Assignment 3.1

Name: Parker Christenson

Date: 09/21/2023

For this assignment, you will refer to the textbook to solve the practice exercises. **Use Python to answer any coding problems (not R, even if indicated in your textbook).** Use Jupyter Notebook, Google Colab, or a similar software program to complete your assignment. Submit your answers as a **PDF or HTML** file. As a best practice, always label your axes and provide titles for any graphs generated on this assignment. Round all quantitative answers to 2 decimal places.

Problem 3.2.

In an exit poll of 1648 voters in the 2020 Senatorial election in Arizona, 51.5% said they voted for Mark Kelly and 48.5% said they voted for Martha McSally

a) Suppose that actually 50% of the population voted for Kelly. If this exit poll had the properties of a simple random sample, find the standard error of the sample proportion voting for him.

b) Under the 50% presumption, are the results of the exit poll surprising? Why? Would you be willing to predict the election outcome? Explain by (i) conducting a simulation; (ii) using the value found in (a) for the standard error.

```
(a) Your answer goes here
```

If the exit poll has the properties of a simple random sample, the standard error SE of the sample proportion p can be calculated using the formula for the standard error of a proportion:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Where:

- p is the true proportion of the population that voted for Kelly (in this case, 0.5 or 50%)
- n is the sample size (in this case, 1648 voters)

Our Test for the Standard Error of the Sample:

$$SE = \sqrt{\frac{0.5 \times 0.5}{1648}}$$

$$SE = \sqrt{\frac{0.25}{1648}}$$

$$SE = \sqrt{0.0001515151515...}$$

$$SE \approx 0.0123$$

So the standard error of the sample proportion voting for Mark Kelly would be approximately 0.0123. With the super small $Standard\ Error$ the test was accurate.

`(b) Your answer goes here`

(I) Conducting the Simulation

In [1]:
```
import numpy as np

# vals
n= 1648
p = 0.5
observed_p = 0.515
num_sims = 100000

# running the sims
extreme_counts = 0
for _ in range(num_sims):
    # Generate a binomial sample
    sample_votes_for_kelly = np.random.binomial(n, p)
    sample_proportion = sample_votes_for_kelly / n

    # Checking to see the pro
    if sample_proportion >= observed_p:
        extreme_counts += 1

# Calc the proportion of sims where the proportion is as extreme as the observed one
proportion_extreme = extreme_counts / num_sims
print(f"The proportion of sims with a sample as the {observed_p} is approx: {proportic
```

```
The proportion of sims with a sample as the 0.515 is approx: 0.11398
```

ii) Using Standard Error

Using the standard error value calculated in part (a), a Z-score can be computed to understand how extreme the observed sample proportion is under the null hypothesis.

The Z-score is calculated using the formula: $Z = \frac{\hat{p} - p_0}{SE}$

Where:

- $\hat{p} = 0.515$ (sample proportion or observed proportion of votes for Kelly)
- $p_0 = 0.5$ (null hypothesis proportion or expected proportion of votes for Kelly)
- $SE \approx 0.0123$ (standard error)

After calculating the Z-score, refer to the standard normal distribution to find the probability $p - value$ of observing a sample proportion as extreme as 51.5%. If this probability is very low,

typically below 0.05 we can reject the null hypothesis and conclude that the exit poll result is surprising.

Conclusion

Based on the results of the simulation and the calculated $p-value$, we decide if the observed proportion is surprising under the presumption of 50%. With that being said, if the results are significant it could inform our prediction regarding the election outcome, while considering the inherent limitations and assumptions in the analysis.

Problem 3.3.

The 49 students in a class at the University of Florida made blinded evaluations of pairs of cola drinks. For the 49 comparisons of Coke and Pepsi, Coke was preferred 29 times. In the population that this sample represents, is this strong evidence that a majority prefers Coke? Use a simulation of a sampling distribution to answer.

