

# STAT 630: Homework 4

Due: October 12th, 2023

1. Simulate the M&M activity in R. Let the true proportion of blue M&M's be  $p = 0.19$ . Suppose there are three different sized bags of M&M's:  $n = 25, 50$ , and  $100$ . Use the `rbinom()` function to create 1,000 different samples for each of the three different sample sizes ( $n$ ). Show your code below. You do not need to print any output.

```
p <- 0.19
n <- c(25, 50, 100)
n_samp <- 1000
samples_sz_25 <- rbinom(n_samp, 25, p)
samples_sz_50 <- rbinom(n_samp, 50, p)
samples_sz_100 <- rbinom(n_samp, 100, p)
```

2. Compute the mean and standard deviation of the sampling distribution for each sample size based on your 1000 simulations. Additionally, calculate the theoretical mean and standard deviation for each size based on the CLT for proportions. Put the information in the table below. You can use this table or create your own. If you choose to use this table, note that the vertical bars | must be lined up.

	Simulation	Theoretical
n = 25	$\bar{x}_{\hat{p}_{25}} = 4.665$ $s_{\hat{p}_{25}} = 1.948$	$\mu_{\hat{p}_{25}} = 0.19$ $\sigma_{\hat{p}_{25}} = 0.078$
n = 50	$\bar{x}_{\hat{p}_{50}} = 9.474$ $s_{\hat{p}_{50}} = 2.735$	$\mu_{\hat{p}_{50}} = 0.19$ $\sigma_{\hat{p}_{50}} = 0.055$
n = 100	$\bar{x}_{\hat{p}_{100}} = 18.894$ $s_{\hat{p}_{100}} = 3.942$	$\mu_{\hat{p}_{100}} = 0.19$ $\sigma_{\hat{p}_{100}} = 0.039$

Calculate mean and standard deviation of the sampling distribution for each sample size based on 1000 simulations

```
xbar_sz_25 <- round(mean(samples_sz_25), 3)
xbar_sz_25
```

```
## [1] 4.779
```

```
sd_sz_25 <- round(sd(samples_sz_25), 3)
sd_sz_25
```

```
## [1] 1.959
```

```
xbar_sz_50 <- round(mean(samples_sz_50), 3)
xbar_sz_50
```

```
## [1] 9.58
```

```
sd_sz_50 <- round(sd(samples_sz_50), 3)
sd_sz_50
```

```
## [1] 2.705
```

```
xbar_sz_100 <- round(mean(samples_sz_100),3)
xbar_sz_100
```

```
## [1] 18.962
```

```
sd_sz_100 <- round(sd(samples_sz_100),3)
sd_sz_100
```

```
## [1] 3.771
```

Calculate the theoretical mean and standard deviation for each size based on the CLT for proportions.

(\*) For theoretical mean ( $\mu$ ), the mean of the sample means ( $\mu$ ) is equal to the population proportion ( $p$ ) = 0.19.

```
sd_theo <- numeric(length(n))

for (i in seq_along(n)) {
  sd_theo[i] <- sqrt((p * (1 - p)) / n[i])
}

results_df_sd <- data.frame(Sample_Size = n, mu = p, Sd = sd_theo)
results_df_sd
```

```
##   Sample_Size   mu      Sd
## 1          25 0.19 0.07846018
## 2          50 0.19 0.05547973
## 3         100 0.19 0.03923009
```

3. Compare  $\sigma_{\hat{p}_{25}}$  with  $\sigma_{\hat{p}_{50}}$ . Compare  $\sigma_{\hat{p}_{25}}$  with  $\sigma_{\hat{p}_{100}}$ . How much do they differ by?

