THE STUDY OF WINE QUALITY PART C

In [72]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as stats
from scipy.stats import f_oneway
from scipy.stats import chi2_contingency
```

In [2]:
```
df = pd.read_csv(r"E:\Linder_college\Statistical_Methods\Data_Sets\winequality-red.csv"
                 ,sep=';')
```

Q1. Produce summary statistics of "residual.sugar" and use its median to divide the data into two groups A and B. We want to test if "density" in Group A and Group B has the same population mean. Please answer the following questions.

Summary statistics for residual sugar

In [140…
```
df['residual sugar'].describe()
```

Out[140]:
```
count    1599.000000
mean        2.538806
std         1.409928
min         0.900000
25%         1.900000
50%         2.200000
75%         2.600000
max        15.500000
Name: residual sugar, dtype: float64
```

Dividing the data based on median of residual sugar

In [5]:
```
median_residual_sugar = np.median(df['residual sugar'])
```

In [11]:
```
def get_group(row):
    if row['residual sugar'] > median_residual_sugar:
        row['Group'] = 'B'
    else:
        row['Group'] = 'A'
    return row['Group']

df['Group'] = df.apply(get_group, axis=1)
```

In [13]:
```
df1 = df[df['Group']=='A']
df2 = df[df['Group']=='B']
```

In [133…
```
df2.describe()
```

Out[133]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH
count	716.000000	716.000000	716.000000	716.000000	716.000000	716.000000	716.000000	716.000000	716.000000
mean	8.726816	0.530957	0.304106	3.333729	0.088163	16.872207	51.234637	0.997484	3.302067