`Your answer goes here`

In [2]:
```
# vals
n = 49
observed_coke_preferences = 29
num_simulations = 100000
p = 0.5

# running the simulation
extreme_counts = 0
for _ in range(num_simulations):
    sample_coke_preferences = np.random.binomial(n, p)

    # checking the observed pref
    if sample_coke_preferences >= observed_coke_preferences:
        extreme_counts += 1

# calculating the proportion of sims where the pref is as extreme as the observed
proportion_extreme = extreme_counts / num_simulations

print(f"The proportion of simulations with a preference for Coke as extreme as {observ
```

```
The proportion of simulations with a preference for Coke as extreme as 29 is approxim
ately 0.1274.
```

Problem 3.5.

The example in Section 3.1.4 simulated sampling distributions of the sample mean to determine how precise $\bar{Y}$ for $n = 25$ may estimate a population mean μ.

a) Find the theoretical standard error of $\bar{Y}$ for the scenario values of $\sigma = 5$ and 8. How do they compare to the standard deviations of the 100,000 sample means in the simulations?

```
Your answer goes here
```

To find the theoretical standard error of the sample mean, $\bar{Y}$, the following formula is used:

$$SE(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

Where:

- $SE(\bar{Y})$ is the standard error of the sample mean.
- σ is the population standard deviation.
- n is the sample size.

Given $\sigma = 5$ and $\sigma = 8$, (assuming the sample size, (n), is known), you can substitute the values to find the theoretical standard error of $\bar{Y}$ for each case. After finding the theoretical standard errors, you can then compare them to the standard deviations of the 100,000 sample means obtained from the simulations to see how well the theoretical values align with the empirical ones.

Example Calculation

For $\sigma = 5$ and a known sample size n:

$$SE(\bar{Y}) = \frac{5}{\sqrt{n}}$$

For $\sigma = 8$ and a known sample size n:

$$SEbarY) = \frac{8}{\sqrt{n}}$$

b) In the first scenario, we chose $\sigma = 5$ under the belief that if $\mu = 20$, about 2/3 of the sample values would fall between 15 and 25. For the gamma distribution with $(\mu, \sigma) = (20, 5)$, show that the actual probability between 15 and 25 is 0.688.

```
Your answer goes here
```

The Gamma distribution is characterized by two parameters: the shape parameter k and the scale parameter θ. These parameters are related to the mean μ and the standard deviation σ of the distribution as follows:

$$\mu = k \cdot \theta$$

$$\sigma = \sqrt{k} \cdot \theta$$

Given $\mu = 20$ and $\sigma = 5$, we can solve for k and θ:

To find k and θ, we change the equations:

#

$$k = \left(\frac{\sigma}{\theta}\right)^2$$

#

Subbing the value of k from the first equation into the second equation to find θ:

#

$$20 = \theta \cdot \sqrt{\left(\frac{5}{\theta}\right)^2}$$

#

#

$$20 = \frac{5}{\theta} \cdot 5$$

#

$$\theta = \frac{5}{4}$$

Subbing back the value of θ to find k:

$$k = \left(\frac{5}{\frac{5}{4}}\right)^2 = 16$$

After having found the params $k = 16$ and $\theta = \frac{5}{4}$ we can now get the prob that a value from this gamma distribution falls inbetween 15 and 25. The probability can be found using the CDF of the Gamma distribution:

$$P(15 < X < 25) = P(X < 25) - P(X < 15)$$

In [3]:
```
from scipy.stats import gamma

# vals
k = 16
theta = 5 / 4

# calc the probability
lower_bound = 15
upper_bound = 25
probability = gamma.cdf(upper_bound, a=k, scale=theta) - gamma.cdf(lower_bound, a=k, s

# print statement
print(f"The actual probability that a value falls between 15 and 25 is approximately {
```

```
The actual probability that a value falls between 15 and 25 is approximately 0.688.
```

Problem 3.8.

Construct the sampling distribution of the sample proportion of heads, for flipping a balanced coin (a) once; (b) twice; (c) three times; (d) four times. Describe how the shape changes as the number of flips n increases. What would happen if n kept growing? Why?

```
Your answer goes here
```

When flipping a coin, the probability of getting a heads or a tails is both (p) is 0.5. The sampling distribution of the sample proportion of heads can be constructed for varying number of flips ($ n $. The possible values of the sample proportion for each scenario are the ratios of the number of heads to the total number of flips.