$\sigma_{\hat{p}_{25}}$  is greater than  $\sigma_{\hat{p}_{50}}$

$\sigma_{\hat{p}_{25}} - \sigma_{\hat{p}_{50}} = 0.023$

$\sigma_{\hat{p}_{25}}$  is greater than  $\sigma_{\hat{p}_{100}}$

$\sigma_{\hat{p}_{25}} - \sigma_{\hat{p}_{100}} = 0.039$

```
sd_theo[25] <- sqrt((p * (1 - p)) / 25)
sd_theo[25]
```

```
## [1] 0.07846018
```

```
sd_theo[50] <- sqrt((p * (1 - p)) / 50)
sd_theo[50]
```

```
## [1] 0.05547973
```

```
m = sqrt((p * (1 - p)) / 25) - sqrt((p * (1 - p)) / 50)
m
```

```
## [1] 0.02298045
```

```
sd_theo[100] <- sqrt((p * (1 - p)) / 100)
sd_theo[100]
```

```
## [1] 0.03923009
```

```
q = sqrt((p * (1 - p)) / 25) - sqrt((p * (1 - p)) / 100)
q
```

```
## [1] 0.03923009
```

4. Suppose that out of all Olympic athletes, 70% of them train for more than 40 hours per week. Suppose a researcher took a sample of 250 athletes.

a. What proportion of the sample would be expected to train for more than 40 hours per week?

70% of the athletes were trained:  $p = 0.7$

sample size:  $n = 250$  athletes.

Number of athletes were trained for more than 40 hrs per week of the sample: 175 athletes

proportion of the sample would be expected to train for more than 40 hours per week:  $175/250 = 0.7$

```
p_bar <- 0.7 # Probability of an athlete training for more than 40 hours per week
n_samp_sz <- 250 # Sample size
```

```
# Calculate the expected value x_bar
x_bar <- n_samp_sz * p_bar
x_bar
```

```
## [1] 175
```

```
#Calculate the proportion of the sample
prop_samp <- x_bar/n_samp_sz
prop_samp
```

```
## [1] 0.7
```

b. What is the sampling distribution of the sample proportion?

The sampling distribution of the sample proportion is the distribution of sample proportions (or sample percentages) that you would get if you took random samples of a given size from a population, and calculated the proportion of each sample.

Reference: Sampling Distribution of the Sample Proportion, p-hat. (n.d.). [https://bolt.mph.ufl.edu/6050-6052/module-9/sampling-distribution-of-p-hat/#:~:text=is%20described%20next,-The%20Sampling%20Distribution%20of%20the%20Sample%20Proportion,the%20population%20proportion%20\(p\).](https://bolt.mph.ufl.edu/6050-6052/module-9/sampling-distribution-of-p-hat/#:~:text=is%20described%20next,-The%20Sampling%20Distribution%20of%20the%20Sample%20Proportion,the%20population%20proportion%20(p).)

For example, if you were interested in the proportion of a population that use Instagram, you could not survey all of the people in the world or in a certain country to ask, so you take multiple random samples of a given place, calculate the proportion of each sample that supports the candidate, and observe the distribution of these proportions.

c. What is the probability that more than 35% of our sampled athletes train for more than 40 hours per week?

sample proportion:  $p_{\text{prob}} : 35\%$

$P_{35}$  is denoted as  $P(\hat{p} \leq 35)$

$P$  is denoted as  $P(\hat{p} > 35)$

The probability that more than 35% of our sampled athletes train for more than 40 hours per week:  
 $P(\hat{p} > 35) = 1 - P(\hat{p} \leq 35)$

```
p_bar = 0.7
```

```
n_samp_sz = 250
```

```
p_prob <- 0.35
```

```
P_35 <- round(pnorm(0.35), 3)
```

```
P_35
```

```
## [1] 0.637
```

```
P_more_than_35 <- 1 - P_35
```

```
P_more_than_35
```

```
## [1] 0.363
```

d. Calculate a 90% confidence interval for the true proportion of Olympic athletes who train more than 4

Confidence level =  $1 - \alpha = 0.9$

Alpha =  $1 - 0.9 = 0.1$

Z\_star = 1.644

```
alpha <- 1-0.9
```

```
z_star <- qnorm(1-alpha/2)
```

```
z_star
```

```
## [1] 1.644854
```

```
p_bar <- 0.7
```

```
n_samp_sz <- 250
```

```
ci_low <- p_bar - z_star * sqrt(p_bar*(1-p_bar)/n_samp_sz)
```

```
ci_low
```

```
## [1] 0.6523276
```

```
ci_high <- p_bar + z_star * sqrt(p_bar*(1-p_bar)/n_samp_sz)
```

```
ci_high
```

```
## [1] 0.7476724
```

```
qnorm(0.95)
```

```
## [1] 1.644854
```

#### 5. Simulating confidence intervals

a. Compute 1000 confidence intervals for each of the three simulation settings from Question 1.

