Wine Project (Part B)

The Study of Wine Quality

In [10]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import random
```

In [53]:
```
# Reading the data
df = pd.read_csv(r"E:\Linder_college\Statistical_Methods\Data_Sets\winequality-red.csv",
```

In [91]:
```
#Displaying the data
df.head()
```

Out[91]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Question 1

Suppose the population mean of the variable “density” is µ , do the following inferences:

- a. Provide an estimate of µ based on the sample
- b. Use the Central Limit Theorem (CLT) to quantify the variability of your estimate;
- c. Use the CLT to give a 95% confidence interval for µ.
- d. Use the bootstrap method to do parts b and c, and compare the results with those obtained from the CLT. State your findings.

a. Provide an estimate of µ based on the sample.

In [55]:
```
sample_density = list(df['density'])

mean_sample_density = np.mean(sample_density)

print("Estimate of µ based on the sample for variable density: ",mean_sample_density)
```

```
Estimate of µ based on the sample for variable density:  0.9967466791744841
```

Use the Central Limit Theorem (CLT) to quantify the variability of your estimate;

In [60]:
```
std_sample_density = np.std(sample_density)

#calculating the variability of estimate
var_density = std_sample_density/np.sqrt(len(sample_density))


print("Variability of estimate mean for density: ",var_density)
```

```
Variability of estimate mean for density:  4.7183339619034974e-05
```

Use the CLT to give a 95% confidence interval for μ.

In [61]:
```
#finding the confidence interval

#upper limit
print("Upper Limit of 95% Confidence interval: ",mean_sample_density+2*var_density)
```

```
Upper Limit of 95% Confidence interval:  0.9968410458537222
```

In [62]:
```
#Lower limit
print("Lower Limit of 95% Confidence interval: ",mean_sample_density-2*var_density)
```

```
Lower Limit of 95% Confidence interval:  0.996652312495246
```

95% confidence interval for density

(0.996652312495246 , 0.9968410458537222)

Use the bootstrap method to do parts b and c, and compare the results with those obtained from the CLT. State your findings.

In [65]:
```
# Bootstrap Method

sample_mean_density_bts = []
for i in range(2000):
    x = np.random.choice(sample_density, size=len(sample_density), replace=True)
    avg = np.mean(x)
    sample_mean_density_bts.append(avg)

#print(np.mean(sample_mean_density_bts))

print("Sample mean for density variability by bootstrap method ", np.mean(sample_mean_de
print("\n")

#np.std(sample_mean_density_bts)

print("Variability estimate for density variability by bootstrap method ", np.std(sample
print("\n")

np.quantile(sample_mean_density_bts, q=[0.025,0.975])

print("95 % Confidence interval for density variable using bootstrap method ",
      np.quantile(sample_mean_density_bts, q=[0.025,0.975]))
```

```
Sample mean for density variability by bootstrap method  0.9967467381769857


Variability estimate for density variability by bootstrap method  4.73972581647487e-05


95 % Confidence interval for density variable using bootstrap method  [0.99665453 0.9968
4033]
```

After compairing the variability and 95% confidence interval for density using the Central Limit Theorem and bootstrap method we see that both the methods provide almost similar results.

Question 2

- a. Provide an estimate of μ based on the sample;
- b. Noting that the sample distribution of “residual sugar” is highly skewed, can we use the CLT to quantify the variability of your estimate? Can we use the CLT to give a 95% confidence interval for μ? If yes, please give your solution. If no, explain why.
- c. Use the bootstrap method to do part b. Is the bootstrap confidence intervalsymmetric? (hint: check the bootstrap distribution; see p. 25-26 in Lecture 4).

a. Provide an estimate of μ based on the sample

In [66]:
```
# storing the values of density from the dataframe in a list

sample_sugar = list(df['residual sugar'])

mean_sample_sugar = np.mean(sample_sugar)

print("Estimate of μ based on the sample for variable residual sugar: ",mean_sample_suga
```

```
Estimate of μ based on the sample for variable residual sugar:  2.53880550343965
```

b. Noting that the sample distribution of “residual sugar” is highly skewed, can we use the CLT to quantify the variability of your estimate? Can we use the CLT to give a 95% confidence interval for μ? If yes, please give your solution. If no, explain why.

As we know that regardless of the type or skewness of the data, the sample mean approximately follows the same distribution, which is normal for the central limit theorem so we can use the CLT to quantify the variability

In [67]:
```
std_sample_sugar = np.std(sample_sugar)