(a) Coin one time

For $n = 1$, there are only two possible outcomes, head (H) or tail (T), each with a probability of 0.5.

The Sampling distribution is: $P(H) = 0.5$ $P(T) = 0.5$

(b) Coin two times

For $n = 2$, there are only four possible outcomes: $HeadsHeads$, $HeadsTails$, $TailsHeads$, $TailsTails$. The sample proportions of heads are 1, 0.5, 0.5, 0 respectively. The sampling distribution is: $P(1) = 0.25$ $P(0.5) = 0.5$ $P(0) = 0.25$

(c) Coin three times

For $n = 3$, there are eight possible outcomes: $Heads\ Heads$ Heads,HeadsHeadsTails,Heads TailsHeads,HeadsTailsTails,TailsHeadsHeads,TailsHeadsTails,TailsTails$Heads$, $Tails\ Tails\ Tails$, with the sample proportions of heads being 1, 0.67, 0.67, 0.33, 0.33, 0.33, 0.

(d) Flipping a Coin Four times

For $n = 4$, the number of possible outcomes is $2^4 = 16$.

Shape Changes and Larger n

As the number of flips (n) increases the sampling distribution of the sample proportion of heads becomes more symmetric and will tend to show a normal distribution.

If n keeps growing, the sampling distribution would continue to approximate a normal distribution more closely, with the mean of the distribution being equal to the true proportion of heads $p = 0.5$ and the standard deviation $Standard\ Error$ (SE) given by the formula:

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

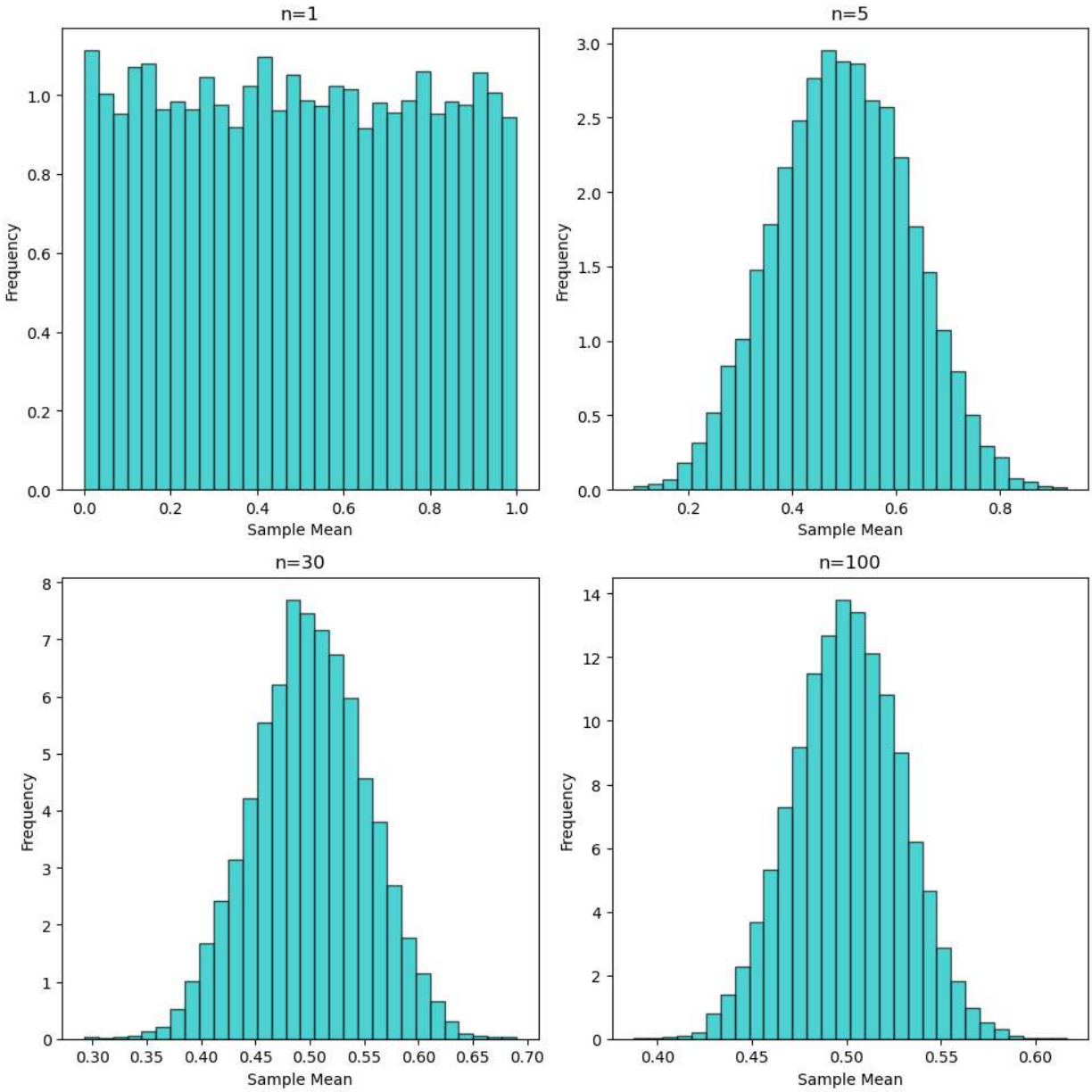
Where $SE(\hat{p})$ is the standard error of the sample proportion and $hatp$ is the sample proportion of heads. As n increases, the standard error decreases, and the sampling distribution becomes narrower, showing more probability around the true proportion of 0.5. With that being said, to get a normal looking distribution, the more n grows, the better the distribution will look. The fewer tests n is, the data will look more skewed and in some cases more scattered.