Compute 95% confidence intervals

```
# p <- 0.19
```

```
#
```

```
# n_samp <- 1000
```

```
phats_25 <- rbinom(1000, 25, 0.19) / 25
```

```
length(phats_25)
```

```
## [1] 1000
```

```
ci_low_25 <- phats_25 - qnorm(.975) * sqrt(phats_25*(1 - phats_25)/25)
```

```
ci_high_25 <- phats_25 + qnorm(.975) * sqrt(phats_25*(1-phats_25)/25)
```

```
phats_50 <- rbinom(1000, 50, 0.19) / 50
```

```
ci_low_50 <- phats_50 - qnorm(.975) * sqrt(phats_50*(1 - phats_50)/50)
```

```
ci_high_50 <- phats_50 + qnorm(.975) * sqrt(phats_50*(1-phats_50)/50)
```

```
phats_100 <- rbinom(1000, 100, 0.19) / 100
ci_low_100 <- phats_100 - qnorm(.975) * sqrt(phats_100*(1 - phats_100)/100)
ci_high_100 <- phats_100 + qnorm(.975) * sqrt(phats_100*(1-phats_100)/100)
```

b. Plot the confidence intervals for each simulation setting. Include a line to mark true proportion. C

```
color <- rep(NA, n_samp)
```

```
# Plot the confidence intervals for simulation of size 25
```

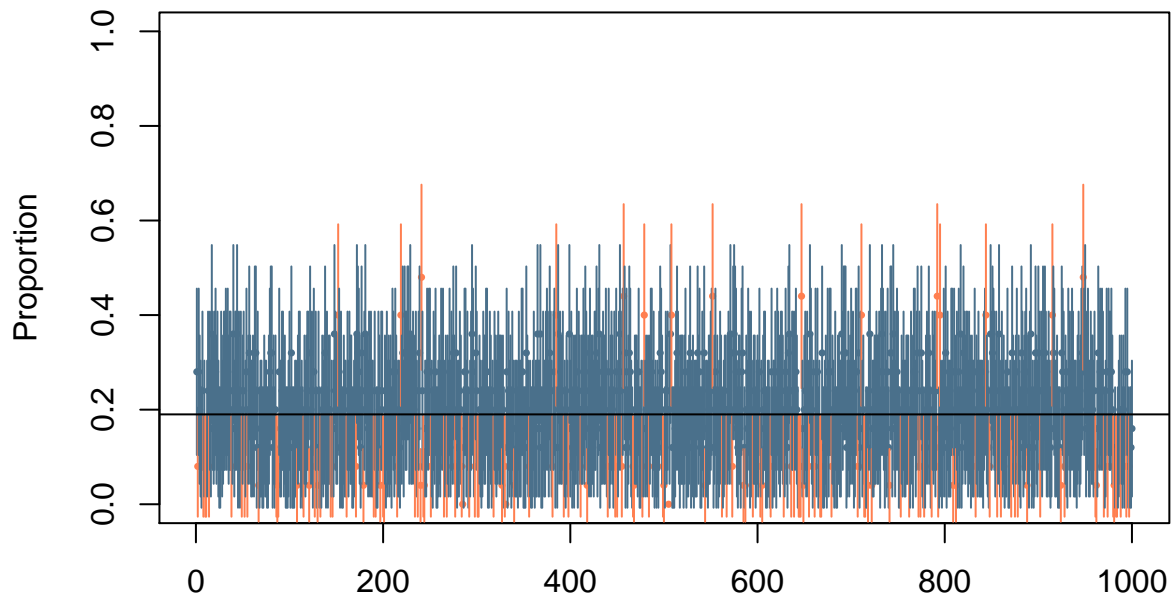
```
for(i in 1:n_samp){
  if(is.na(ci_low_25[i]) == FALSE & p > ci_low_25[i] & is.na(ci_high_25[i]) == FALSE & p < ci_high_25[i])
    color[i] <- "skyblue4"
  }
  else color[i] <- "coral"
}
table(color)
```

```
## color
##      coral skyblue4
##      141      859
```

```
x <- 1:n_samp
# print(length(x))

plot(x, phats_25, ylim = c(0,1),
     pch = 16, cex = 0.5, col = color,
     main = "1000 CI's for size 25",
     xlab = "",
     ylab = "Proportion")
segments(x, ci_low_25, x, ci_high_25,
         col = color)
abline(h = p, col = "black")
```

