Tugas Besar IF2220 Probabilitas dan Statistika

April 18, 2023

1 Soal 1

Menulis deskripsi statistika (Descriptive Statistics) dari semua kolom pada data yang bersifat numerik

[1]:
```
import pandas as pd
```

[2]:
```
df = pd.read_csv("anggur.csv")
df
```

[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	5.90	0.4451	0.1813	2.049401	0.070574	
1	8.40	0.5768	0.2099	3.109590	0.101681	
2	7.54	0.5918	0.3248	3.673744	0.072416	
3	5.39	0.4201	0.3131	3.371815	0.072755	
4	6.51	0.5675	0.1940	4.404723	0.066379	
..	…	…	…	…	…	
995	7.96	0.6046	0.2662	1.592048	0.057555	
996	8.48	0.4080	0.2227	0.681955	0.051627	
997	6.11	0.4841	0.3720	2.377267	0.042806	
998	7.76	0.3590	0.3208	4.294486	0.098276	
999	5.87	0.5214	0.1883	2.179490	0.052923	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	16.593818	42.27	0.9982	3.27	0.71	
1	22.555519	16.01	0.9960	3.35	0.57	
2	9.316866	35.52	0.9990	3.31	0.64	
3	18.212300	41.97	0.9945	3.34	0.55	
4	9.360591	46.27	0.9925	3.27	0.45	
..	…	…	…	…	…	
995	14.892445	44.61	0.9975	3.35	0.54	
996	23.548965	25.83	0.9972	3.41	0.46	
997	21.624585	48.75	0.9928	3.23	0.55	
998	12.746186	44.53	0.9952	3.30	0.66	
999	16.203864	24.37	0.9983	3.29	0.70	

	alcohol	quality
0	8.64	7

```
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …       …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

[3]:
```
df_result = pd.DataFrame()
counter = 1
for col in df.columns.drop('quality'):
    df_each = pd.DataFrame({
        'Name': col,
        'Mean': df[col].mean(),
        'Median': df[col].median(),
        'Modus': df[col].value_counts().idxmax(),
        'Standar deviasi': df[col].std(),
        'Variansi': df[col].var(),
        'Range': df[col].max() - df[col].min(),
        'Q1': df[col].quantile(0.25),
        'Q2': df[col].quantile(0.5),
        'Q3': df[col].quantile(0.75),
        'Jarak Antar Kuartil': df[col].quantile(0.75) - df[col].quantile(0.25),
        'Skewness': df[col].skew(),
        'Kurtosis': df[col].kurt()
    }, index=[counter])
    counter += 1

    df_result = pd.concat([df_result, df_each])

df_result
```

[3]:

	Name	Mean	Median	Modus	Standar deviasi	\
1	fixed acidity	7.152530	7.150000	6.540000	1.201598	
2	volatile acidity	0.520839	0.524850	0.554600	0.095848	
3	citric acid	0.270517	0.272200	0.301900	0.049098	
4	residual sugar	2.567104	2.519430	2.049401	0.987915	
5	chlorides	0.081195	0.082167	0.070574	0.020111	
6	free sulfur dioxide	14.907679	14.860346	16.593818	4.888100	
7	total sulfur dioxide	40.290150	40.190000	40.610000	9.965767	
8	density	0.995925	0.996000	0.996100	0.002020	
9	pH	3.303610	3.300000	3.340000	0.104875	

```
10             sulphates   0.598390   0.595000   0.590000         0.100819
11               alcohol  10.592280  10.610000  10.310000         1.510706

     Variansi      Range         Q1         Q2         Q3  \
1    1.443837   8.170000   6.377500   7.150000   8.000000
2    0.009187   0.665200   0.456100   0.524850   0.585375
3    0.002411   0.292900   0.237800   0.272200   0.302325
4    0.975977   5.518200   1.896330   2.519430   3.220873
5    0.000404   0.125635   0.066574   0.082167   0.095312
6   23.893519  27.267847  11.426717  14.860346  18.313098
7   99.316519  66.810000  33.785000  40.190000  47.022500
8    0.000004   0.013800   0.994600   0.996000   0.997200
9    0.010999   0.740000   3.230000   3.300000   3.370000
10   0.010164   0.670000   0.530000   0.595000   0.670000
11   2.282233   8.990000   9.560000  10.610000  11.622500

    Jarak Antar Kuartil  Skewness  Kurtosis
1              1.622500 -0.028879 -0.019292
2              0.129275 -0.197699  0.161853
3              0.064525 -0.045576 -0.104679
4              1.324544  0.132638 -0.042980
5              0.028738 -0.051319 -0.246508
6              6.886381  0.007130 -0.364964
7             13.237500 -0.024060  0.063950
8              0.002600 -0.076883  0.016366
9              0.140000  0.147673  0.080910
10             0.140000  0.149199  0.064819
11             2.062500 -0.018991 -0.131732
```

[ ]:

2 Soal 2

Menampilkan visualisasi plot distribusi dalam bentuk histogram dan boxplot untuk setiap kolom

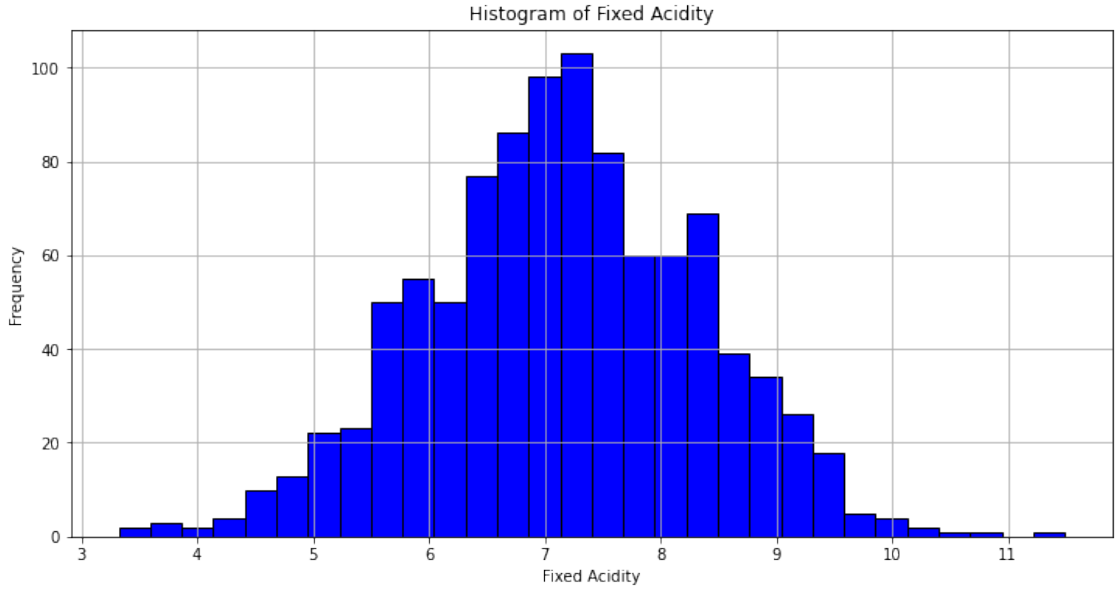
[4]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

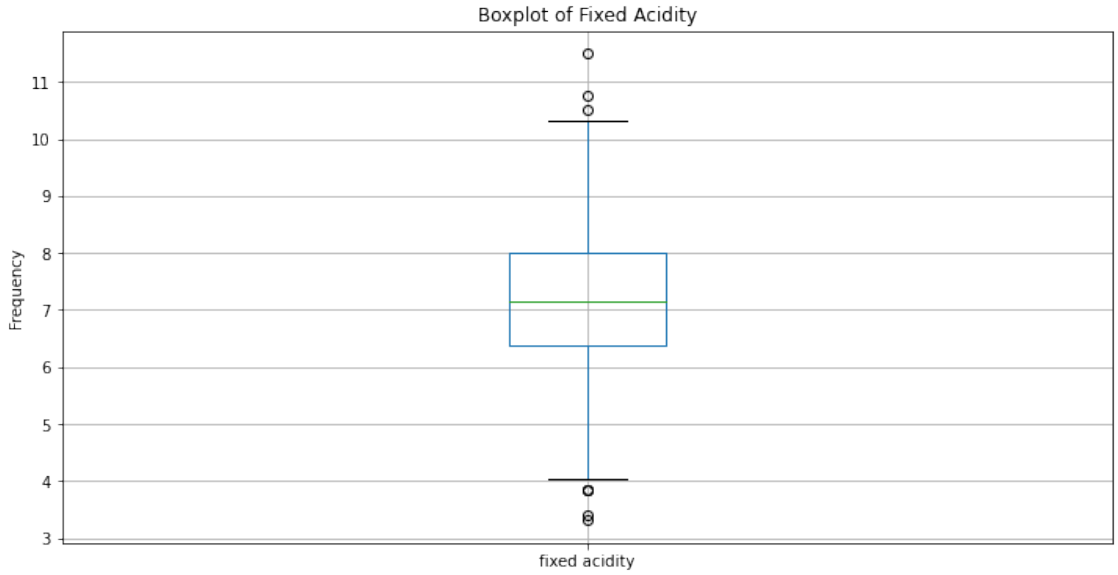
[5]:
```
df = pd.read_csv("anggur.csv")
```

2.1 Fixed Acidity

[6]:
```
df.hist(
    column="fixed acidity",
    bins=30,
    figsize=(12, 6),
    color="blue",
```

```
    edgecolor="black",
)
plt.title("Histogram of Fixed Acidity")
plt.xlabel("Fixed Acidity")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="fixed acidity",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Fixed Acidity")
plt.ylabel("Frequency")
plt.show()
```



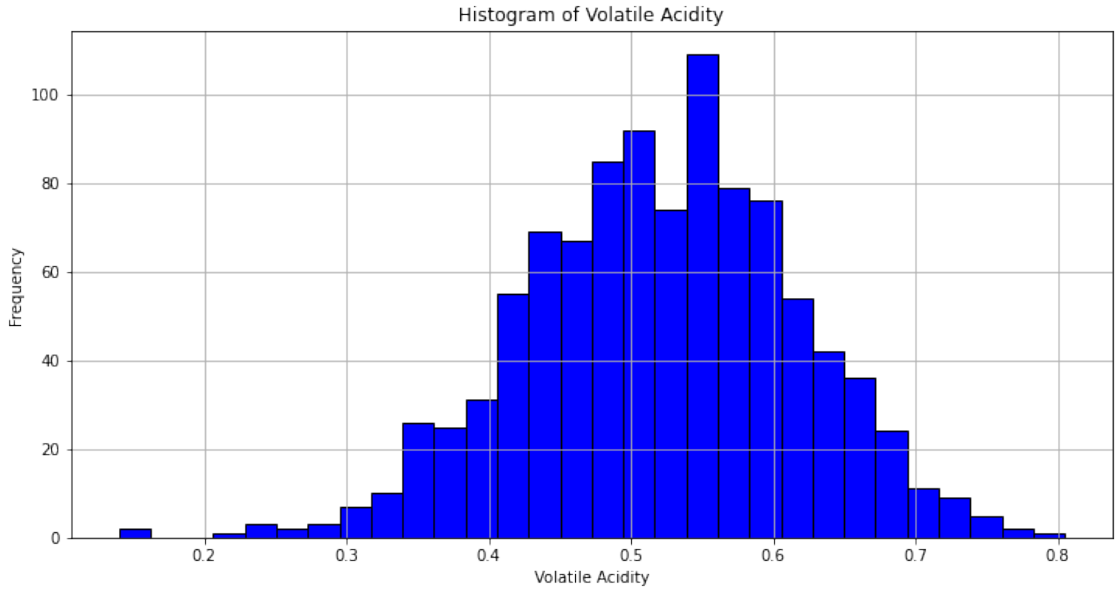


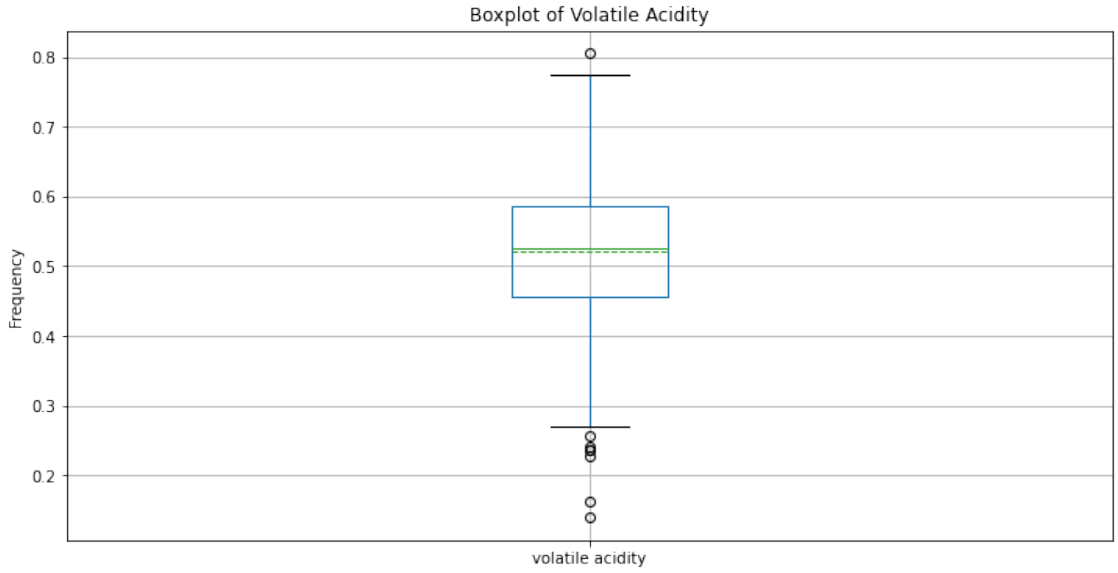
Berdasarkan histogram tersebut terlihat bahwa distribusi Fixed Acidity memiliki distribusi secara normal serta dapat dilihat dari boxplot terapat beberapa outlier dengan di rentang 3.5 - 3.9 dan 10 - 11.5, terlihat median berada di sekitar 7 dengan quartil pertama di sekitar 6

2.2 Volatile Acidity

[7]:
```
df.hist(
    column="volatile acidity",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Volatile Acidity")
plt.xlabel("Volatile Acidity")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="volatile acidity",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Volatile Acidity")
plt.ylabel("Frequency")
```