Problem 3.13.

Simulate random sampling from a uniform population distribution with several n values to illustrate the Central Limit Theorem.

`Your answer goes here`

In [4]:
```
import matplotlib.pyplot as plt

np.random.seed(0)

# number of sims
num_simulations = 10000

# sample sizes
sample_sizes = [1, 5, 30, 100]
fig, axs = plt.subplots(2, 2, figsize=(10, 10))

# flatten
axs = axs.flatten()

# for loop to iterate over different vals of 'n'
for i, n in enumerate(sample_sizes):
    # list to store the sample mean vals
    sample_means = []

    for _ in range(num_simulations):
        # pulling the 'n' samples from a uniform dist
        samples = np.random.uniform(0, 1, n)
        # bringing the samples to the list
        sample_means.append(np.mean(samples))

    # plotting the hist
    axs[i].hist(sample_means, bins=30, density=True, color='c', edgecolor='k', alpha=0
    axs[i].set_title(f'n={n}')
    axs[i].set_xlabel('Sample Mean')
    axs[i].set_ylabel('Frequency')


# showing theh plot
plt.tight_layout()
plt.show()
```



Problem 3.14.

On each bet in a sequence of bets, you win 1 dollar with probability 0.50 and lose 1 dollar (i.e., win negative 1 dollar) with probability 0.50. Let Y denote the total of your winnings and losings after 100 bets. Giving your reasoning, state the approximate distribution of Y.

`Your answer goes here`

We are going to make X the amount of winning bets, Y the total amount of winning and loosing bets.

The prob of winning will be $p = 0.5$ and the prob of loosing will be $q = 0.5$.

The Formula will look like this:

$$Y = X - (100 - X) = 2X - 100$$

Binomial Distribution Parameters:

number of trails: $n = 100$ probabilty of sucess: $p = 0.5$ probabilty of failure: $q = 0.5$

Since the odds of loosing and winning are the same, p and q are going to be equal to each other.

Binomial Distribution Characteristics are as follows:

$Mean$ = np = $100 \times 0.5 = 50$

$Variance$ = npq = $100 \times 0.5 \times 0.5 = 25$

Getting the Approximate Distribution of Y

The Central Limit Theorem states that the distribution of the sample mean of a random sample of size n taken from a population with mean μ and finite variance σ^2 will be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ for a large n.

With that being said our mean for the approx distribution will look like this:

$E(Y)$ = $2 * E\left(X\right)$ - 100 = $2 * 50$ - 100 = 0

Our Variance for the approx distribution will look like this:

$Var(Y)$ = $4 * Var\left(X\right)$ = $4 * (100 * .05 * .05)$ = 100

Lastly, our standard deviation for the approx distribution will look like this:

$\sqrt{100}$ = 10

Problem 3.15.