## 1000 CI's for size 25



```
# Plot the confidence intervals for simulation of size 50
```

```
for(i in 1:n_samp){
  if(is.na(ci_low_50[i]) == FALSE & p > ci_low_50[i] & is.na(ci_high_50[i]) == FALSE & p < ci_high_50[i]){
    color[i] <- "cyan4"
  }
  else color[i] <- "plum"
}
table(color)
```

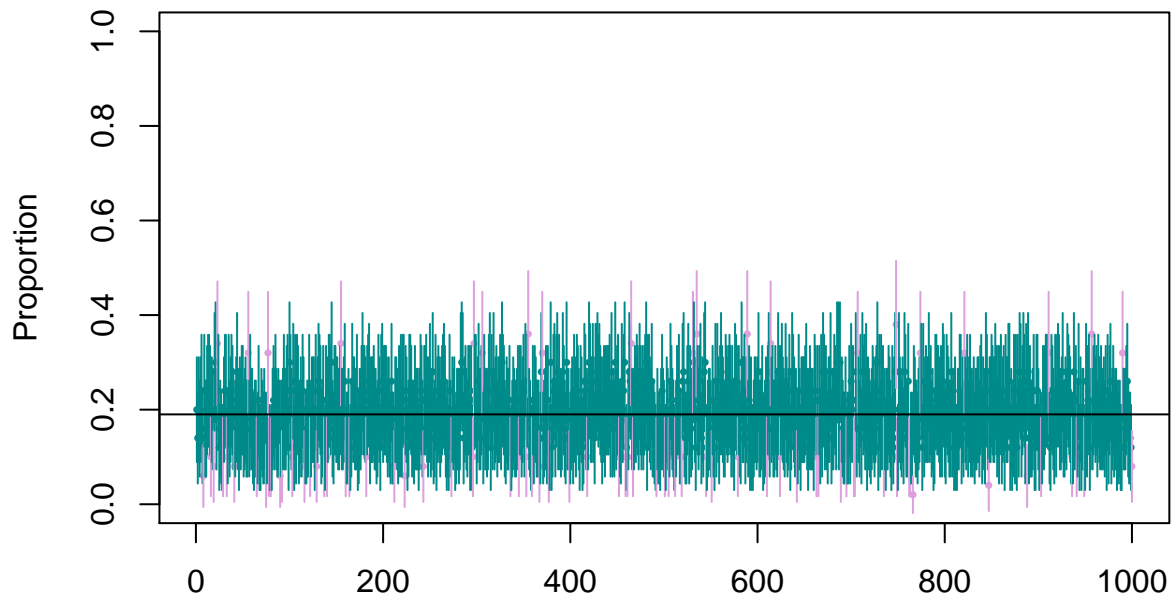
```
## color
## cyan4 plum
## 913 87
```

```
x <- 1:n_samp
print(length(x))
```

```
## [1] 1000
```

```
plot(x, phats_50, ylim = c(0,1),
     pch = 16, cex = 0.5, col = color,
     main = "1000 CI's for size 50",
     xlab = "",
     ylab = "Proportion")
segments(x, ci_low_50, x, ci_high_50,
         col = color)
abline(h = p, col = "black")
```