std	1.857371	0.171351	0.202590	1.796556	0.036953	11.122840	35.446287	0.001963	0.143458
min	4.700000	0.160000	0.000000	2.250000	0.012000	3.000000	7.000000	0.990200	2.740000

25%	7.300000	0.400000	0.120000	2.400000	0.073000	8.000000	24.000000	0.996287	3.200000
50%	8.300000	0.520000	0.305000	2.600000	0.082000	15.000000	43.000000	0.997455	3.300000
75%	9.900000	0.640000	0.480000	3.400000	0.093250	23.000000	71.000000	0.998700	3.390000
max	15.900000	1.185000	1.000000	15.500000	0.610000	72.000000	289.000000	1.003690	3.850000

In [16]:
```
np.mean(df1['density'])
```

Out[16]:
```
0.9961489580973952
```

In [17]:
```
np.mean(df2['density'])
```

Out[17]:
```
0.997483812849162
```

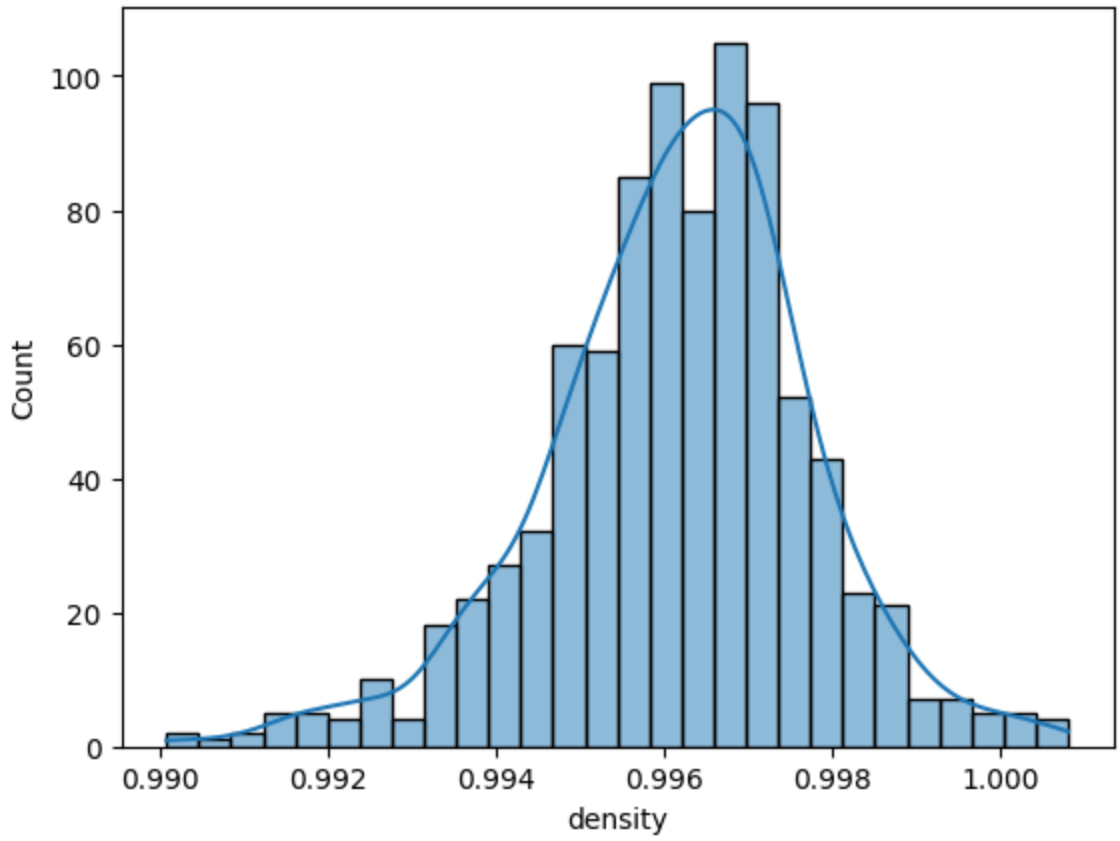
1. State the null hypothesis

- Null Hypothesis: There is no difference in the population mean of Density for Group A and Group B

Use visualization tools to inspect the hypothesis. Do you think the hypothesis is right or not?

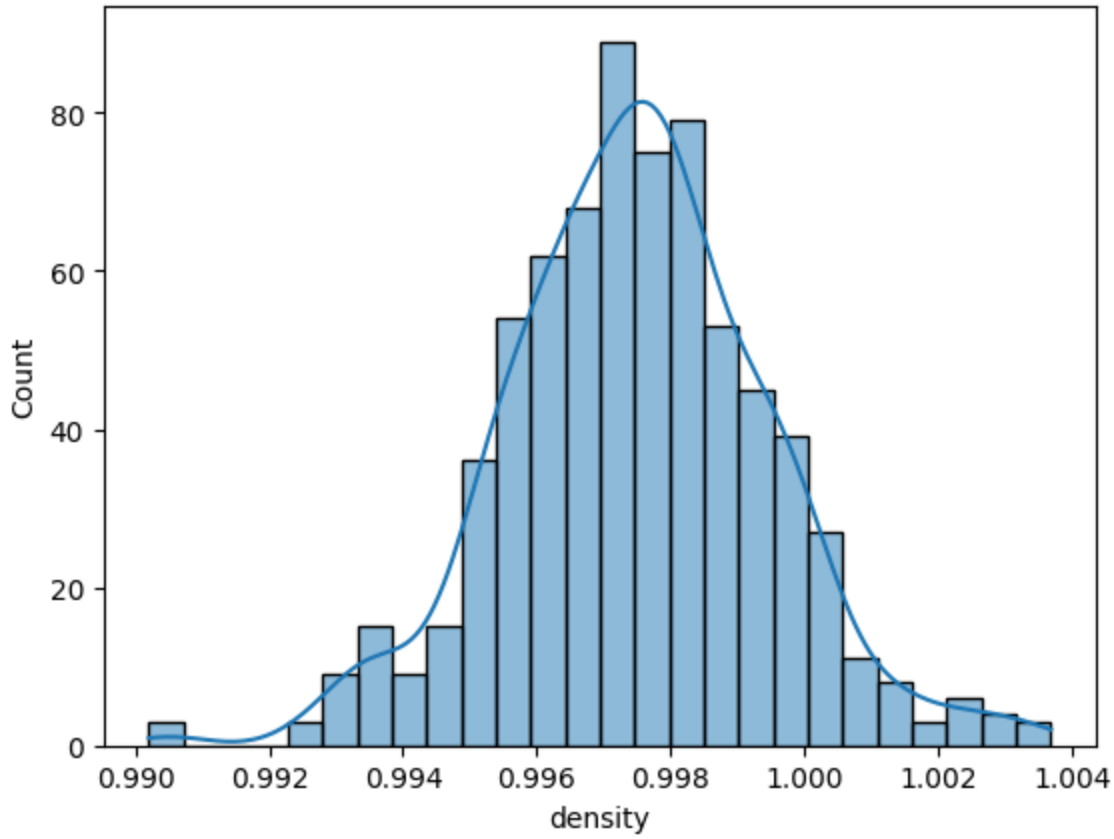
In [52]:
```
sns.histplot(df1['density'], kde= True)
```

Out[52]:
```
<Axes: xlabel='density', ylabel='Count'>
```



In [19]:
```
sns.histplot(df2['density'], kde= True)
```

Out[19]:
```
<Axes: xlabel='density', ylabel='Count'>
```



Utilizing the visualization we do see a difference, we also find that their means are different

c. What test are you going to use?

We will be using the two tailed T test.

What is the p-value?

In [130…
```
print(stats.ttest_ind(df1['density'], df2['density'],equal_var = False))
```