According to a General Social Survey, in the United States the population distribution of Y = number of good friends (not including family members) has a mean of about 5.5 and a standard deviation of about 3.9.

a) Is it plausible that this population distribution is normal? Explain.

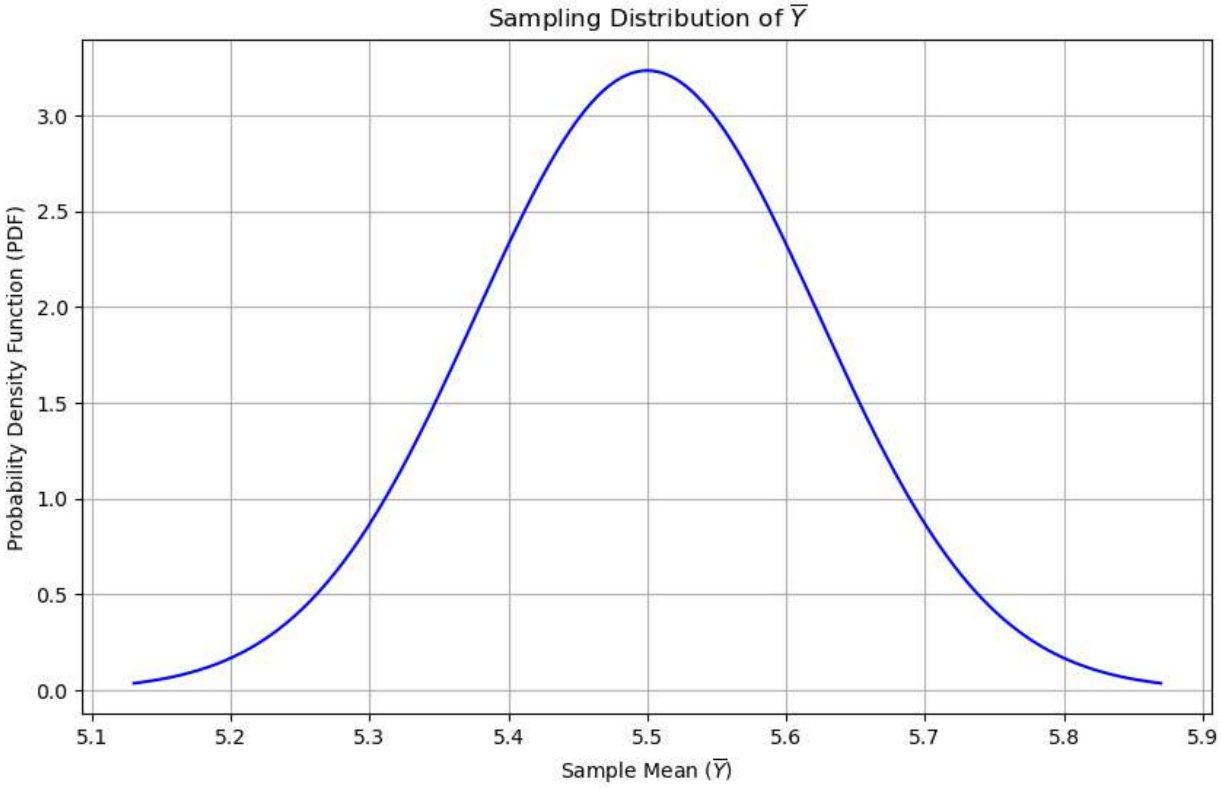
```
Your answer here
```

It is plausible to have a distribution like that with a mean of 5.5 and a standard deviation of 3.9. But, we would need to have more information on that specific data set to be able to determine if it is normal for the data set to be normal. But if you were able to get some of the figures, you might be able to use non-parametric tests to expand the data set, and see if the data is normally distributed.

b) If a new survey takes a simple random sample of 1000 people, describe the sampling distribution of $\bar{Y}$ by giving its shape and approximate mean and standard error.