## 1000 CI's for size 50



```
# Plot the confidence intervals for simulation of size 100

for(i in 1:n_samp){

  if(is.na(ci_low_100[i]) == FALSE & p > ci_low_100[i] & is.na(ci_high_100[i]) == FALSE & p < ci_high_100[i]){

    color[i] <- "darkolivegreen4"
  }
  else color[i] <- "lightcoral"
}
table(color)

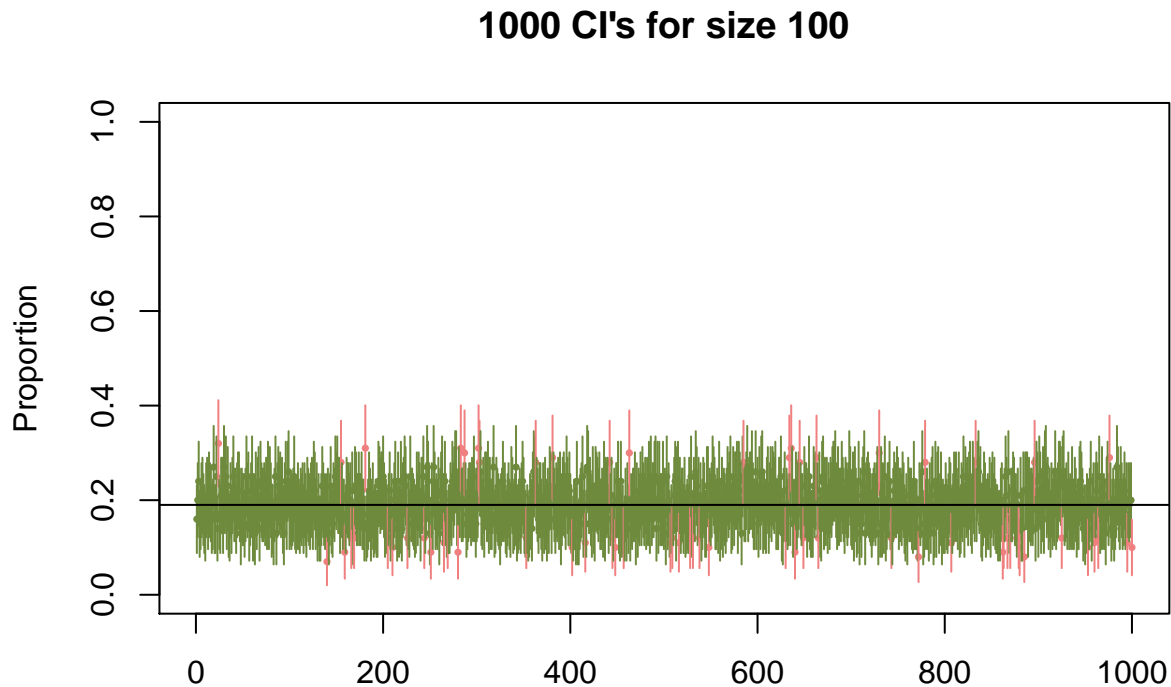
## color
## darkolivegreen4      lightcoral
##              934              66

x <- 1:n_samp
print(length(x))

## [1] 1000

plot(x, phats_100, ylim = c(0,1),
     pch = 16, cex = 0.5, col = color,
     main = "1000 CI's for size 100",
     xlab = "",
     ylab = "Proportion")
segments(x, ci_low_100, x, ci_high_100,
         col = color)
```

```
abline(h = p, col = "black")
```



c. For each simulation setting, calculate how many confidence intervals contained the true proportion.

A is denoted as the number of confidence intervals of sample size 25 contained the true proportion B is denoted as the number of confidence intervals of sample size 50 contained the true proportion C is denoted as the number of confidence intervals of sample size 100 contained the true proportion

```
p <- 0.19

#the number of confidence intervals of sample size 25 simulation contained the true proportion
A <- sum(ci_low_25 <= p & ci_high_25 >= p)
A

## [1] 859

#the number of confidence intervals of sample size 50 simulation contained the true proportion
B <- sum(ci_low_50 <= p & ci_high_50 >= p)
B

## [1] 913

#the number of confidence intervals of sample size 100 simulation contained the true proportion
C <- sum(ci_low_100 <= p & ci_high_100 >= p)
C

## [1] 934
```

6. As kitten season is coming up (yes that is a real thing), the number of newborn kittens found outside without their mom increases during this part of the year. Suppose the Hayward Humane Society has



taken in 70 kittens. They gave 10 fosters each 7 kittens. The fosters collect the weight of each of their 7 newborn kittens and calculate the mean weight for their litter.