```
Ttest_indResult(statistic=-14.697361284248352, pvalue=1.6548158268854152e-45)
```

P value is 1.6548158268854152e-45

What is your conclusion?

As the p value is quite less than 0.05 we can say that the null hypothesis is rejected, there is a difference in population mean of Group A and Group B

Does your conclusion imply that there is an association between “density” and “residual.sugar”?

Since there is difference between population mean of group A and group B, we can conclude that there is an association between density and residiual sugar.

Q2. Produce summary statistics of “residual.sugar” and use its 1st, 2nd, and 3rd quantiles to divide the data into four groups A, B, C, and D. We want to test if “density” in the four groups has the same population mean. Please answer the following questions.

In [90]:
```
df_1 = pd.read_csv(r"E:\Linder_college\Statistical_Methods\Data_Sets\winequality-red.csv
                   ,sep=';')
```

Summary statistics for residual sugar

In [141…]:
```
df_1['residual sugar'].describe()
```

Out[141]:
```
count    1599.000000
mean        2.538806
std         1.409928
min         0.900000
25%         1.900000
50%         2.200000
75%         2.600000
max        15.500000
Name: residual sugar, dtype: float64
```

In [91]:
```
df_1['residual sugar'].quantile([0.25, 0.5, 0.75])
```

Out[91]:
```
0.25    1.9
0.50    2.2
0.75    2.6
Name: residual sugar, dtype: float64
```

In [92]:
```
q1 = 1.9
q2 = 2.2
q3 = 2.6
```

In [93]:
```
def get_group1(row):
    if row['residual sugar'] <= q1:
        row['Group'] = 'A'
    elif (row['residual sugar'] > q1) & (row['residual sugar'] <= q2) :
        row['Group'] = 'B'
    elif (row['residual sugar'] > q2) & (row['residual sugar'] <= q3) :
        row['Group'] = 'C'
    else:
        row['Group'] = 'D'

    return row['Group']