```
plt.show()
```

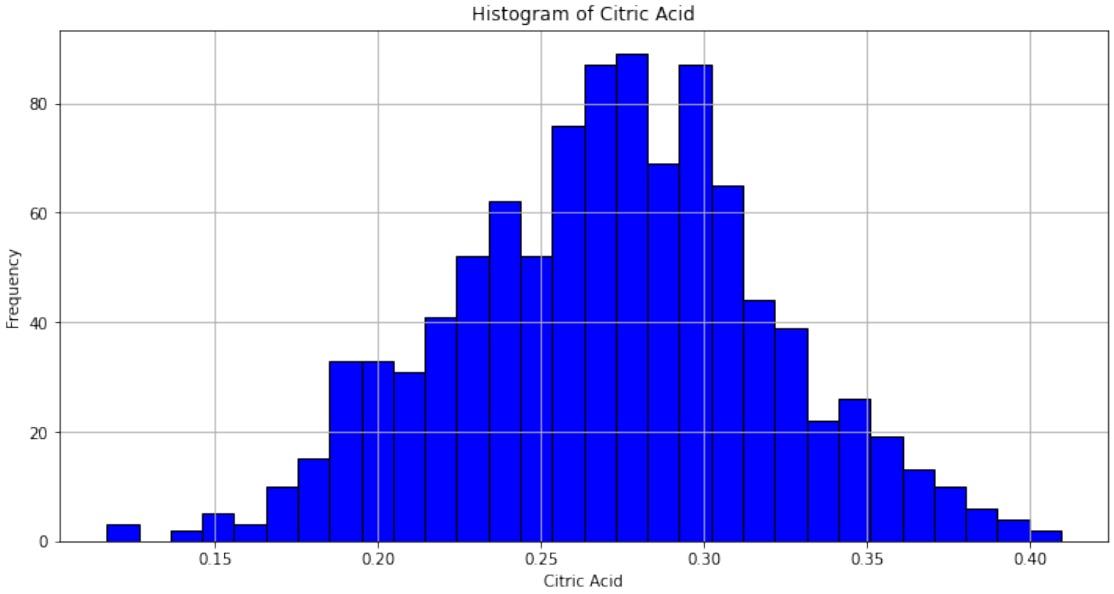


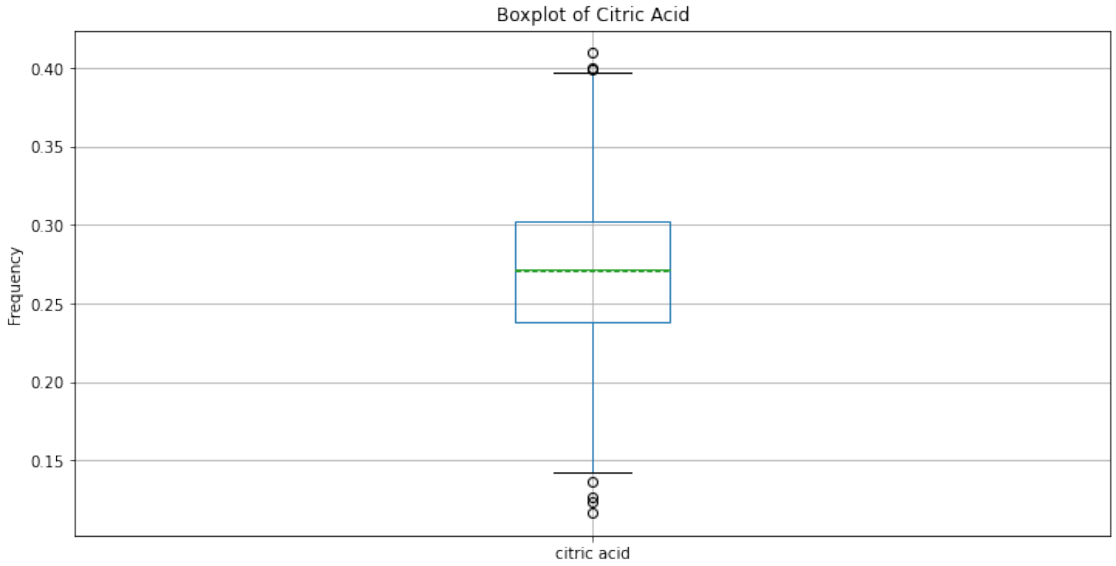


Berdasarkan histogram tersebut terlihat bahwa distribusi Volatile Acidity memiliki kecenderungan ke arah kiri (Negatively skewed) serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 0.19 - 0.28 dan 0.81 - 0.83, terlihat median berada di sekitar 0.52 dengan quartil peratama di sekitar 0.45

2.3 Citric Acid

[8]:
```
df.hist(
    column="citric acid",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Citric Acid")
plt.xlabel("Citric Acid")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="citric acid",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Citric Acid")
plt.ylabel("Frequency")
plt.show()
```

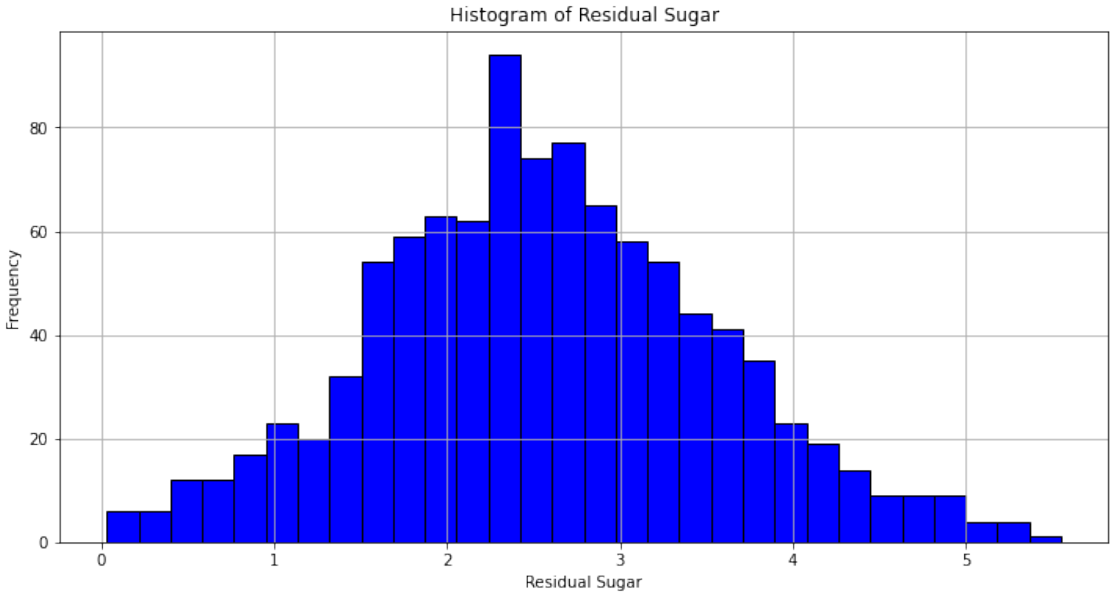


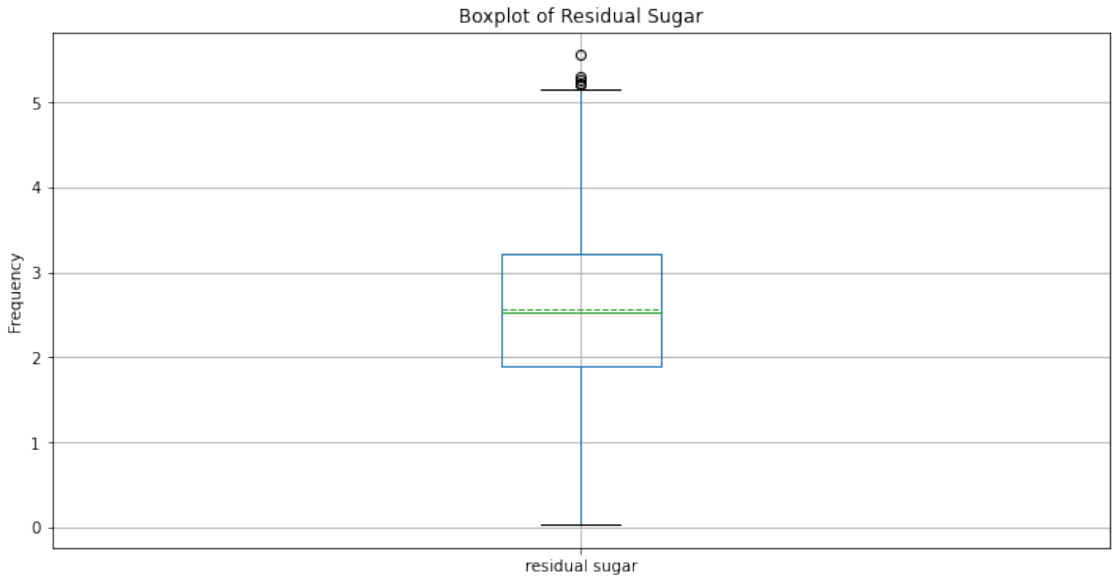


Berdasarkan histogram tersebut terlihat bahwa distribusi Citric Acid memiliki distribusi secara normal serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 0.0 - 0.14 dan 0.4 - 0.45, terlihat median berada di sekitar 0.52 dengan quartil peratama di sekitar 0.2

2.4 Residual Sugar

[9]:
```
df.hist(
    column="residual sugar",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Residual Sugar")
plt.xlabel("Residual Sugar")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="residual sugar",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Residual Sugar")
plt.ylabel("Frequency")
plt.show()
```



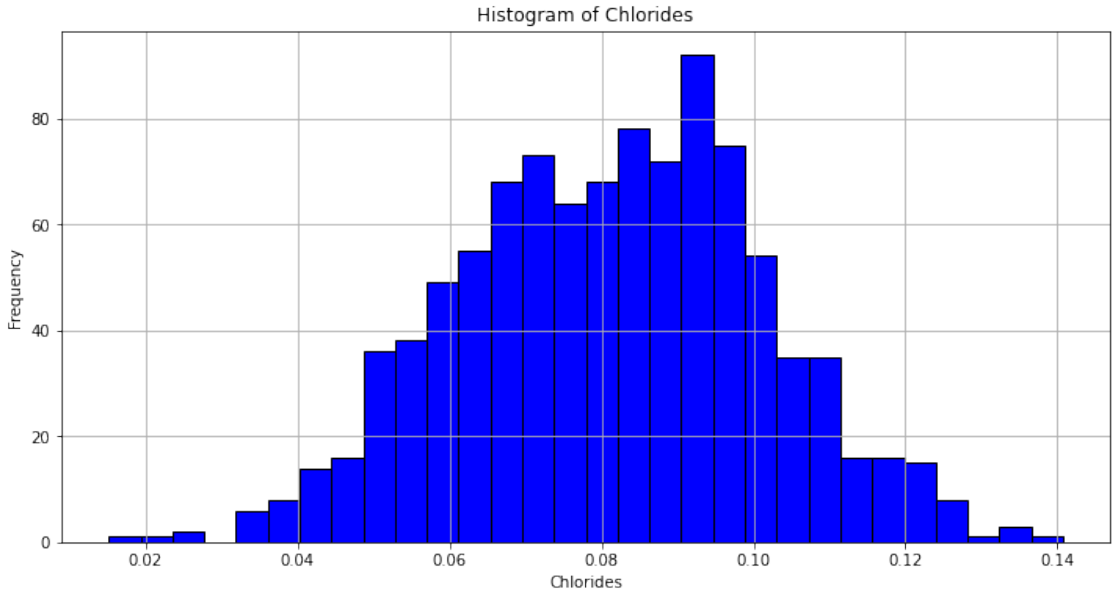


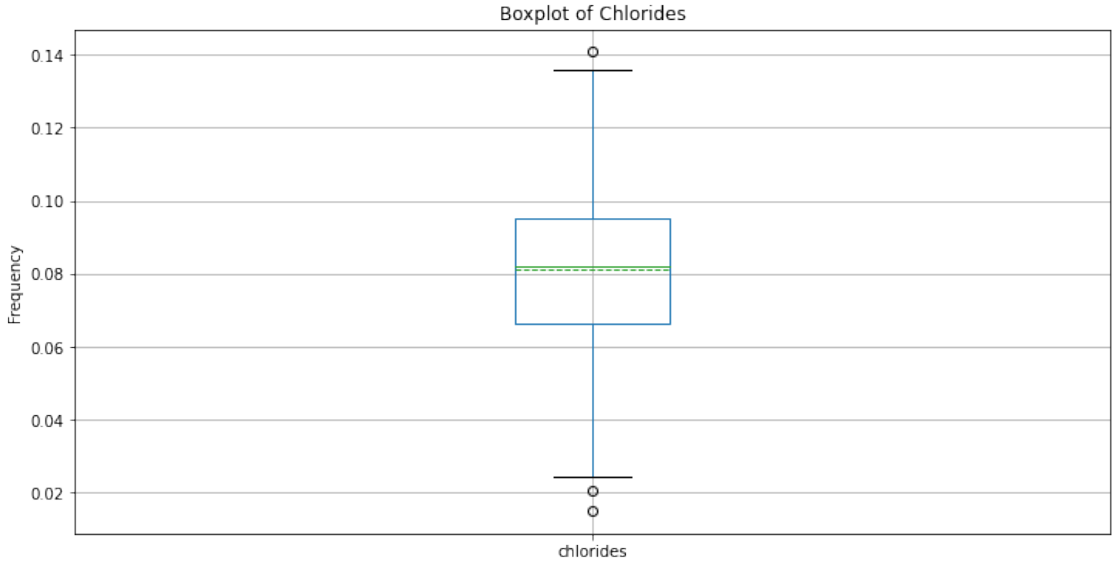
Berdasarkan histogram tersebut terlihat bahwa distribusi Residual Sugar memiliki distribusi secara normal serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 5.1 - 5.8, terlihat median berada di sekitar 2.5 dengan quartil peratama di sekitar 1.89

2.5 Chlorides

[10]:
```
df.hist(
    column="chlorides",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Chlorides")
plt.xlabel("Chlorides")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="chlorides",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Chlorides")
plt.ylabel("Frequency")
```