Your answer here

In [6]:
```
import math
from scipy.stats import norm

# vals
mu = 5.5  #  mean
sigma = 3.9  # standard deviation
n = 1000  # sample size

# standard error
SE = sigma / math.sqrt(n)

# generating the values
x = np.linspace(mu - 3*SE, mu + 3*SE, 1000)
y = norm.pdf(x, mu, SE)

# Plot the histogram
plt.figure(figsize=(10,6))
plt.plot(x, y, color='blue')
plt.title('Sampling Distribution of $\overline{Y}$')
plt.xlabel('Sample Mean ($\overline{Y}$)')
plt.ylabel('Probability Density Function (PDF)')
plt.grid(True)
plt.show()

# Output the calculated Standard Error
print(f'Standard Error: {SE:.4f}')
```




```
Standard Error: 0.1233
```

c) Suppose that actually the mean of 5.5 and standard deviation of 3.9 are not population values but are based on a sample of 1000 people. Treating results as a simple random sample,

give an interval of values within which you can be very sure that the population mean falls. Explain your reasoning.

`Your answer here`

Confidence Interval for Population Mean

Given the mean $\bar{x}$ of 5.5, and the sample standard deviation s of 3.9, and a sample size n of 1000 people.

Standard Error

The standard error (SE) is calculated using the formula: $SE = \frac{s}{\sqrt{n}}$ $SE = \frac{3.9}{\sqrt{1000}}$

Confidence Interval

For a large sample size ($n > 30$), we can use the z-distribution to construct the confidence interval for the population mean. For being "very sure," let's construct a 99% confidence interval. The z-value for a 99% confidence interval is approximately 2.576.

The margin of error (ME) is calculated as follows: $ME = z \cdot SE$ $ME = 2.576 \cdot \frac{3.9}{\sqrt{1000}}$

Finally, the confidence interval (CI) for the population mean (μ) is given by: $CI : \bar{x} \pm ME$
$CI : 5.5 \pm 2.576 \cdot \frac{3.9}{\sqrt{1000}}$

Therefor we can be 99% confident that the sample, represents the population mean.

Explination

This interval provides a range of values within which we can be 99% confident that the true population mean number of good friends in the United States falls with the given params. With that being said, the wider the interval, the more uncertain we are about where the true population mean lies and our confidence score will go down. However, the narrower the interval, the more certain we are about the true population mean.

Problem 3.18.

Sunshine City, which attracts primarily retired people, has 90,000 residents with a mean age of 72 years and a standard deviation of 12 years. The age distribution is skewed to the left. A random sample of 100 residents of Sunshine City has $\bar{y} = 70$ and $s = 11$.

a) Describe the center and spread of the (i) population distribution, (ii) sample data distribution. What shape does the sample data distribution probably have? Why?

`Your answer here`

(i)

The Center of the population is 72 years old, which is the mean.

The spread of the population is 12 years old, which is the also the standard devation.

The shape of the age of the distribution for sunshine city is skewed to the left, because the mean of the residents is the age of 72. Which means the majority of the residents are around the age of 72.

(ii)

The Sample data distribution is 70, which is the mean age for the samples we randomly selected.

The spread of the sampple data distribution is 11, which is the standard deviation of the samples we randomly selected.

The shape of the data distribution is skewed to the left, because the mean of the samples is 70. The shape might have some left skewness, but since the data was randomly pulled, and the people in the sample are not the same as the population, the shape might be a little different and 100 people who have been randomly selected may not be able to invoke, or show the Central Limit Theorem. The data distribution might be a little different than the population distribution, just moved slightly to the right.

b) Find the center and spread of the sampling distribution of $\bar{Y}$ for $n = 100$**.** What shape does it have and what does it describe?

`Your answer here`

b) Center and Spread of the Sampling Distribution of $\bar{Y}$

Center:

The center of the sampling distribution of $\bar{Y}$ is equal to the population mean μ:

$$\mathrm{E}(\bar{Y}) = \mu = 72 \text{ years}$$

Spread:

The spread of the sampling distribution of $\bar{Y}$ is calculated using the standard error (SE) of the mean, which is given by the formula:

$$\mathrm{SE}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

Given $n = 100$ and $\sigma = 12$ years, we have:

$$\mathrm{SE}(\bar{Y}) = \frac{12}{\sqrt{100}} = \frac{12}{10} = 1.2 \text{ years}$$

Shape:

Due to the Central Limit Theorem, the sampling distribution of $\bar{Y}$ will be approximately normally distributed, regardless of the shape of the population distribution, since the sample size $n = 100$ is sufficiently large.

