ which is satisfied as N is the number of young Americans between the age of 18 and 29 which is obviously greater than 3000. SATISFIED

∴, All conditions of the central limit theorem are met.

Sampling distribution of the sampling proportion:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

mean $= \mu = p = 0.58$

standard error $= \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.58(1-0.58)}{300}}$

$= 0.028496$

$\hat{p} \sim N(0.58, 0.028)$

c) $z\ score = \frac{x-\mu}{\sigma}$

$\hat{p} = \frac{181}{300} = 0.603333$

$z\ score = 0.603333 - 0.58$

$$\text{Z score} = \frac{0.603333 - 0.58}{\frac{300}{0.028496}}$$

$$= 0.821$$

For a value to be surprising/unusual it should have a z score greater than 2 or less than -2. However this is not the case here. Hence, it would not be surprising to find that 181 people in the sample watched television primarily through streaming service.

d) $$\text{Z score} = \frac{0.65 - 0.58}{0.028496} = 2.456485 = 2.46$$

using the z-score table

probability associated with the z score = $1 - 0.9931$

$= 0.0069$

$= 0.69\%$

∴ the probability that more than 65% of the sample watched television primarily through streaming service is 0.007

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.

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.

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) To verify the central limit theorem, it must satisfy the following three conditions

1) Random and Independent: We must assume that the survey is well designed and the sampling is done correctly. The sample is selected randomly from the population and the observations are independent of each other.

2) Large sample - Sample size $n = 800$ and $\hat{p} = 0.82$ then $n\hat{p} = 656$ which is is greater than or equal to 10 $\therefore$, $n\hat{p} \geq 10$

$n(1-\hat{p}) \geq 10$: $800 \times 0.18 = 144 \leftarrow$ greater than 10

3) Big Population - N (which is the population size) is the number of adults in a certain country to satisfy this question $N \geq 10n$

Sample size $n = 800$. Total population $\geq 10n = 8000$ which is true as the total number of adults in a country would be greater than $10n$.

$\therefore$ all conditions of central limit theorem are met.

b) $\hat{p} = 0.82$

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.82(1-0.82)}{800}} = 0.01358$$

$z^* = 1.96$

confidence interval:

$\hat{p} \pm z^* \times SE$

$0.82 \pm 1.96 \times 0.01358$

0.82 ± 0.0266168

$0.793 < p < 0.847$

c)

z^* for 95% confidence interval is 1.96

z^* for a 90% confidence interval is 1.645

As the confidence level is controlled by z^*. An increase in z^* will produce a wider confidence interval.

Since, $1.645 < 1.96$

there is a decrease in z^*

Hence, a 90% confidence interval based on this sample will be narrower than the 95% interval.

Confidence Level	z^*
90%	1.645
95%	1.96
99%	2.576