```
plt.show()
```

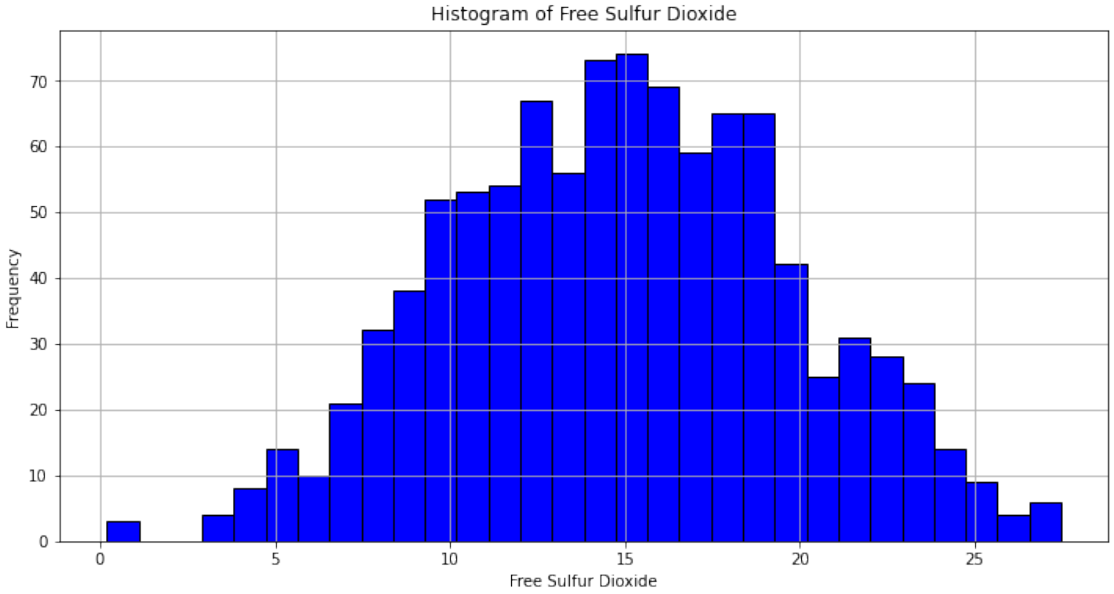


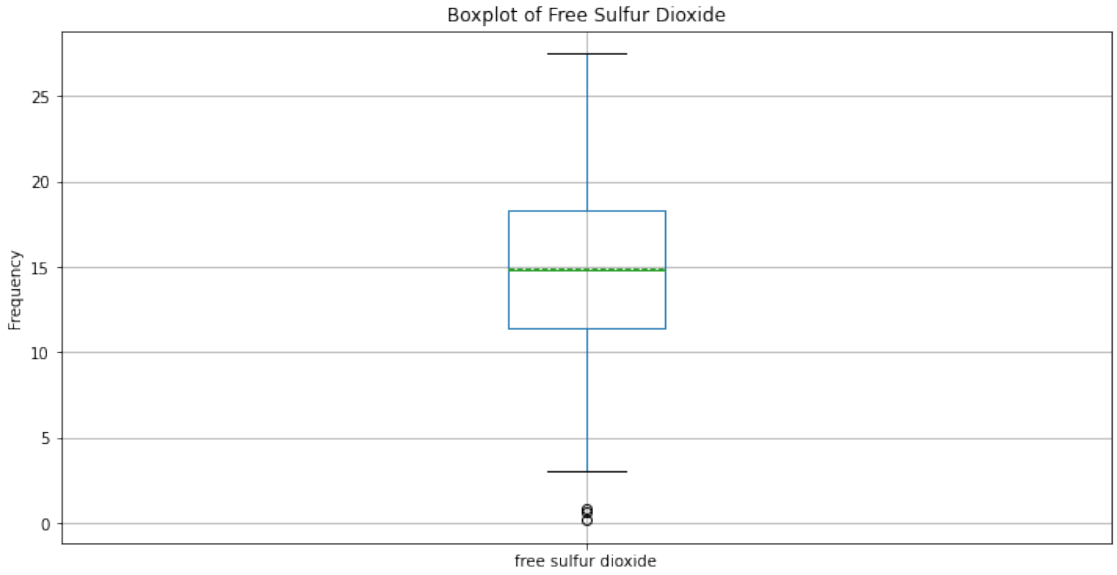


Berdasarkan histogram tersebut terlihat bahwa distribusi Chlorides memiliki kecenderungan ke arah kiri (Negatively skewed) serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 0.01 - 0.22 dan 0.14 - 0.15, terlihat median berada di sekitar 0.082 dengan quartil peratama di sekitar 0.06

2.6 Free Sulfur Dioxide

[11]:
```
df.hist(
    column="free sulfur dioxide",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Free Sulfur Dioxide")
plt.xlabel("Free Sulfur Dioxide")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="free sulfur dioxide",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Free Sulfur Dioxide")
plt.ylabel("Frequency")
plt.show()
```

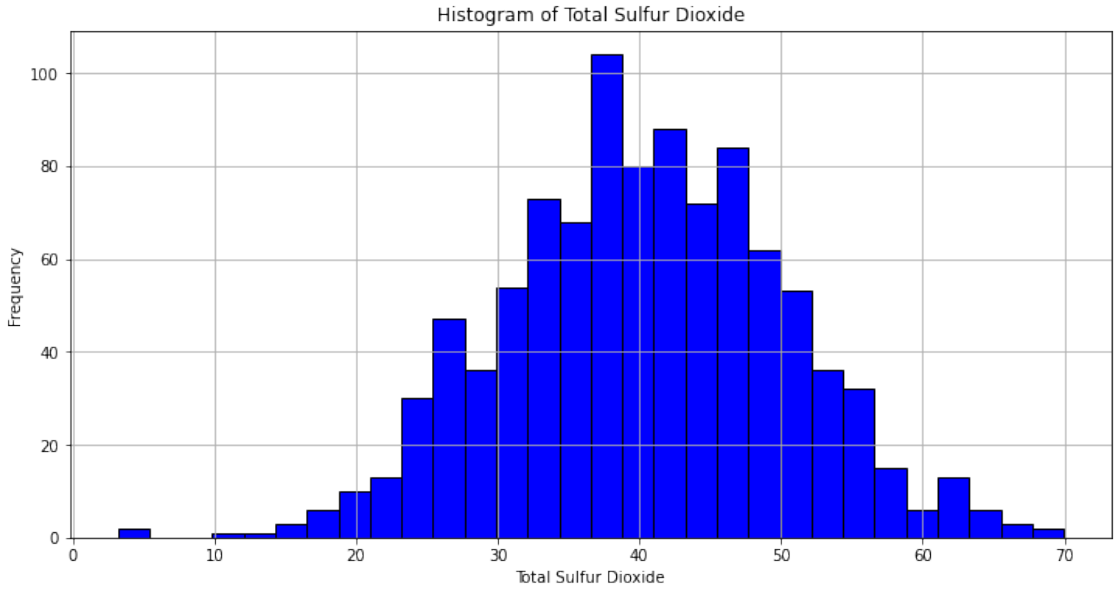


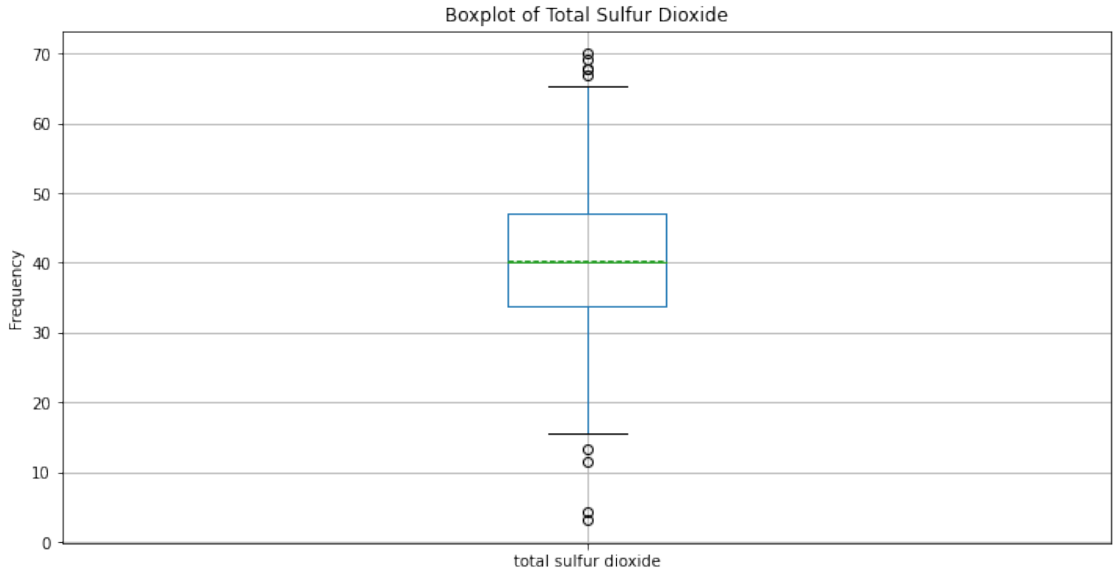


Berdasarkan histogram tersebut terlihat bahwa distribusi Free Sulfur Dioxide memiliki kecenderungan ke arah kiri (Negatively skewed) serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 0.0 - 2, terlihat median berada di sekitar 15 dengan quartil peratama di sekitar 11.4

2.7 Total Sulfur Dioxide

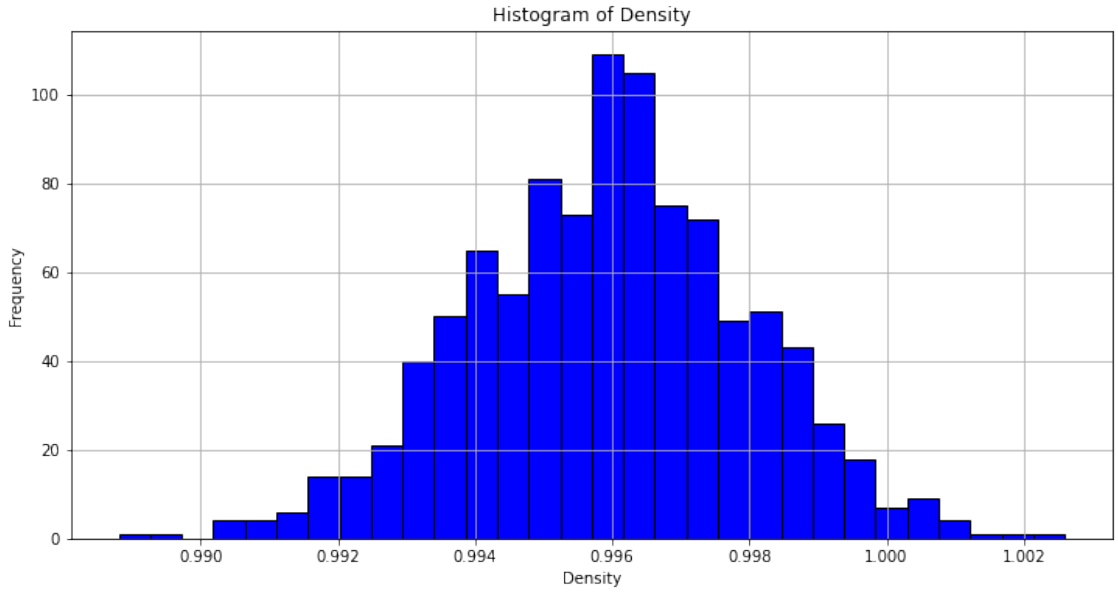
[12]:
```
df.hist(
    column="total sulfur dioxide",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Total Sulfur Dioxide")
plt.xlabel("Total Sulfur Dioxide")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="total sulfur dioxide",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Total Sulfur Dioxide")
plt.ylabel("Frequency")
plt.show()
```

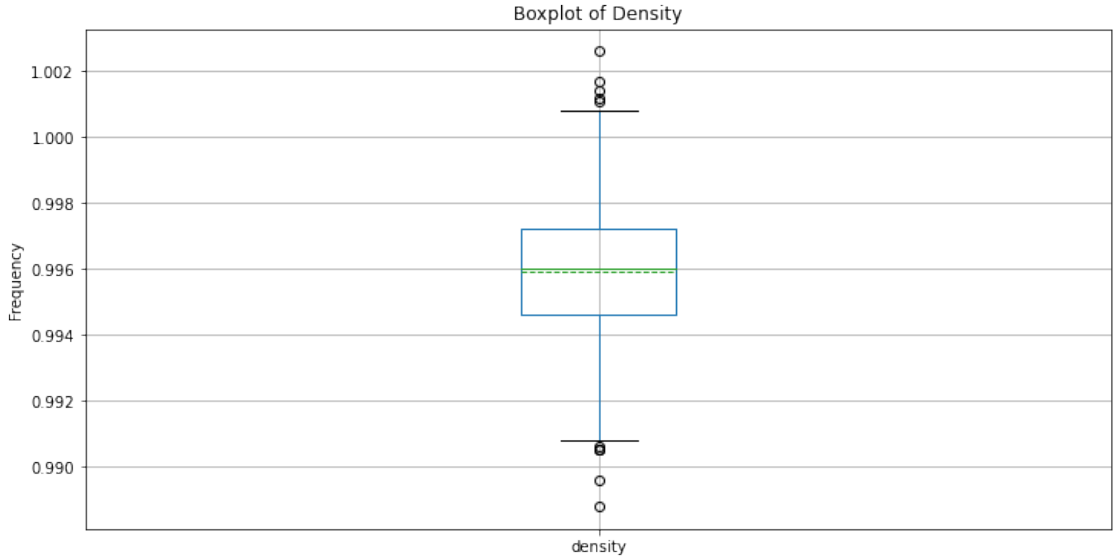




Berdasarkan histogram tersebut terlihat bahwa distribusi Total Sulfur Dioxide memiliki kecenderungan ke arah kiri (Negatively skewed) serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 1 - 15 dan 65 - 70, terlihat median berada di sekitar 40 dengan quartil peratama di sekitar 33.78

[13]:
```
df.hist(
    column="density",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Density")