df_1['Group'] = df_1.apply(get_group1, axis=1)
```

In [94]:
```
df_1
```

Out[94]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Gr
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	
…	…	…	…	…	…	…	…	…	…	…	…	…	
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5	
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6	
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6	
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5	
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6	

1599 rows × 13 columns

In [95]:
```
dfa = df_1[df_1['Group']=='A']
dfb = df_1[df_1['Group']=='B']
dfc = df_1[df_1['Group']=='C']
dfd = df_1[df_1['Group']=='D']
```

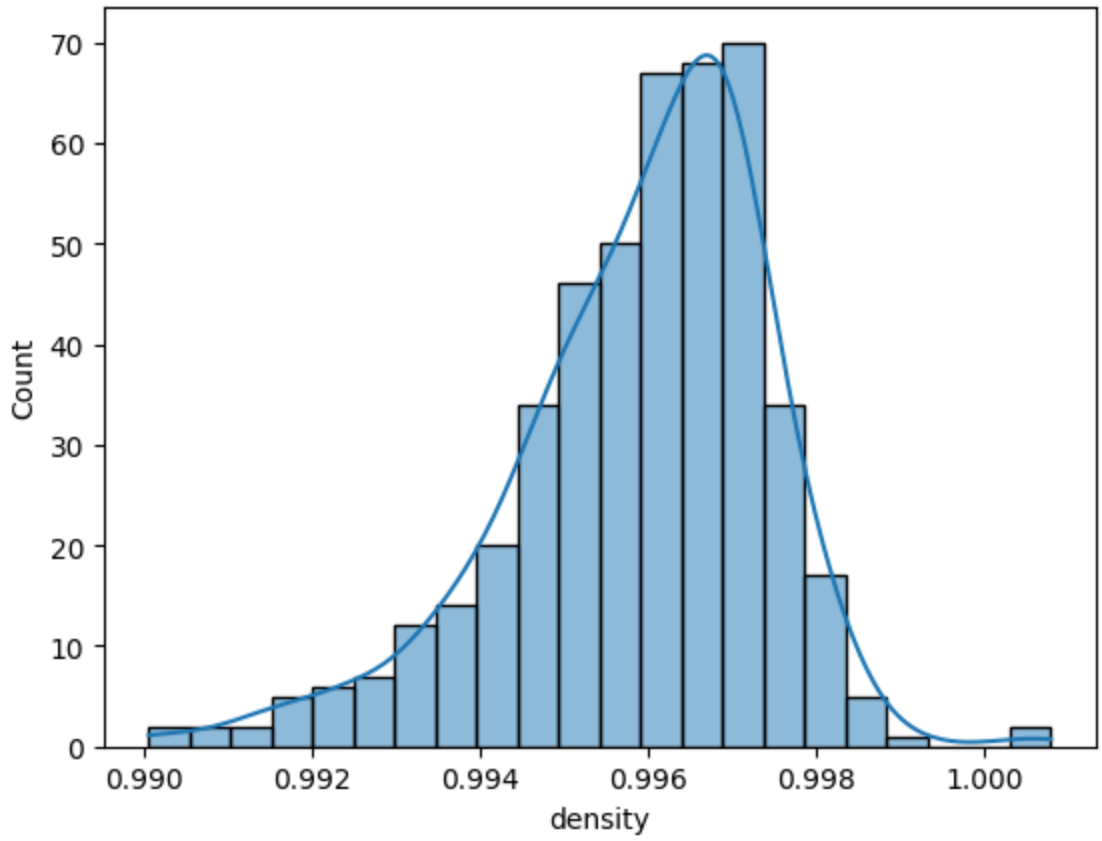
- State the null hypothesis

Our Null hypothesis in this case is that Density for all the four groups will have the same popultion mean

Q2. Use visualization tools to inspect the hypothesis. Do you think the hypothesis is right or not?

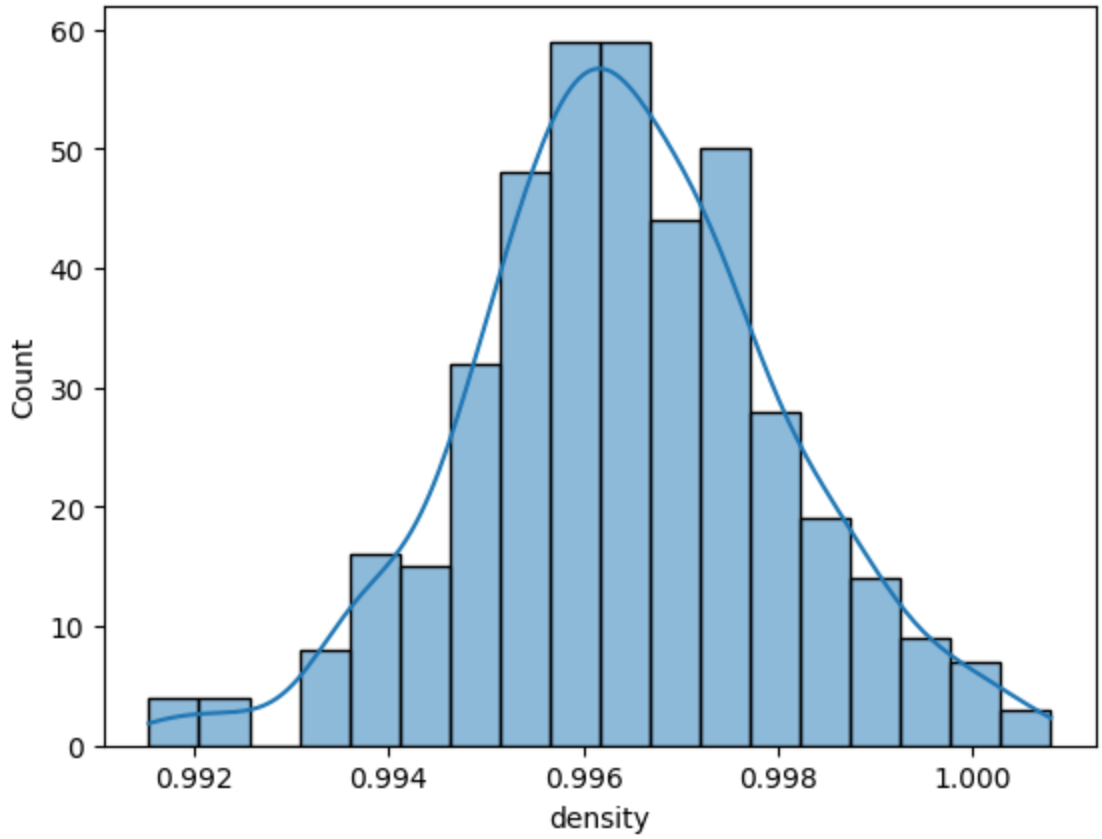
In [97]:
```
sns.histplot(dfa['density'], kde= True)
```

Out[97]:
```
<Axes: xlabel='density', ylabel='Count'>
```



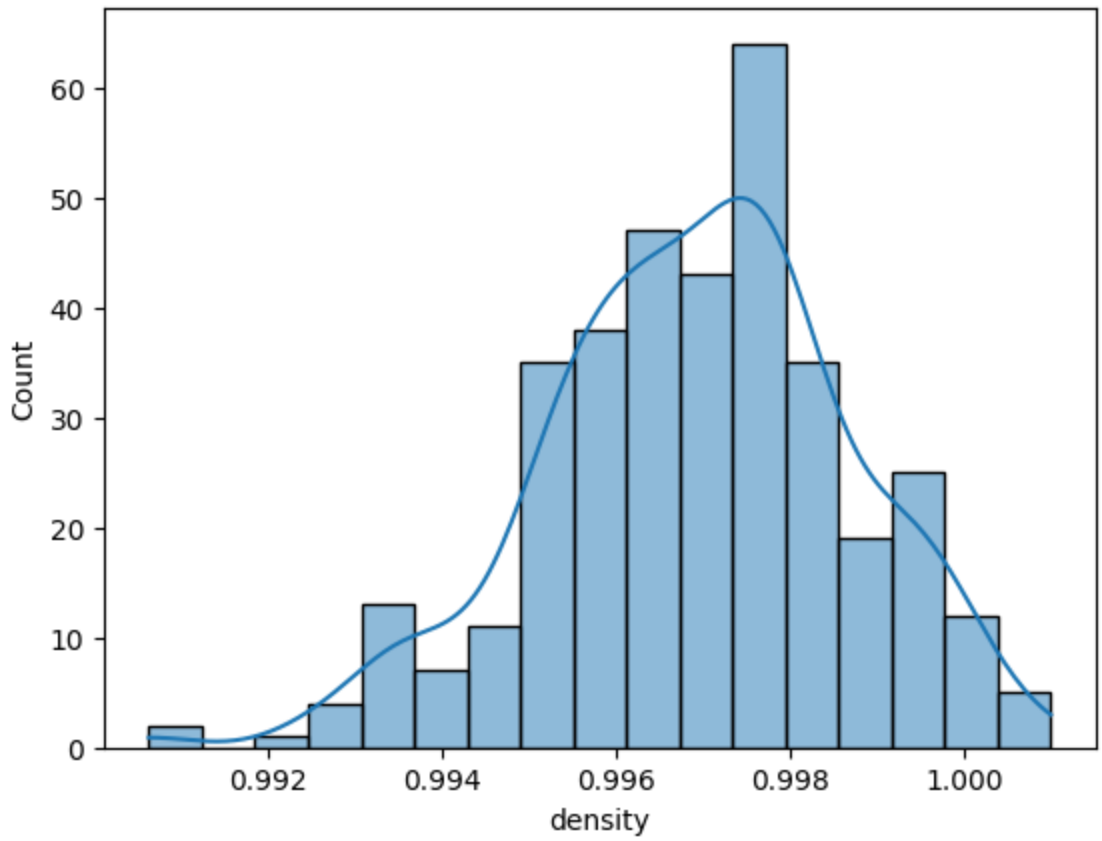