Description of the shape of the sampling distribution of $\bar{Y}$:

The sampling distribution of $\bar{Y}$ describes the distribution of sample means that would be obtained if we repeatedly drew samples of size 100 from the population of Sunshine City residents. It gives us an idea of how much the sample mean $\bar{Y}$ is likely to vary from multiple samples. It's providing insights into the precision of $\bar{Y}$ as an estimate of μ.

c) Explain why it would not be unusual to sample a person of age 60 in Sunshine City, but it would be highly unusual for the sample mean to be 60, for a random sample of 100 residents.

`Your answer here`

When sampling the age, it would not be unlikely to see the age of 60, because the population is skewed to the left, and the mean is 72. So the majority of the population is around the age of 72.

When sampling the mean, it would be highly unlikely to see the mean of 60, because the mean of the population is 72, and the sample size is 100. When thinking about the age of people that live in the city, 68% of the population is between the ages of 60 and 84. So it would be highly unlikely to see the mean of 60, because the mean of the population is 72, and the sample size is 100.

However, it is possible that our sample size is not large enough to fit the Central Limit Theorem. If we had a larger sample size of maybe 1000 residents, a mean of 72 would be more likely. Our ages would also have a very diffrent shape, and very well could be distributed diffrently than the population when pulling just 100 people as a sample.

d) Describe the sampling distribution of $\bar{Y}$: (i) for a random sample of size $n = 1$; (ii) if you sample all 90,000 residents.

`Your answer here`

In [ ]:

Problem 3.21.

In your school, suppose that GPA has an approximate normal distribution with $\mu = 3.0, \sigma = 0.40$. Not knowing μ, you randomly sample $n = 25$ students to estimate it. Using simulation for this application, illustrate the difference between a sample data distribution and the sampling distribution of Y .

`Your answer here`

Sample Data Distribution

In this problem it is asking for a sample data distribution. This distribution will have its own mean $\bar{y}$ and standard deviations. The estimates for the population parameters μ and σ.

Sampling Distribution of $\bar{Y}$

On the other hand, if we were to repeatedly draw different samples of 25 students and compute the sample mean for each the distribution of these sample means would form the sampling distribution of $\bar{Y}$**.

According to the Central Limit Theorem, the sampling distribution of $\bar{Y}$ will be approximately normally distributed, regardless of the shape of the population distribution, provided the sample size is rather large. With that being said, the mean of this distribution will be equal to the population mean ($\mu = 3.0$), and the $Standard\ Error$ of the distribution will be as follows:

$$\text{SE}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

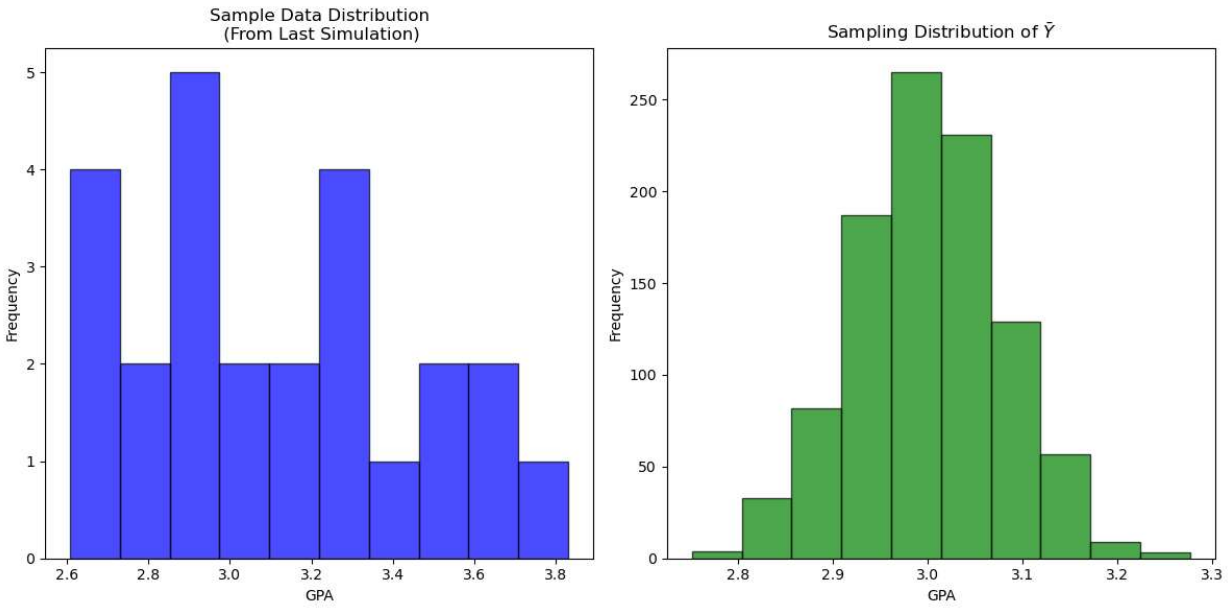
$$\text{SE}(\bar{Y}) = \frac{0.40}{\sqrt{25}}$$