plt.xlabel("Density")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="density",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Density")
plt.ylabel("Frequency")
plt.show()
```

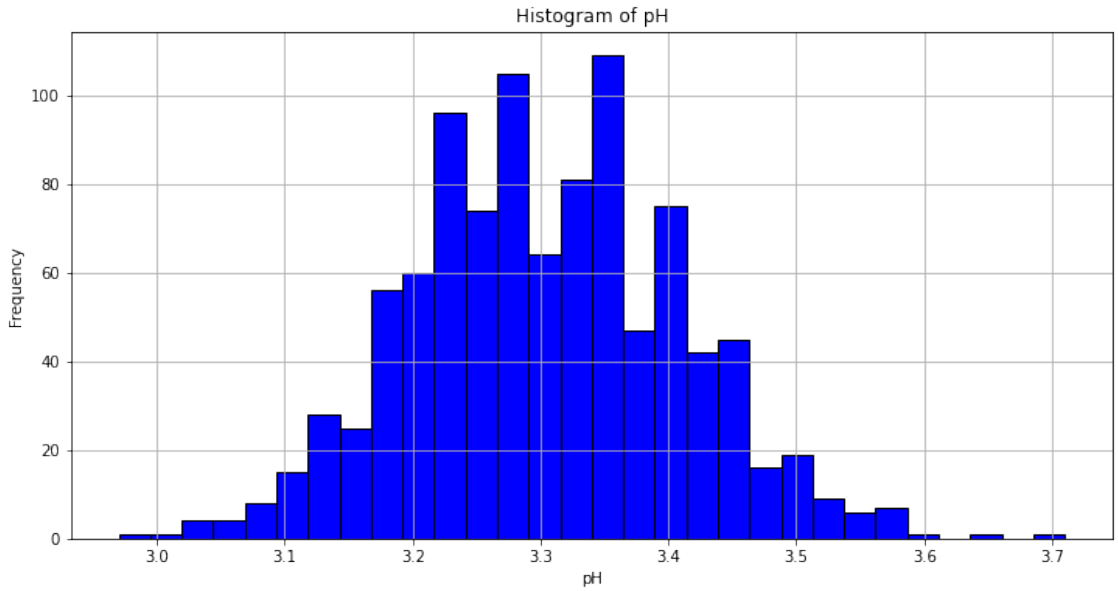


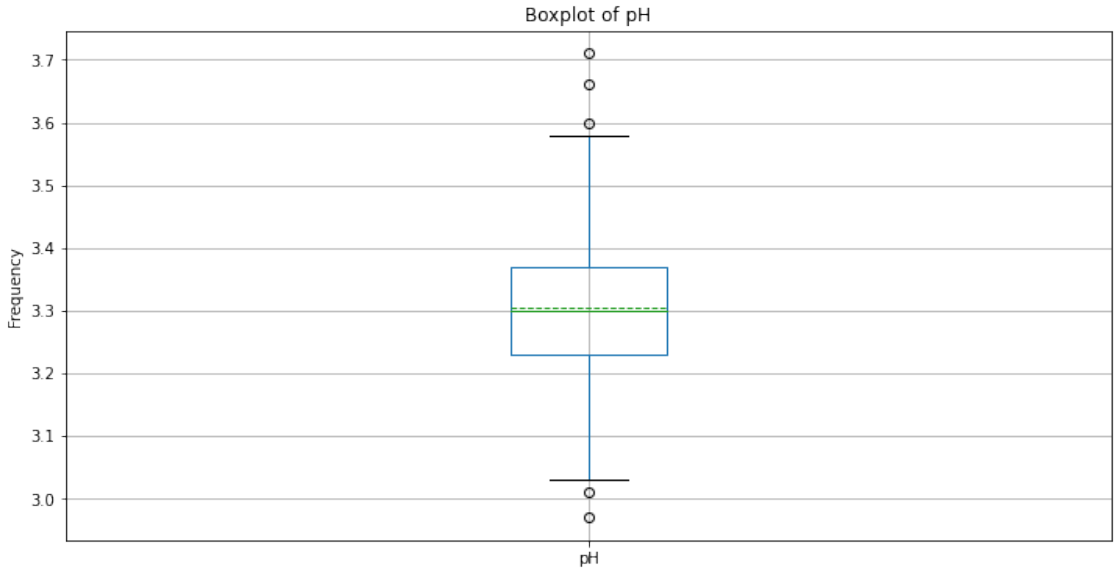


Berdasarkan histogram tersebut terlihat bahwa distribusi Density memiliki distribusi secara normal serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 0.988 - 0.991 dan 1.001 - 1.003, terlihat median berada di sekitar 0.996 dengan quartil peratama di sekitar 0.99

[14]:
```
df.hist(
    column="pH",
    bins=30,
    figsize=(12, 6),
```

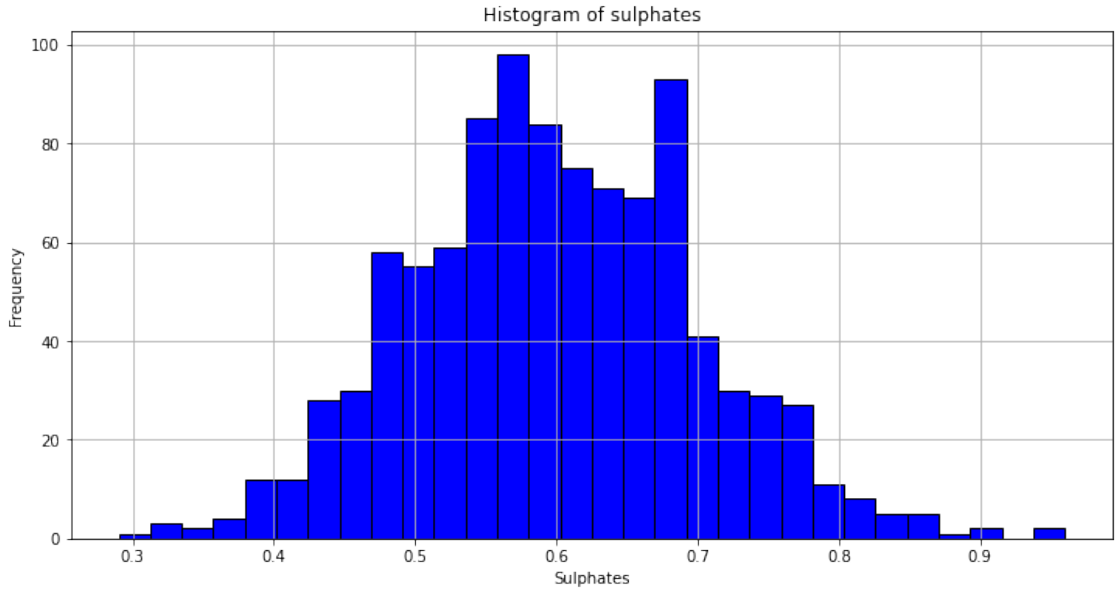
```
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of pH")
plt.xlabel("pH")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="pH",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of pH")
plt.ylabel("Frequency")
plt.show()
```

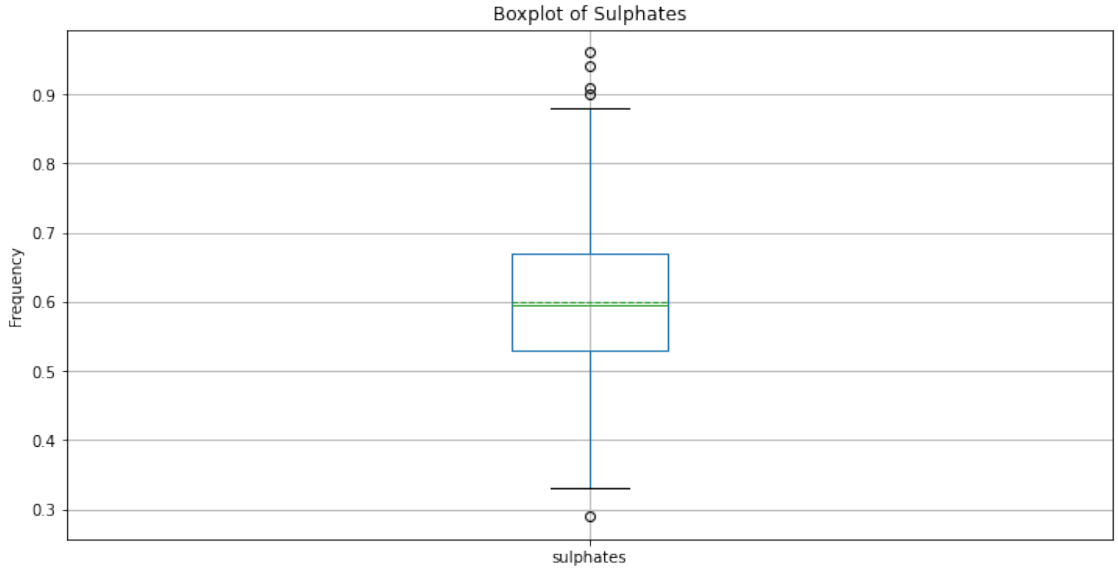




Berdasarkan histogram tersebut terlihat bahwa distribusi pH memiliki kecenderungan ke arah kanan (Positively skewed) serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 2.8 - 3.05 dan 3.6 - 3.71, terlihat median berada di sekitar 3.33 dengan quartil peratama di sekitar 3.23

[15]:
```
df.hist(
    column="sulphates",
    bins=30,
    figsize=(12, 6),
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of sulphates")
plt.xlabel("Sulphates")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="sulphates",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Sulphates")
plt.ylabel("Frequency")
plt.show()
```

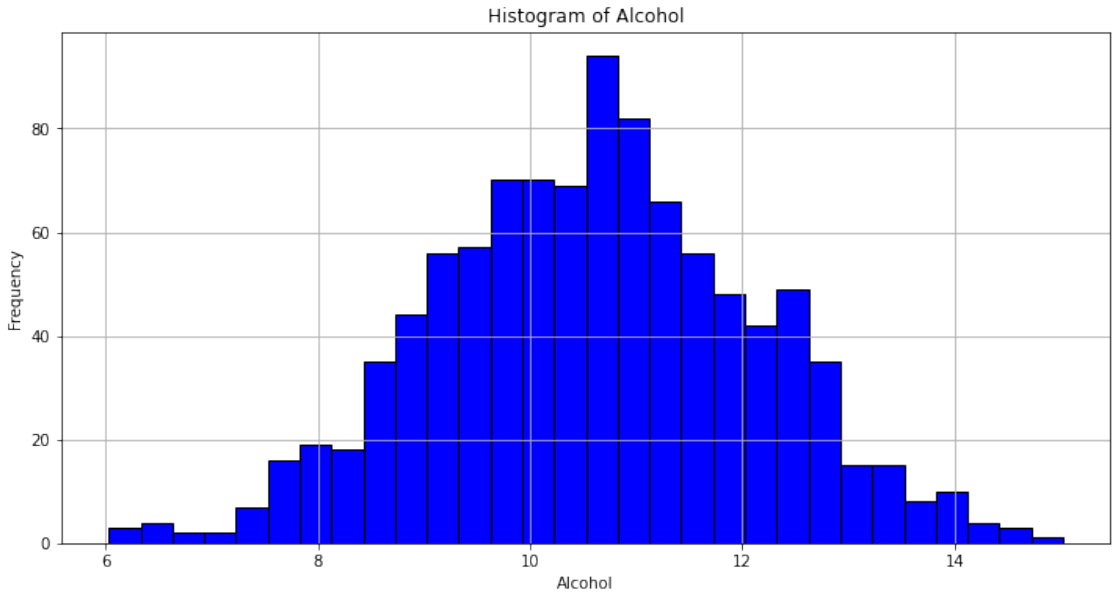


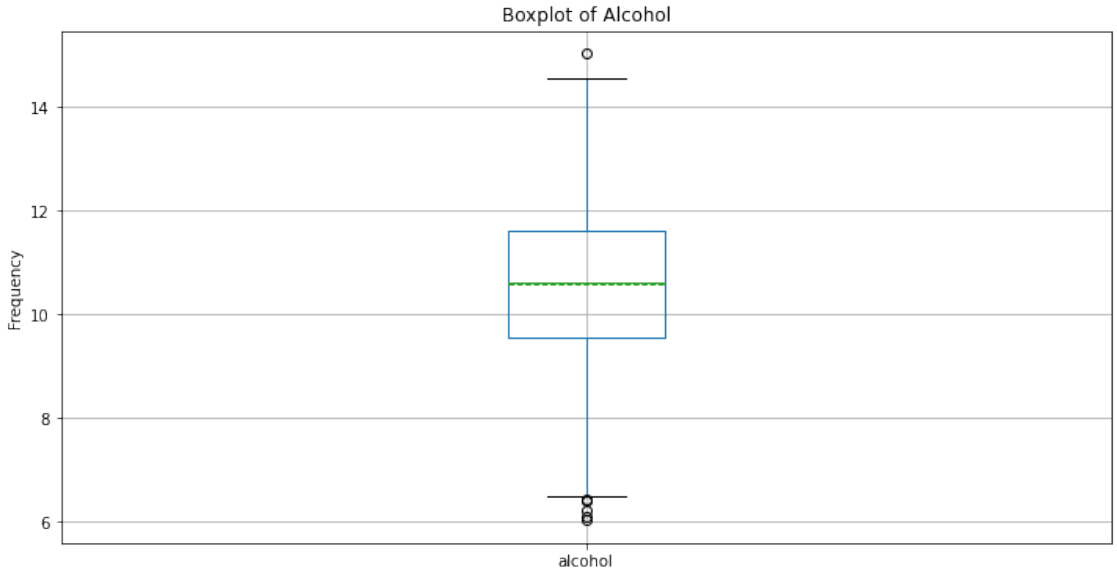


Berdasarkan histogram tersebut terlihat bahwa distribusi Shulpates memiliki kecenderungan ke arah kanan (Positively skewed) serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 0.29 dan 0.89 - 0.98, terlihat median berada di sekitar 0.6 dengan quartil peratama di sekitar 0.53

[16]:
```
df.hist(
    column="alcohol",
    bins=30,
    figsize=(12, 6),
```