#calculating the std deviation of mean
var_sugar = std_sample_sugar/np.sqrt(len(sample_sugar))

print("Variability of estimate mean for residual sugar: ",var_sugar)
```

```
Variability of estimate mean for residual sugar:  0.035248194594653 37
```

In [68]:
```
#finding the confidence interval

#upper limit

print("Upper Limit of 95% Confidence interval: ",mean_sample_sugar+2*var_sugar)
```

```
Upper Limit of 95% Confidence interval:  2.609301892628957
```

In [69]:
```
#Lower limit

print("Lower Limit of 95% Confidence interval: ",mean_sample_sugar-2*var_sugar)
```

```
Lower Limit of 95% Confidence interval:  2.468309114250343
```

95% confidence interval for residual_sugar

(2.468309114250343 , 2.609301892628957)

c. Use the bootstrap method to do part b. Is the bootstrap confidence intervalsymmetric?

In [70]:
```
# Bootstrap Method

sample_mean_sugar_bts = []
for i in range(2000):
    x = np.random.choice(sample_sugar, size=len(sample_sugar), replace=True)
    avg = np.mean(x)
    sample_mean_sugar_bts.append(avg)


print("Sample mean for residual sugar variability by bootstrap method ", np.mean(sample_
print("\n")

#np.std(sample_mean_density_bts)

print("Variability estimate for residual sugar variability by bootstrap method ", np.std
print("\n")

np.quantile(sample_mean_sugar_bts, q=[0.025,0.975])

print("95 % Confidence interval for residual sugar variable using bootstrap method ",
      np.quantile(sample_mean_sugar_bts, q=[0.025,0.975]))
```

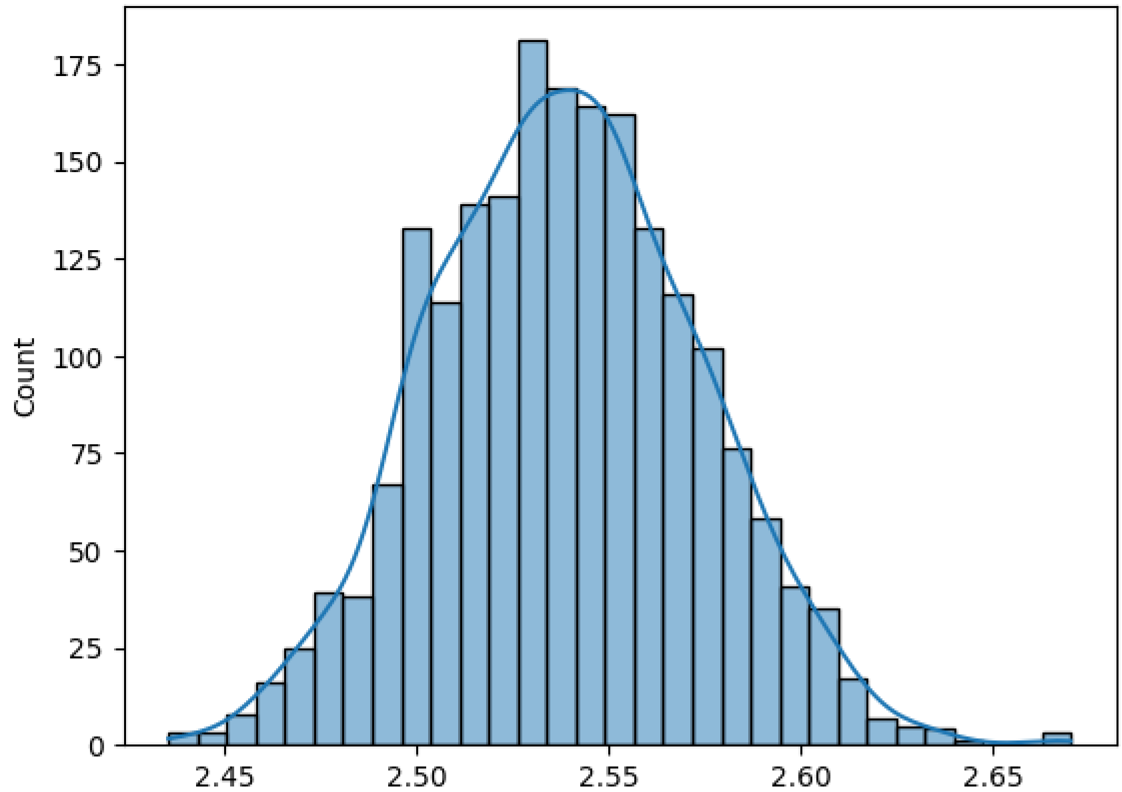
```
Sample mean for residual sugar variability by bootstrap method  2.538300594121326