$$\text{SE}(\bar{Y}) = 0.08$$

The sample data distribution represents the distribution of individual GPAs in a specific sample of students, while the sampling distribution of $\bar{Y}$** represents the distribution of sample means we would get if we repeatedly sampled different groups of 25 students from the population.

In [16]:
```
### Simulation for this problem


# vals
mu = 3.0
sigma = 0.40
n = 25
num_simulations = 1000

np.random.seed(42)
sample_means = []

# Sim the number of samples
for _ in range(num_simulations):
    sample = np.random.normal(loc=mu, scale=sigma, size=n)
    sample_mean = np.mean(sample)
    sample_means.append(sample_mean)

# plotting the hist
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.hist(sample, bins=10, color='blue', edgecolor='k', alpha=0.7)
```

```
plt.title('Sample Data Distribution\n (From Last Simulation)')
plt.xlabel('GPA')
plt.ylabel('Frequency')

# plotting the run of means
plt.subplot(1, 2, 2)
plt.hist(sample_means, bins=10, color='green', edgecolor='k', alpha=0.7)
plt.title('Sampling Distribution of $\\bar{Y}$')
plt.xlabel('GPA')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()

# calculating the mean and standard deviations
mean_sampling_distribution = np.mean(sample_means)
std_dev_sampling_distribution = np.std(sample_means)

print(f"Mean Sampling Distribution: {mean_sampling_distribution:.2f}")
print(f"Standard Deviation Sampling Distribution : {std_dev_sampling_distribution:.2f}
```




```
Mean Sampling Distribution: 3.00
Standard Deviation Sampling Distribution : 0.08
```

Mean of the Sampling Distribution of $barY$: 3.00

And

Standard Deviation of the Sampling Distribution of $\bar{Y}$: 0.08

Problem 3.26.

When sample data were used to rank states by brain cancer rates, Ellenberg (2014) noted that the highest ranking state (South Dakota) and the nearly lowest ranking state (North Dakota) had relatively small sample sizes. Also, when schools in North Carolina were ranked by their average improvement in test scores, the best and the worst schools were very small schools. Explain how

these results could merely reflect how the variability of sample means and proportions depends on the sample size.

`Your answer here`

For the smaller sample sizes, the variability of sample means and proportions depends on the sample size. The smaller the sample size, the more variability there is. The larger the sample size, the less variability there is. The smaller sample sizes in this case is the amount of people that live in south dakota. The state is not heavily populated, and the state may have abnormally high cancer rates from the farming equipment. But the sample size is so small, that it is hard to tell if the cancer rates are actually higher than the rest of the country. And in the case of Schools in North Carolina, the smaller schools may have a higher test score, but the sample size is so small, that it is hard to tell if the test scores are actually higher than the rest of the country. The rate in which the scores fluctuate is based soley off of the sample size. It would be a lot harder to move the average of a larger sample size, than it would be to move the average of a smaller sample size.