```
    color="blue",
    edgecolor="black",
)
plt.title("Histogram of Alcohol")
plt.xlabel("Alcohol")
plt.ylabel("Frequency")
plt.show()

df.boxplot(
    column="alcohol",
    figsize=(12, 6),
    meanline=True,
    showmeans=True,
)
plt.title("Boxplot of Alcohol")
plt.ylabel("Frequency")
plt.show()
```





Berdasarkan histogram tersebut terlihat bahwa distribusi Alcohol memiliki distribusi secara normal serta dapat dilihat dari boxplot terapat beberapa outlier di rentang 6.0 - 6.5 dan 15, terlihat median berada di sekitar 10.5 dengan quartil peratama di sekitar 9.56

3 Soal 3

Menentukan setiap kolom numerik berdistribusi normal atau tidak. Gunakan normality test yang dikaitkan dengan histogram plot.

[17]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import shapiro, normaltest
```

[18]:
```
df = pd.read_csv("anggur.csv")
df
```

[18]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	5.90	0.4451	0.1813	2.049401	0.070574	
1	8.40	0.5768	0.2099	3.109590	0.101681	
2	7.54	0.5918	0.3248	3.673744	0.072416	
3	5.39	0.4201	0.3131	3.371815	0.072755	
4	6.51	0.5675	0.1940	4.404723	0.066379	
..	…	…	…	…	…	
995	7.96	0.6046	0.2662	1.592048	0.057555	
996	8.48	0.4080	0.2227	0.681955	0.051627	

```
997           6.11            0.4841       0.3720        2.377267   0.042806
998           7.76            0.3590       0.3208        4.294486   0.098276
999           5.87            0.5214       0.1883        2.179490   0.052923

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
3              18.212300                 41.97   0.9945  3.34       0.55
4               9.360591                 46.27   0.9925  3.27       0.45
..                   …                     …      …   …       …
995            14.892445                 44.61   0.9975  3.35       0.54
996            23.548965                 25.83   0.9972  3.41       0.46
997            21.624585                 48.75   0.9928  3.23       0.55
998            12.746186                 44.53   0.9952  3.30       0.66
999            16.203864                 24.37   0.9983  3.29       0.70

     alcohol  quality
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …      …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

3.1 Metode Pengujian

Normalitas distribusi dapat dilihat secara sekilas dari histogram yang dihasilkan. Namun pada percobaan ini juga dilakukan beberapa persamaan yang menguji normalitas data sebagai pembanding

3.1.1 Uji Visual

Seperti disebutkan sebelumnya, uji visual dengan melihat histogram dapat dilakukan dengan menyamakan dengan ciri - ciri visual grafik normal, yakni

1. Menyerupai lonceng (bel).
2. Puncak terdapat pada tengah grafik yang merupakan median dan juga merupakan rata - rata.
3. Bentuk grafik simetris pada sebelah kiri dan kanan puncak.
4. Luas daerah sebelah kiri dan kanan puncak adalah 50%.

Ada pula uji visual telah dilakukan pada nomor sebelumnya sehingga tidak akan dilakukan kembali

pada nomor ini

3.1.2 Uji Shapiro-Wilk

Uji Shapiro-Wilk adalah sebuah metode atau rumus perhitungan sebaran data yang digunakan untuk menguji normalitas secara efektif dan valid digunakan untuk sampel berjumlah kecil. Metode Shapiro-wilk menggunakan data dasar yang belum diolah dalam tabel frekuensi, diurut, kemudian dibagi dalam dua kelompok untuk dikonversi dalam Shapiro-Wilk. Setelah itu dilakukan transformasi dalam nilai Z untuk mendapatkan luasan kurva normal.

Persamaan yang digunakan pada uji shapiro-wilk adalah sebagai berikut:

$$T = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

dengan keterangan

x_n adalah angka ke n terkecil pada data

a_n adalah koefisien tes Shapiro-Wilk ke n

Hasil tes diuji dilakukan dengan signifikansi $ = 0.05$, jika nilai p yang didapatkan kurang dari alpha, maka H_0 ditolak dan distribusi tidak normal

Tes shapiro-wilk memiliki ketelitian tinggi untuk data yang berjumlah sedikit (kira - kira $x < 2000$)

3.1.3 Uji D'Agostino-Pearson

Uji lain yang dilakukan adalah tes D'Agostino-Pearson yang merupakan fungsi normaltest dari library scipy. Tes D'Agostino-Pearson dilakukan dengan menggabungkan hasil tes skewness dan kurtosis D'Agostino.

Persamaan yang digunakan pada uji d'agostino-pearson adalah sebagai berikut:

$$K^2 = Z_s^2 + Z_k^2$$

dengan keterangan

Z_s^2 adalah nilai z dari tes skewness d'agostino yang dikuadratkan

Z_k^2 adalah yang dikuadratkan

Hasil tes diuji dilakukan dengan signifikansi $ = 0.05$, jika nilai p yang didapatkan kurang dari alpha, maka H_0 ditolak dan distribusi tidak normal

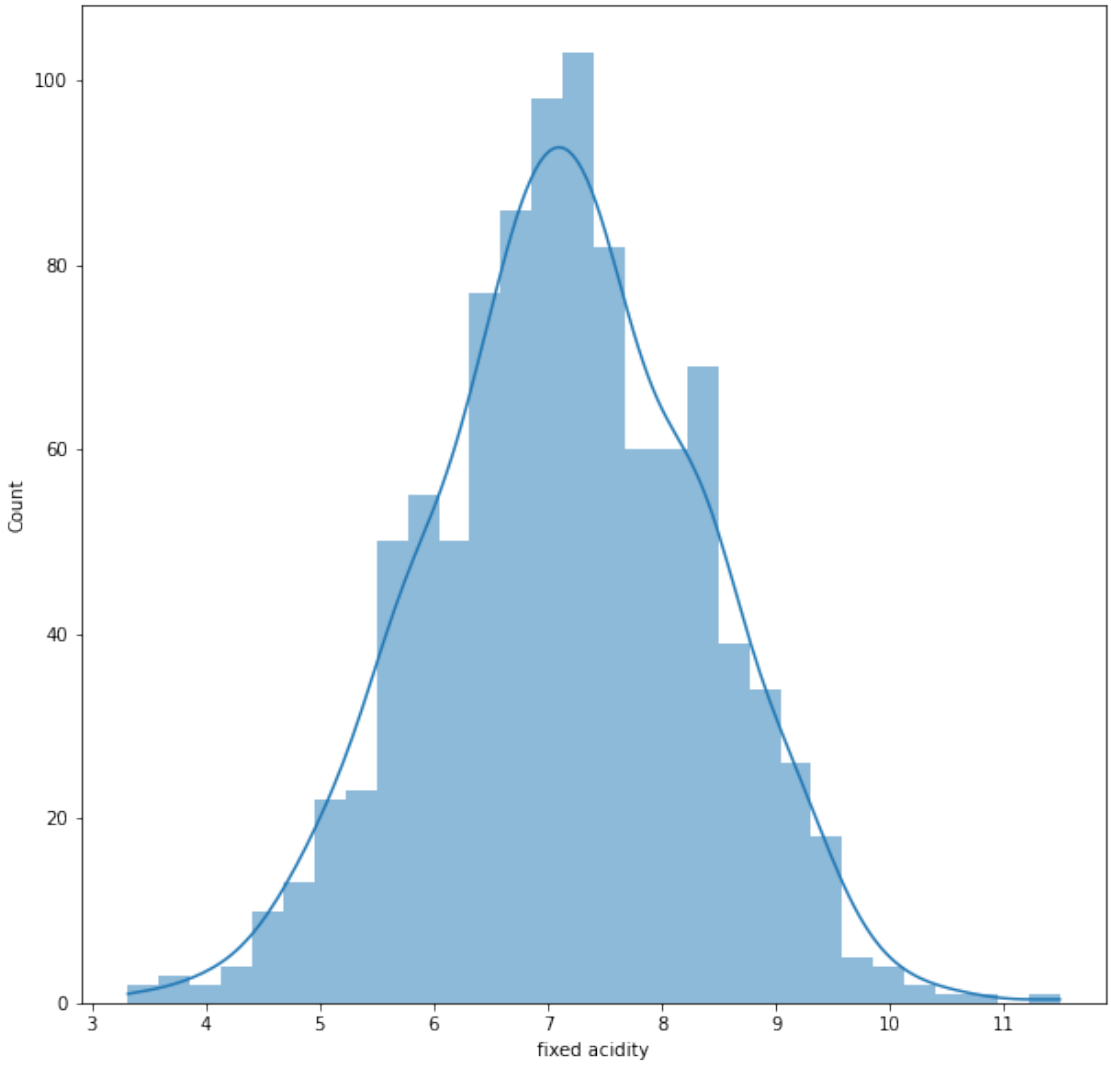
Tes D'Agostino-Pearson memiliki ketelitian yang lebih rendah dibandingkan shapiro-wilk untuk data yang berjumlah sedikit (kira - kira $x < 2000$) namun lebih tinggi untuk data yang lebih banyak

3.1.4 Perbedaan Uji D'Agostino-Pearson dan Uji Shapiro-Wilk

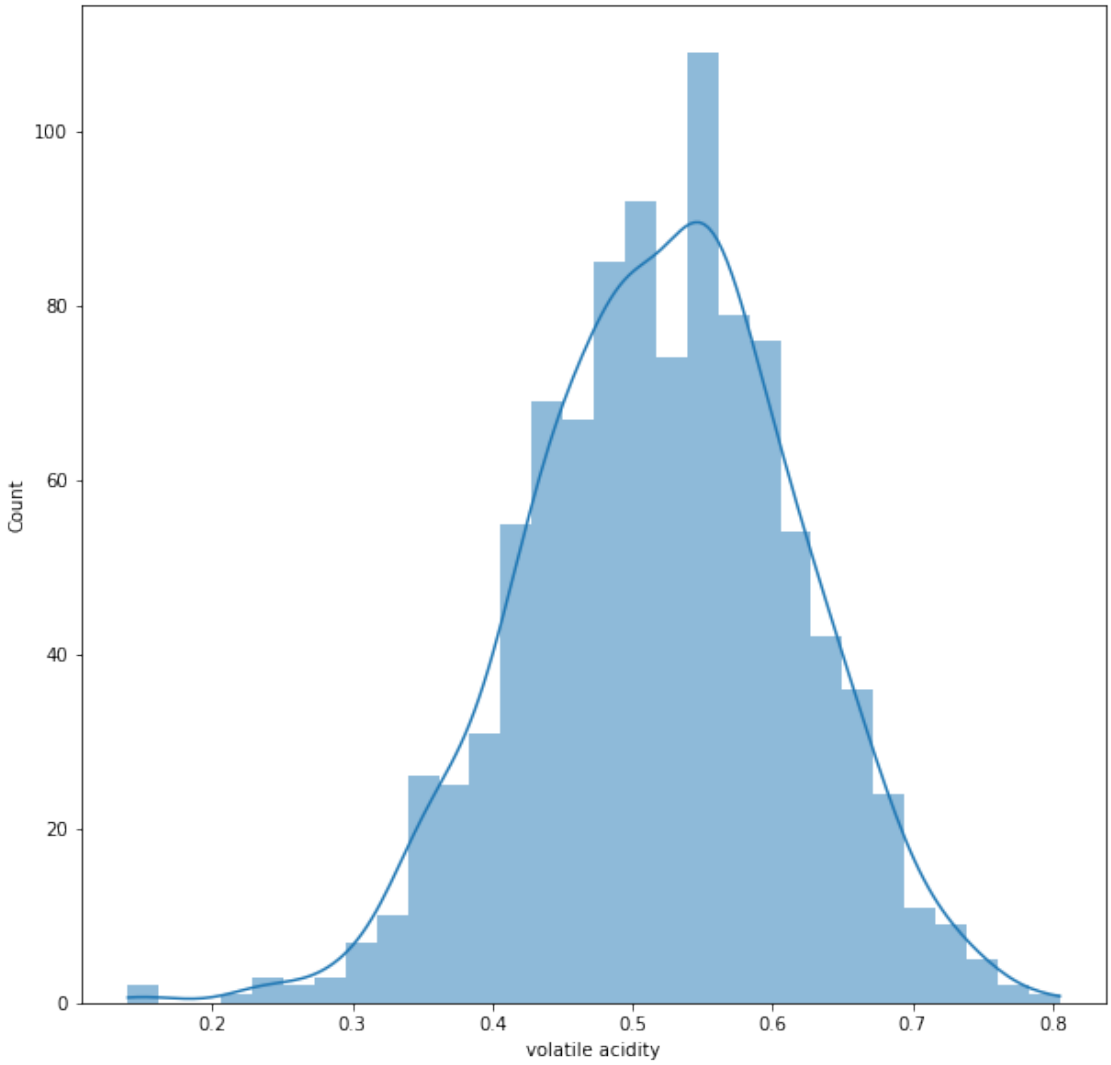
Perbedaan uji dapat terlihat jelas dari metodologi cara pengujian. Namun, pada umumnya - Tes shapiro-wilk memiliki ketelitian tinggi untuk data yang berjumlah sedikit (kira - kira $x < 2000$) - Tes D'Agostino-Pearson memiliki ketelitian yang lebih rendah dibandingkan shapiro-wilk untuk data yang berjumlah sedikit (kira - kira $x < 2000$), tetapi lebih tinggi untuk data yang lebih banyak.

Dalam pengujian memungkinkan grafik untuk ternilai normal untuk satu pengujian dan tidak normal pada pengujian lainnya karena beberapa perbedaan ini.

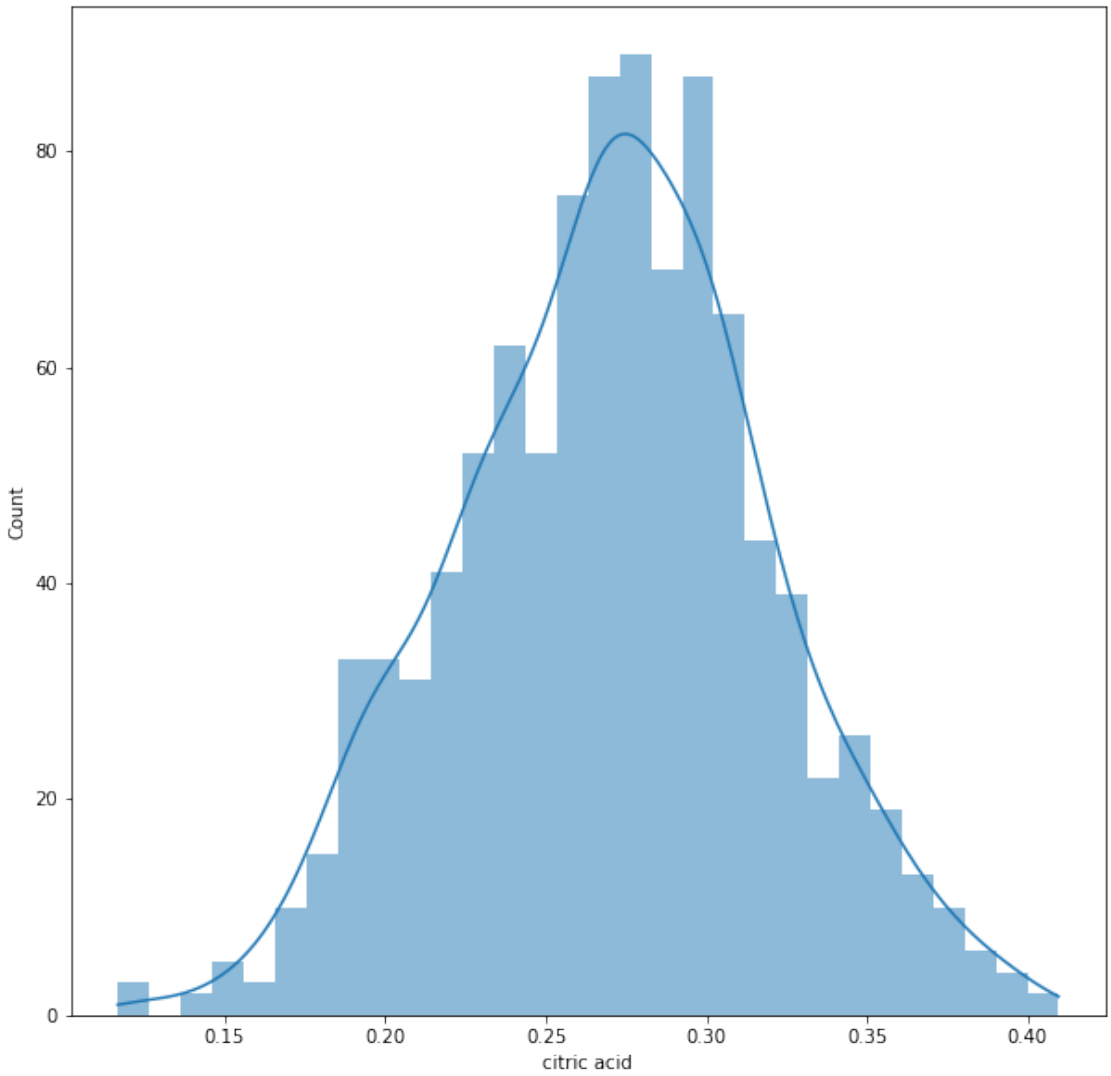
[19]:
```
df_result = pd.DataFrame()
counter = 1
for col in df.columns.drop('quality'):

    plt.figure(figsize=(10, 10))

    sns.histplot(
        df[col],
        bins=30,
        kde=True,
        linewidth = 0,
        alpha=0.5,
    )
    plt.show()