In [98]:
```
sns.histplot(dfb['density'], kde= True)
```

Out[98]:
```
<Axes: xlabel='density', ylabel='Count'>
```



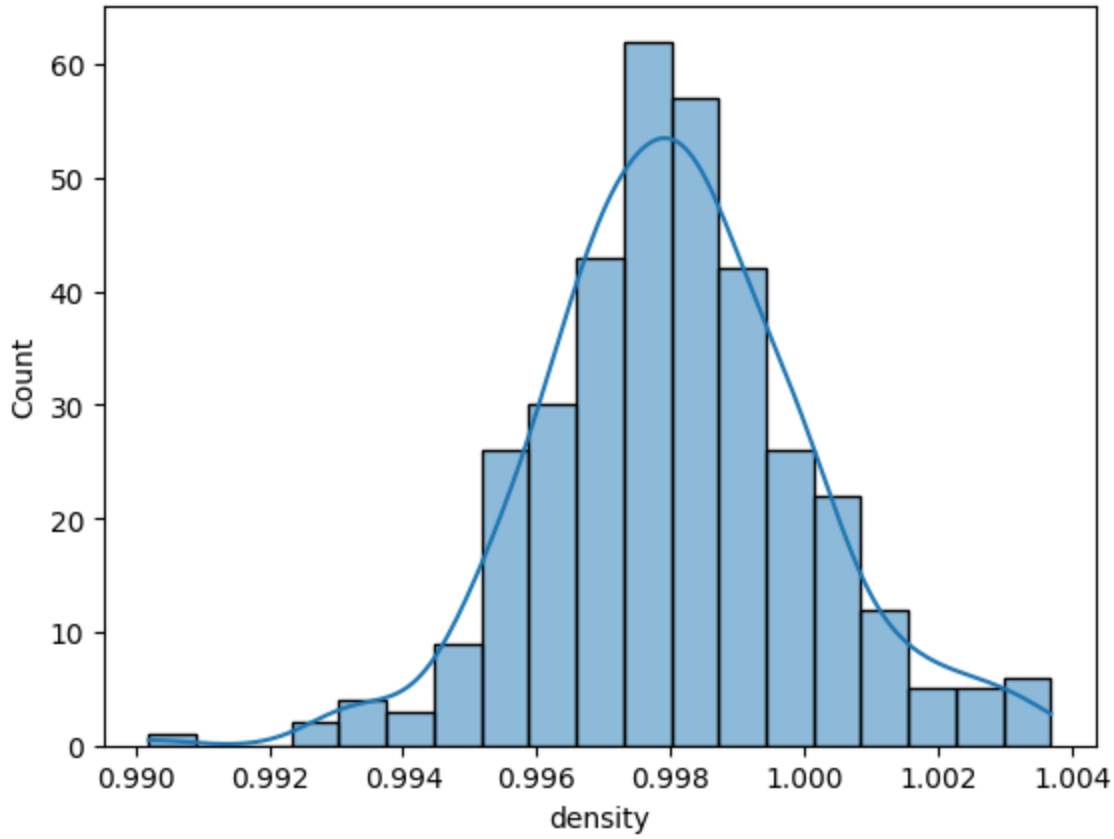
In [99]:
```
sns.histplot(dfc['density'], kde= True)
```

Out[99]:
```
<Axes: xlabel='density', ylabel='Count'>
```



In [100…
```
sns.histplot(dfd['density'], kde= True)
```

Out[100]:
```
<Axes: xlabel='density', ylabel='Count'>
```



We do see variations in the visualization moreover we do find their means to be different.

What test are you going to use?

We will be using the Anova Test

What is the p-value?

In [135…
```
# Conduct the one-way ANOVA
f_oneway(dfa['density'], dfb['density'], dfc['density'], dfd['density'])
```

Out[135]:
```
F_onewayResult(statistic=112.79982975153055, pvalue=3.0656834068470876e-66)
```

The P value is 3.0656834068470876e-66

What is your conclusion?

Since the P value is less than 0.05 we can say that there is difference in population mean of density for the four groups, and we reject the null hypothesis

Does your conclusion imply that there is an association between “density” and “residual.sugar”? Compare your result here with that in Question 1. Do you think increasing the number of groups help identify the association? Would you consider dividing the data into 10 groups so as to help the discovery of the association? Why?

Yes we can conclude that there is an association between density and residual sugar as the p value is less than 0.05, I do feel that dividing the data helps to identify the assosiation. But as we did get the assosiation after dividing the data into two groups there shouldnt be a need to divide the data even further as we have already identified the assosiation.

Q3. Create a 2 by 4 contingency table using the categories A, B, C, D of “residual.sugar” and the binary variable “excellent” you created in Part B. Note that you have two factors: the categorical levels of “residual.sugar” (A, B, C and D) and an indicator of excellent wines (yes or no)

In [102…
```
df_2 = pd.read_csv(r"E:\Linder_college\Statistical_Methods\Data_Sets\winequality-red.csv
                   ,sep=';')
```

In [103…
```
def p(row):
    if row['quality'] >= 7:
        row['p'] = 1
    else:
        row['p'] = 0
    return row['p']