Let the random variable  $X$  = the weight of a newborn kitten. Label each description as either a population, sample, or sampling distribution.

- The distribution of all newborn kitten weights: Population Distribution
- The distribution of 10 mean weights of newborn kittens from each foster's litter.

Sampling distribution

- The distribution of 7 newborn kitten weights cared for by a single foster.

Sample

## Sample

Recall the Bechdel dataset from class. (Link).

- Using the data, perform a 5-step hypothesis test using a confidence interval to determine if the percentage of movies that pass the Bechdel test in 2013 is different from 50%.

```
tuesdata <- tidyTuesdayR::tt_load('2021-03-02')

##
## Downloading file 1 of 1: `youtube.csv`

tuesdata <- tidyTuesdayR::tt_load(2021, week = 11)

##
## Downloading file 1 of 2: `raw_bechdel.csv`
## Downloading file 2 of 2: `movies.csv`

bechdel <- tuesdata$movies
```

### 1. Write the hypotheses.

$H_0$  : The percentage of movies that pass the Bechdel test in 2013 is equal to 50%

$H_A$  : the percentage of movies that pass the Bechdel test in 2013 is different from 50%

### 2. Check conditions.

The observational units randomly selected

Since we have  $n\_movies * p\_0 > 10$

```
p_0 <- 0.5

n_movies <- nrow(bechdel)
n_movies
```

```
## [1] 1794
```

```
n_movies*p_0
```

```
## [1] 897
```

### 3. Calculate test statistic.

```
table(bechdel$binary)
```

```
##
## FAIL PASS
## 991 803

props <- prop.table(table(bechdel$binary))
props
```

```
##
##      FAIL      PASS
## 0.5523969 0.4476031

addmargins(table(bechdel$binary))
```

```
##
## FAIL PASS Sum
## 991 803 1794

# n_movies <- nrow(bechdel)
# n_movies
# p_0 <- 0.5

xbar <- props[2]
xbar
```

```
##      PASS
## 0.4476031

se_movies <- sqrt(0.5*(1-0.5)/n_movies)
se_movies
```

```
## [1] 0.0118048

z_stat <- (xbar - p_0) / se_movies
z_stat
```

```
##      PASS
## -4.438606
```

#### 4. Calculate p-value.

```
p_values_mv <- 2*pnorm(abs(z_stat), lower.tail = FALSE)
p_values_mv
```

```
##      PASS
## 9.054324e-06
```

#### 5. Make a decision and conclude in the context of the problem.

alpha = 0.05

p\_value\_mv = 9.05 > alpha

We fail to reject  $H_0$

We do not have enough evidence that the true proportion of movies that pass the Bechdel test in 2013 is different from 50%

Bonus Question (optional): The CLT (for means) states that if we have a random sample  $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma)$ , then  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}})$ . Prove  $E[\bar{X}] = \mu$  and  $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$ . If you would like to do this by hand, you can upload a picture in addition to the .Rmd and .pdf/.doc.

(\*) Prove  $E[\bar{X}] = \mu$

We have  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

$$\Rightarrow E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right]$$

Since  $X_i$  is distributed by normal distribution  $N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}})$ ,  $E[X_i] = \mu$

$$\text{So, } E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} * n * \mu = \mu$$

$$E[\bar{X}] = \mu$$

(\*) Prove  $SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

$$X_1, X_2, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma)$$

$$\text{So, } \text{Var}(X) = \text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$$

$$\Rightarrow \text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) \quad (\text{Since } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i)$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} * n * \sigma^2 = \frac{\sigma^2}{n} \quad (\text{Since } (X_i \sim N(\mu, \sigma^2) \Rightarrow \text{Var}(X_i) = \sigma^2))$$

$$\Rightarrow \sqrt{\text{Var}(\bar{X})} = SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$