    pvalue1 = shapiro(df[col]).pvalue
    print("p-value shapiro-wilk: " + str(pvalue1))
    if pvalue1 > 0.05:
        print("Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom " +␣
↪col + " memiliki distribusi normal")
    else:
        print("Berdasarkan tes shapiro-wilk, hipotesis nol ditolak, kolom " +␣
↪col + " tidak memiliki distribusi normal")
    pvalue2 = normaltest(df[col]).pvalue
    print("\np-value d'agostino-pearson: " + str(pvalue2))
    if pvalue2 > 0.05:
        print("Berdasarkan tes d'agostino-pearson, hipotesis nol diterima,␣
↪kolom " + col + " memiliki distribusi normal")
    else:
        print("Berdasarkan tes d'agostino-pearson, hipotesis nol ditolak, kolom␣
↪" + col + " tidak memiliki distribusi normal")
```



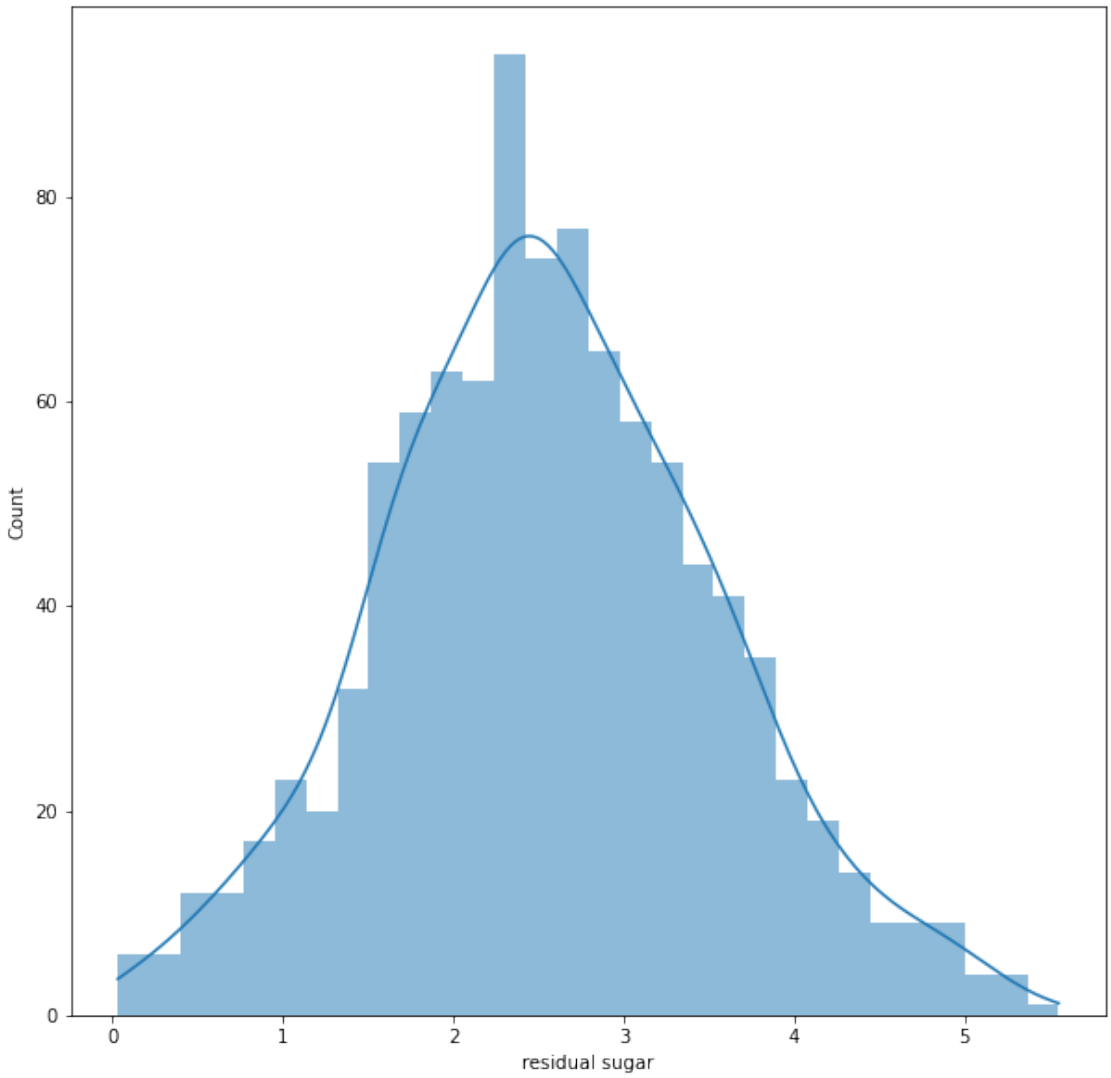

```
p-value shapiro-wilk: 0.8935267925262451
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom fixed acidity
memiliki distribusi normal

p-value d'agostino-pearson: 0.9308584274486692
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom fixed acidity
memiliki distribusi normal
```



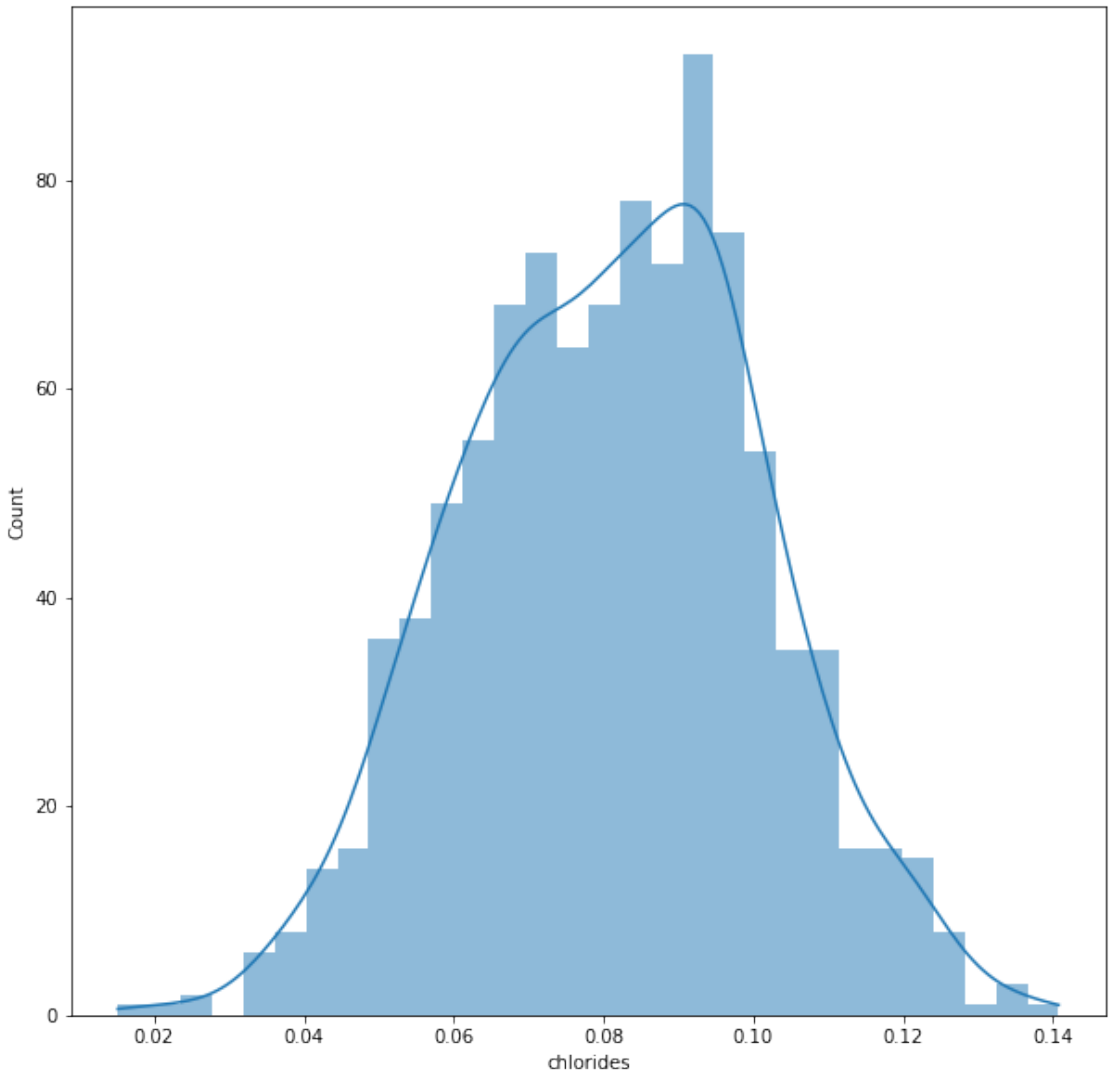

```
p-value shapiro-wilk: 0.05993043631315231
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom volatile acidity
memiliki distribusi normal

p-value d'agostino-pearson: 0.022581461594113835
Berdasarkan tes d'agostino-pearson, hipotesis nol ditolak, kolom volatile
acidity tidak memiliki distribusi normal
```



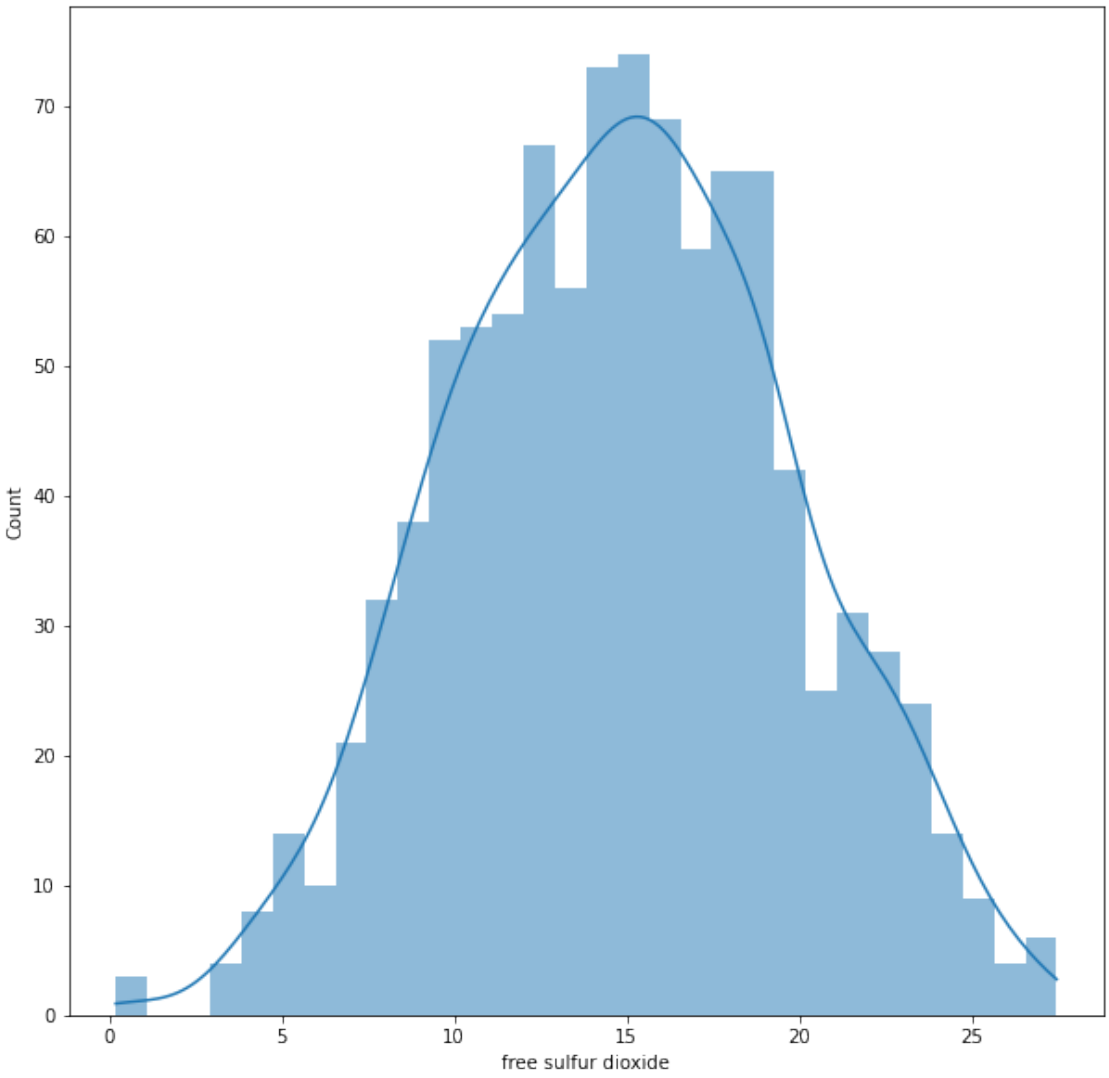

```
p-value shapiro-wilk: 0.26522907614707947
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom citric acid memiliki
distribusi normal

p-value d'agostino-pearson: 0.6816899375976969
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom citric acid
memiliki distribusi normal
```



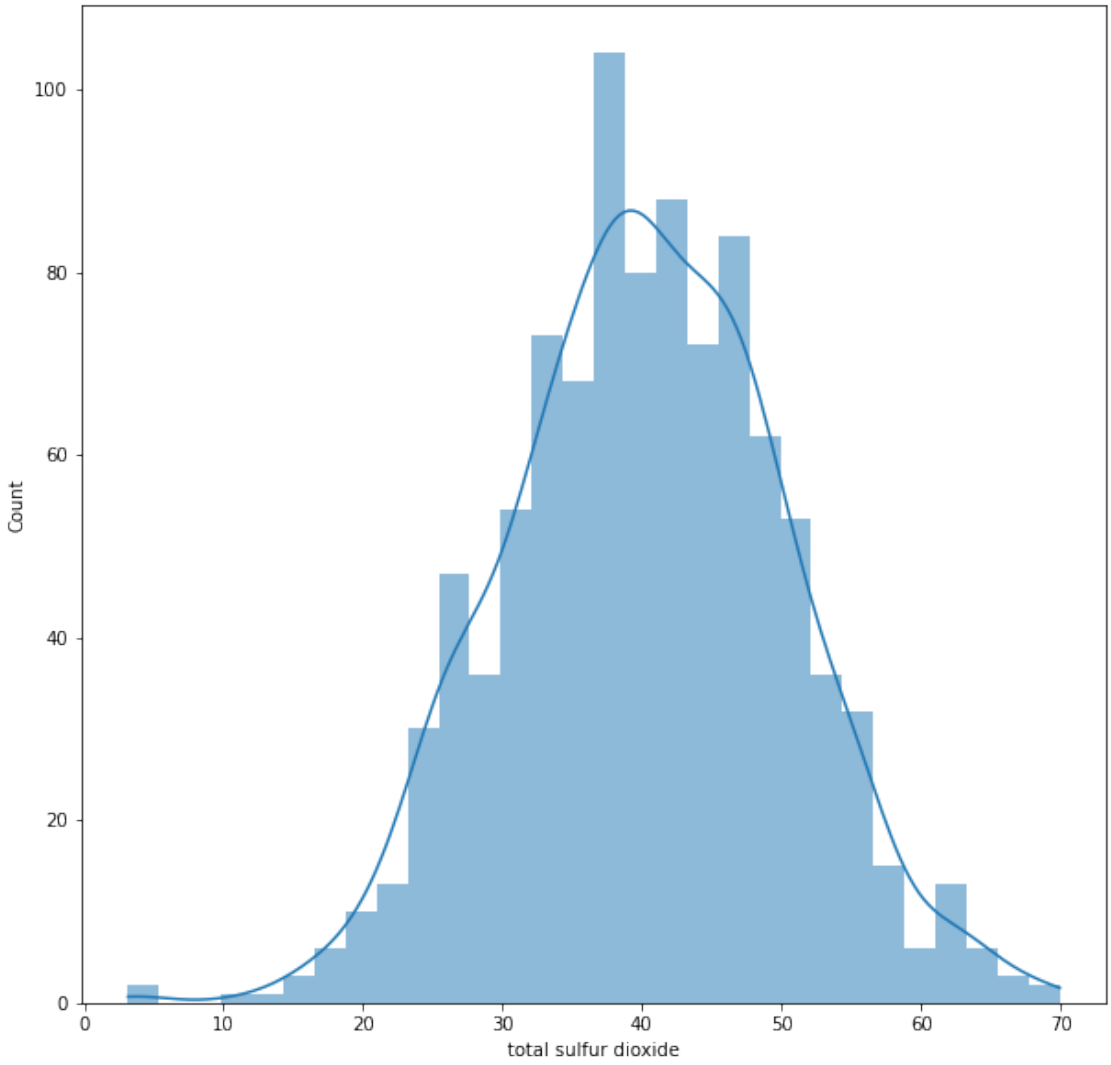