df_2['Excellent'] = df_2.apply(p,axis=1)
```

In [ ]:
```
def get_group1(row):
    if row['residual sugar'] < q1:
        row['Group'] = 'A'
    elif (row['residual sugar'] >= q1) & (row['residual sugar'] < q2) :
        row['Group'] = 'B'
    elif (row['residual sugar'] >= q2) & (row['residual sugar'] < q3) :
        row['Group'] = 'C'
    else:
        row['Group'] = 'D'

    return row['Group']
```

In [106…
```
df_2['Group'] = df_2.apply(get_group1, axis=1)
```

In [107…
```
df_2.head(3)
```

Out[107]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Excelle
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	

In [108…
```
df_3 = df_2[['residual sugar','Excellent','Group']]
```

In [109…
```
df_7 = pd.crosstab(df_3['Group'], df_3['Excellent'])
```

Out[109]:

Excellent	0.0	1.0
Group		
A	411	53
B	367	52
C	308	53
D	296	59

Use the Chi-square test to test if these two factors are correlated or not ?

In [110…
```
stat, p, dof, expected = chi2_contingency(df_7)
```

In [111…
```
# interpret p-value
alpha = 0.05
print("p value is " + str(p))
```

```
p value is 0.13864021156303938
```

In [76]:
```
if p <= alpha:
    print('Dependent (reject Hypothesis)')
else:
    print('Independent (Hypothesis holds true)')
```

```
Independent (Hypothesis holds true)
```

Since p value using the chi square test is greater than 0.05 the hypothesis holds true

- Use the permutation test to do the same and compare the result to that in (a);

In [129…
```
#Using permutation test with 2000 replications

# Define the observed test statistic (chi-squared statistic)
observed_chi2 = chi2_contingency(pd.crosstab(df_3['Group'], df_3['Excellent']))[0]

# Specify the number of permutations
num_permutations = 2000  # Adjust as needed

# Initialize an empty array to store permutation test statistics
permutation_stats = []

# Perform the permutation test
for _ in range(num_permutations):
    # Randomly permute the 'excellent' variable
    permuted_excellent = np.random.permutation(df_3['Excellent'])

    # Create a contingency table and compute the test statistic
    permuted_chi2 = chi2_contingency(pd.crosstab(df_3['Group'], permuted_excellent))[0]

    # Store the test statistic in the permutation_stats array
    permutation_stats.append(permuted_chi2)

# Calculating the p-value
p_value = (np.sum(np.abs(permutation_stats) >= np.abs(observed_chi2)) + 1) / (num_permut

# Output the results
print("Permutation-based p-value:", p_value)
```

```
Permutation-based p-value: 0.14342828585707146
```

The P value using the Chi square test is 0.1386 and the p value using the permutation test is 0.1434

- Can you conclude that “residual sugar” is a significant factor contributing to the excellence of wine? Why?

Since both the test i.e chi square test and permutation test show that there is no corelation between these two variables I would say that residual sugar is not a significant factor contributing to the excellence of wine.