Variability estimate for residual sugar variability by bootstrap method  0.0355627184389
56864


95 % Confidence interval for residual sugar variable using bootstrap method  [2.46987336
2.60771263]
```

In [37]:
```
sns.histplot(sample_mean_sugar_bts, kde= True)
```

Out[37]:
```
<Axes: ylabel='Count'>
```



After plotting the histogram for confidence interval for residual sugar we see that it is symetric

Question 3

We classify those wines as “excellent” if their rating is at least 7. Suppose the population proportion of excellent wines is p. Do the following:

- a. Use the CLT to derive a 95% confidence interval for p;
- b. Use the bootstrap method to derive a 95% confidence interval for p;
- c. Compare the two intervals. Is there any difference worth our attention?
- d. What is the maximum likelihood estimate of p and its standard error?

In [39]:
```
#Excellent wines

def get_p(row):
    if row['quality'] >= 7:
        row['Excellent_Wines'] = 1
    else:
        row['Excellent_Wines'] = 0
    return row['Excellent_Wines']

df_p = df.copy()

df_p['Excellent_Wines'] = df.apply(get_p,axis=1)
```

Use the CLT to derive a 95% confidence interval for p;

In [72]:
```
# storing the values of density from the dataframe in a list

sample_excellent_wines = list(df_p['Excellent_Wines'])

mean_sample_excellent_wines = np.mean(sample_excellent_wines)
```

```
print("Estimate of µ based on the sample for p: ",mean_sample_excellent_wines)
```

```
Estimate of µ based on the sample for p:  0.1357098186366479
```

Use the Central Limit Theorem (CLT) to quantify the variability of your estimate;

In [90]:

```
std_sample_excellent_wines = np.std(sample_excellent_wines)

#calculating the std deviation of mean
var_excellent_wines = std_sample_excellent_wines/np.sqrt(len(sample_excellent_wines))

print("Variability of estimate mean for p: ",var_excellent_wines)
```

```
Variability of estimate mean for p:  0.008564681018695619
```

Use the CLT to give a 95% confidence interval for µ.

In [74]:

```
#finding the confidence interval

#upper limit

print("Upper Limit of 95% Confidence interval: ",mean_sample_excellent_wines+2*var_excel
```

```
Upper Limit of 95% Confidence interval:  0.15283918067403915
```

In [75]:

```
#Lower limit

print("Lower Limit of 95% Confidence interval: ",mean_sample_excellent_wines-2*var_excel
```

```
Lower Limit of 95% Confidence interval:  0.11858045659925667
```

95% confidence interval for p

(0.11858045659925667,0.15283918067403915)

In [76]:

```
# Bootstrap Method

sample_excellent_wines_bts = []
for i in range(2000):
    x = np.random.choice(sample_excellent_wines, size=len(sample_excellent_wines), repla
    avg = np.mean(x)
    sample_excellent_wines_bts.append(avg)


print("Sample mean for p by bootstrap method ", np.mean(sample_excellent_wines_bts))
print("\n")

#np.std(sample_mean_density_bts)

print("Variability estimate for p by bootstrap method ", np.std(sample_excellent_wines_b
print("\n")

#np.quantile(sample_mean_density_bts, q=[0.025,0.975])

print("95 % Confidence interval for p using bootstrap method ",
      np.quantile(sample_excellent_wines_bts, q=[0.025,0.975]))


#print(np.mean(sample_excellent_wines_bts))

#np.std(sample_excellent_wines_bts)
```

```
#np.quantile(sample_excellent_wines_bts, q=[0.025,0.975])
```

```
Sample mean for p by bootstrap method  0.13572858036272673


Variability estimate for p by bootstrap method  0.00852213186985463


95 % Confidence interval for p using bootstrap method  [0.11944966 0.15322076]
```

Compare the two intervals. Is there any difference worth our attention?

Comparing the 95% confidence intervals for both bootstrap method and Central limit theorem we see that the confidence interval provided by Central Limit theorem is 11.85% to 15.28% and confidence interval provided by bootstrap is 11.94 % to 15.32%. Hence we can infer that the confidence interval provided by both methods is almost similar.