```
p-value shapiro-wilk: 0.044918645173311234
Berdasarkan tes shapiro-wilk, hipotesis nol ditolak, kolom residual sugar tidak
memiliki distribusi normal

p-value d'agostino-pearson: 0.22466703321310558
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom residual sugar
memiliki distribusi normal
```



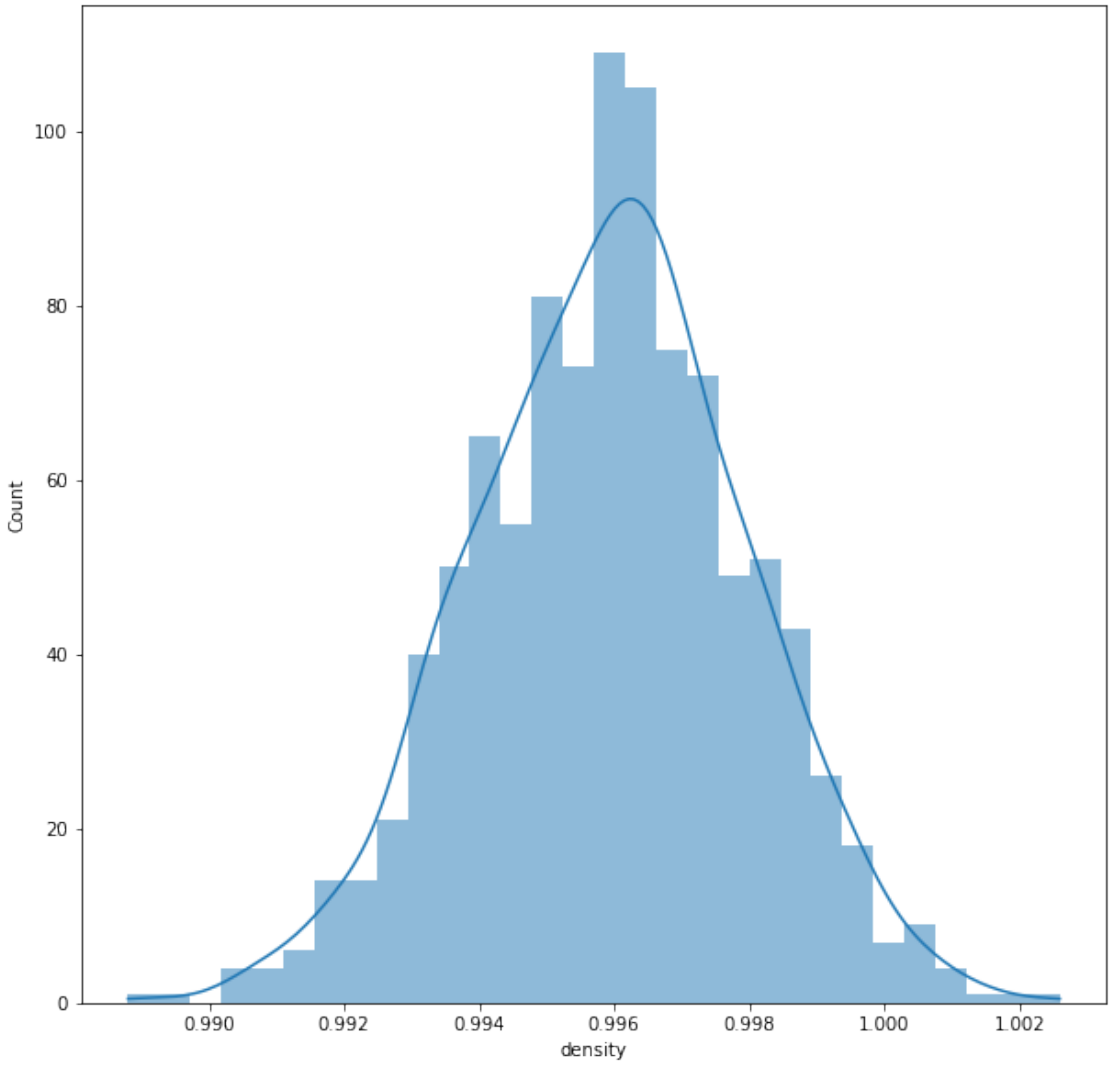

```
p-value shapiro-wilk: 0.17465530335903168
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom chlorides memiliki
distribusi normal

p-value d'agostino-pearson: 0.17048274704296862
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom chlorides
memiliki distribusi normal
```



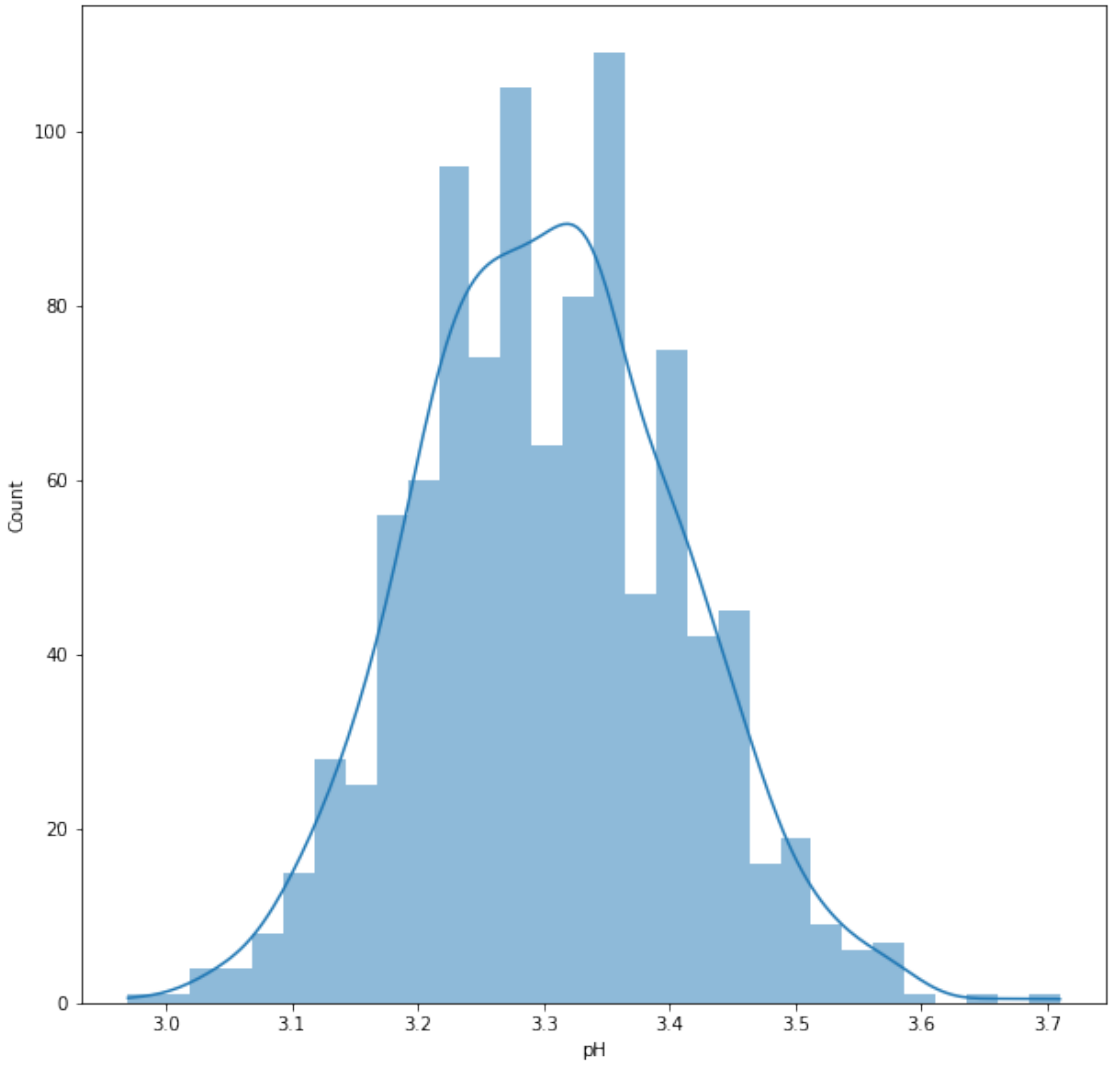

```
p-value shapiro-wilk: 0.04255827143788338
Berdasarkan tes shapiro-wilk, hipotesis nol ditolak, kolom free sulfur dioxide
tidak memiliki distribusi normal

p-value d'agostino-pearson: 0.01743043451827735
Berdasarkan tes d'agostino-pearson, hipotesis nol ditolak, kolom free sulfur
dioxide tidak memiliki distribusi normal
```



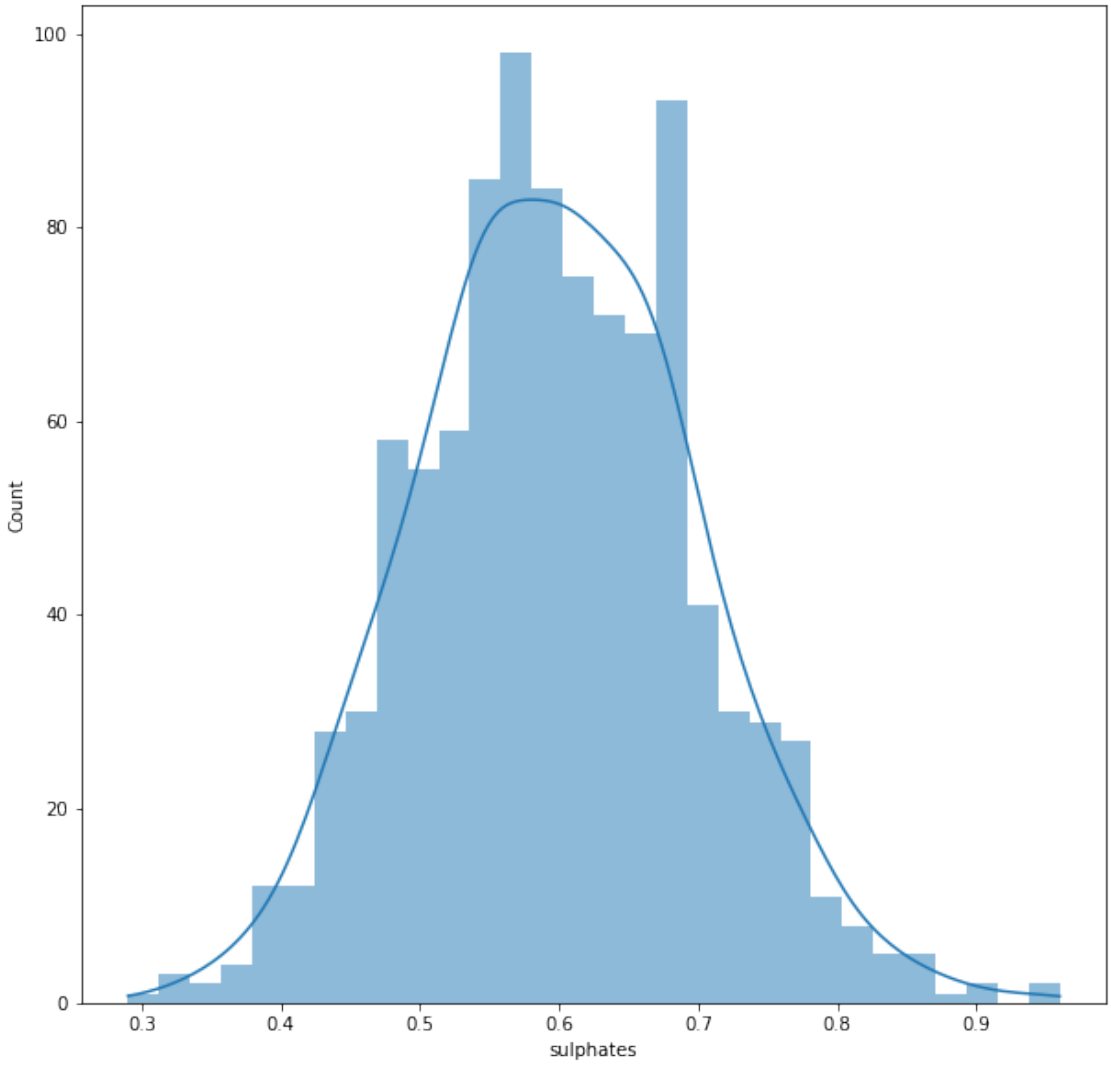

```
p-value shapiro-wilk: 0.5367269515991211
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom total sulfur dioxide
memiliki distribusi normal

p-value d'agostino-pearson: 0.8488846101395726
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom total sulfur
dioxide memiliki distribusi normal
```



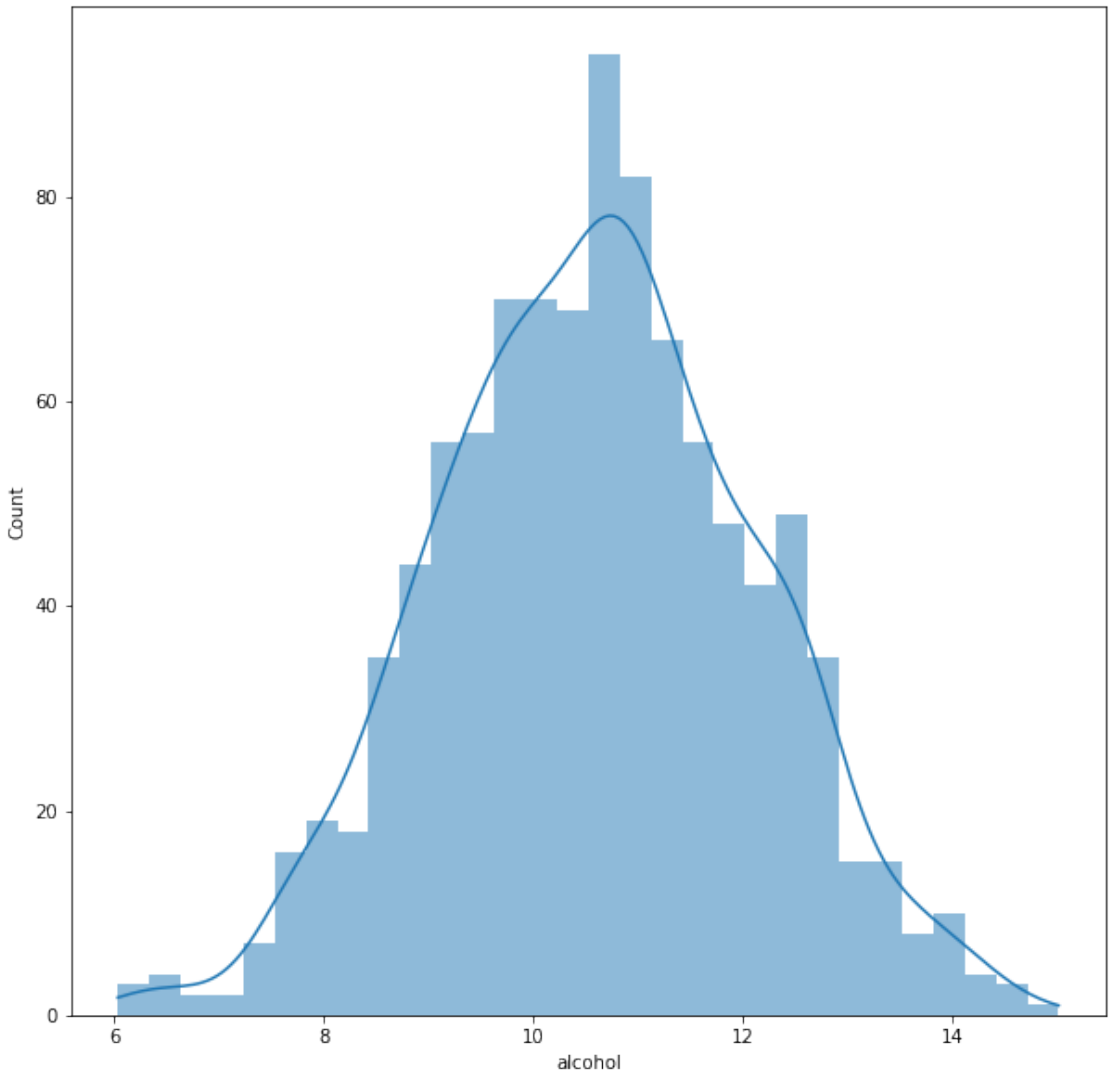

```
p-value shapiro-wilk: 0.8533204793930054
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom density memiliki
distribusi normal

p-value d'agostino-pearson: 0.5985227325531981
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom density
memiliki distribusi normal
```




```
p-value shapiro-wilk: 0.13713516294956207
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom pH memiliki
distribusi normal

p-value d'agostino-pearson: 0.13678740824860436
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom pH memiliki
distribusi normal
```




```
p-value shapiro-wilk: 0.11214283108711243
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom sulphates memiliki
distribusi normal

p-value d'agostino-pearson: 0.13884318628391681
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom sulphates
memiliki distribusi normal
```




```
p-value shapiro-wilk: 0.519870400428772
Berdasarkan tes shapiro-wilk, hipotesis nol diterima, kolom alcohol memiliki
distribusi normal

p-value d'agostino-pearson: 0.6790884901361043
Berdasarkan tes d'agostino-pearson, hipotesis nol diterima, kolom alcohol
memiliki distribusi normal
```

4 Soal 4

Melakukan test hiposis 1 sampel

[20]:
```
import pandas as pd
import scipy.stats as st
from statsmodels.stats.weightstats import ztest
from statsmodels.stats.proportion import proportions_ztest
```

[21]:
```
df = pd.read_csv("anggur.csv")
df
```

[21]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	5.90	0.4451	0.1813	2.049401	0.070574	
1	8.40	0.5768	0.2099	3.109590	0.101681	
2	7.54	0.5918	0.3248	3.673744	0.072416	
3	5.39	0.4201	0.3131	3.371815	0.072755	
4	6.51	0.5675	0.1940	4.404723	0.066379	
..	…	…	…	…	…	
995	7.96	0.6046	0.2662	1.592048	0.057555	
996	8.48	0.4080	0.2227	0.681955	0.051627	
997	6.11	0.4841	0.3720	2.377267	0.042806	
998	7.76	0.3590	0.3208	4.294486	0.098276	
999	5.87	0.5214	0.1883	2.179490	0.052923	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	16.593818	42.27	0.9982	3.27	0.71	
1	22.555519	16.01	0.9960	3.35	0.57	
2	9.316866	35.52	0.9990	3.31	0.64	
3	18.212300	41.97	0.9945	3.34	0.55	
4	9.360591	46.27	0.9925	3.27	0.45	
..	…	…	…	…	…	
995	14.892445	44.61	0.9975	3.35	0.54	
996	23.548965	25.83	0.9972	3.41	0.46	
997	21.624585	48.75	0.9928	3.23	0.55	
998	12.746186	44.53	0.9952	3.30	0.66	
999	16.203864	24.37	0.9983	3.29	0.70	

	alcohol	quality
0	8.64	7
1	10.03	8
2	9.23	8
3	14.07	9
4	11.49	8
..	…	…
995	10.41	8
996	9.91	8
997	9.94	7
998	9.76	8
999	10.17	7

```
[1000 rows x 12 columns]
```

4.1 Nilai rata - rata pH di atas 3.29?

Menggunakan data pH keseluruhan sebagai sampel

$H_0 : \mu = 3.29$

$H_1 : \mu > 3.29$

Hasil tes diuji dilakukan dengan signifikansi $ $ = 0.05$

Uji statistik dapat dilakukan dengan one tailed test ke arah kanan karena mencari $\mu > \mu_0$ pada H_1. Dengan demikian, critical section adalah $z > z_\alpha$

Nilai z didapatkan dengan rumus

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 ditolak jika $z > z_\alpha$ dan diterima jika $z \leq z_\alpha$

[22]:
```
mu_0 = 3.29
alpha = 0.05
z_a = st.norm.ppf(1-alpha)
print("Nilai z_a adalah: ", z_a)
print("Nilai p_a adalah: ", alpha)

z, p = ztest(df['pH'], value=mu_0)
print("\nNilai z adalah: ", z)
print("Nilai p adalah: ", p)

if(z > z_a):
    print()
    print(z, " > ", z_a)
    print(p, " < ", alpha)
    print("\nKarena z > z_a (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪dapat ditolak")
else:
    print()
    print(z, " <= ", z_a)
    print(p, " >= ", alpha)
    print("\nKarena z <= z_a (tidak berada pada daerah kritis) dan p >= p_a,␣
 ↪maka H0 tidak dapat ditolak")
```

```
Nilai z_a adalah:  1.6448536269514722
Nilai p_a adalah:  0.05

Nilai z adalah:  4.1037807933651145
Nilai p adalah:  4.0645260086604666e-05

4.1037807933651145  >  1.6448536269514722
```

```
4.0645260086604666e-05  <  0.05

Karena z > z_a (berada pada daerah kritis) dan p < p_a, maka H0 dapat ditolak
```

Didapatkan $z = 4.1038$ dan $z_\alpha = 1.6448$. Karena $z > z_a$ dan $p < \alpha$, maka z terdapat pada area kritis dan H_0 dapat ditolak.

Kesimpulan: Nilai rata - rata pH di atas 3.29

4.2 Nilai rata-rata Residual Sugar tidak sama dengan 2.50?

Menggunakan data Residual Sugar keseluruhan sebagai sampel

$H_0 : \mu = 2.50$

$H_1 : \mu \neq 2.50$

Hasil tes diuji dilakukan dengan signifikansi $ $ = 0.05$

Uji statistik dapat dilakukan dengan two tailed test ke arah kanan karena mencari $\mu \neq \mu_0$ pada H_1. Dengan demikian, critical section adalah $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$

Nilai z didapatkan dengan rumus

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 ditolak jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ dan diterima jika $-z_{\alpha/2} \geq z \geq z_{\alpha/2}$

[23]:
```
mu_0 = 2.50
alpha = 0.05
z_a2 = st.norm.ppf(1-alpha/2)
print("Nilai z_a adalah: ", z_a2)
print("Nilai -z_a adalah: ", -z_a2)
print("Nilai p_a adalah: ", alpha)

z, p = ztest(df['residual sugar'], value=mu_0)
print("\nNilai z adalah: ", z)
print("Nilai p adalah: ", p)

if(z < -z_a2):
    print()
    print(z, " < ", -z_a2)
    print(p, " < ", alpha)
    print("\nKarena z< -z_a2 (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪dapat ditolak")
elif(z > z_a2):
    print()
    print(z, " > ", z_a2)
    print(p, " < ", alpha)
    print("\nKarena z > z_a2 (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪dapat ditolak")
else:
```

```
    print()
    print(-z_a2, "<= ", z, " <= ", z_a2)
    print(p, " > ", alpha)
    print("\nKarena -z_a2 <=z <= z_a2 (tidak berada pada daerah kritis) dan p␣
↪>= p_a, maka H0 tidak dapat ditolak")
```

```
Nilai z_a adalah:  1.959963984540054
Nilai -z_a adalah:  -1.959963984540054
Nilai p_a adalah:  0.05

Nilai z adalah:  2.1479619435539523
Nilai p adalah:  0.031716778818727434

2.1479619435539523  >  1.959963984540054
0.031716778818727434  <  0.05

Karena z > z_a2 (berada pada daerah kritis) dan p < p_a, maka H0 dapat ditolak
```

Didapatkan $z = 2.1479$ dan $z_{\alpha/2} = 1.9599$. Karena $z > z_a$ dan $p < \alpha$, maka z terdapat pada area kritis bagian kanan dan H_0 dapat ditolak.

Kesimpulan: Nilai rata - rata Residual Sugar tidak sama dengan 2.50

4.3 Nilai rata-rata 150 baris pertama kolom sulphates bukan 0.65?

Menggunakan data Sulphates dengan 150 kolom pertama sebagai sampel

$H_0 : \mu = 0.65$

$H_1 : \mu \neq 0.65$

Hasil tes diuji dilakukan dengan signifikansi $ $ = 0.05$

Uji statistik dapat dilakukan dengan two tailed test ke arah kanan karena mencari $\mu \neq \mu_0$ pada H_1. Dengan demikian, critical section adalah $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$

Nilai z didapatkan dengan rumus

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 ditolak jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ dan diterima jika $-z_{\alpha/2} \geq z \geq z_{\alpha/2}$

[24]:
```
mu_0 = 0.65
alpha = 0.05
z_a2 = st.norm.ppf(1-alpha/2)
print("Nilai z_a adalah: ", z_a2)
print("Nilai -z_a adalah: ", -z_a2)
print("Nilai p_a adalah: ", alpha)

z, p = ztest(df['sulphates'].head(150), value=mu_0)
print("\nNilai z adalah: ", z)
print("Nilai p adalah: ", p)
```

```
if(z < -z_a2):
    print()
    print(z, " < ", -z_a2)
    print(p, " < ", alpha)
    print("\nKarena z< -z_a2 (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪dapat ditolak")
elif(z > z_a2):
    print()
    print(z, " > ", z_a2)
    print(p, " < ", alpha)
    print("\nKarena z > z_a2 (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪dapat ditolak")
else:
    print()
    print(-z_a2, "<= ", z, " <= ", z_a2)
    print(p, " > ", alpha)
    print("\nKarena -z_a2 <=z <= z_a2 (tidak berada pada daerah kritis) dan p >␣
 ↪p_a, maka H0 tidak dapat ditolak")
```

```
Nilai z_a adalah:  1.959963984540054
Nilai -z_a adalah:  -1.959963984540054
Nilai p_a adalah:  0.05

Nilai z adalah:  -4.964843393315918
Nilai p adalah:  6.875652918327359e-07

-4.964843393315918  <  -1.959963984540054
6.875652918327359e-07  <  0.05

Karena z< -z_a2 (berada pada daerah kritis) dan p < p_a, maka H0 dapat ditolak
```

Didapatkan $z = -4.9648$ dan $z_{\alpha/2} = 1.9599$. Karena $z < -z_a$ dan $p < \alpha$, maka z terdapat pada area kritis bagian kiri dan H_0 dapat ditolak.

Kesimpulan: Nilai rata - rata 150 kolom pertama Sulphates tidak sama dengan 0.65

4.4 Nilai rata-rata total sulfur dioxide di bawah 35?

Menggunakan data Total Sulphur Dioxide keseluruhan sebagai sampel

$H_0 : \mu = 35$

$H_1 : \mu < 35$

Hasil tes diuji dilakukan dengan signifikansi $ $= 0.05$$

Uji statistik dapat dilakukan dengan one tailed test ke arah kanan karena mencari $\mu < \mu_0$ pada H_1. Dengan demikian, critical section adalah $z < -z_\alpha$

Nilai z didapatkan dengan rumus

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 ditolak jika $z < z_\alpha$ dan diterima jika $z \geq z_\alpha$

[25]:
```
mu_0 = 35
alpha = 0.05
z_a = st.norm.ppf(1-alpha)
print("Nilai z_a adalah: ", z_a)
print("Nilai p_a adalah: ", alpha)

z, p = ztest(df['total sulfur dioxide'], value=mu_0)

#karena tailed test ke kiri maka p perlu dibalik
p = 1 - p

print("\nNilai z adalah: ", z)
print("Nilai p adalah: ", p)

if(z < z_a):
    print()
    print(z, " < ", z_a)
    print(p, " < ", alpha)
    print("\nKarena z < z_a (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪ditolak")
else:
    print()
    print(z, " >= ", z_a)
    print(p, " >= ", alpha)
    print("\nKarena z >= z_a (tidak berada pada daerah kritis) dan p > p_a,␣
 ↪maka H0 tidak dapat ditolak")
```

```
Nilai z_a adalah:  1.6448536269514722
Nilai p_a adalah:  0.05

Nilai z adalah:  16.786387372296744
Nilai p adalah:  1.0

16.786387372296744  >=  1.6448536269514722
1.0  >=  0.05

Karena z >= z_a (tidak berada pada daerah kritis) dan p > p_a, maka H0 tidak
dapat ditolak
```

Didapatkan $z = 16.7864$ dan $z_\alpha = 1.6448$. Karena $z \geq z_a$ dan $p \geq \alpha$, maka z tidak terdapat pada area kritis dan H_0 tidak dapat ditolak.

Kesimpulan: Nilai rata - rata Total Sulfur Dioxide tidak di bawah 35

4.5 Proporsi nilai total sulfur dioxide yang lebih dari 40, adalah tidak sama dengan 50% ?

Menggunakan data Total Sulphur Dioxide keseluruhan sebagai sampel dan data Total Sulphur Dioxide dengan nilai lebih dari 40 sebagai sampel yang memenuhi kondisi

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

Hasil tes diuji dilakukan dengan signifikansi $ $ = 0.05$

Uji statistik dapat dilakukan dengan uji proporsi two tailed test ke arah kanan karena mencari $p \neq p_0$ pada H_1. Tes statistik dapat dilakukan dengan menggunakan tes binomial didekati normal karena statistik memiliki karakteristik normal berdasarkan tes yang dilakukan pada no 3. Dengan demikian, critical section adalah $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$

Nilai z didapatkan dengan rumus

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

H_0 ditolak jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ dan diterima jika $-z_{\alpha/2} \geq z \geq z_{\alpha/2}$

[26]:
```
p_0 = 0.5
alpha_2 = 0.05/2
z_a2 = st.norm.ppf(1-alpha/2)
print("Nilai z_a/2 adalah: ", z_a2)
print("Nilai -z_a/2 adalah: ", -z_a2)
print("Nilai p_a adalah: ", alpha)

satisfy = len(df[df['total sulfur dioxide'] > 40])
total = len(df)

z, p = proportions_ztest(satisfy, total, p_0)
print("\nNilai z adalah: ", z)
print("Nilai p adalah: ", p)

if(z < -z_a2):
    print()
    print(z, " < ", z_a2)
    print(p, " < ", alpha)
    print("\nKarena z< -z_a2 (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪ditolak")
elif(z > z_a2):
    print()
    print(z, " > ", z_a2)
    print(p, " < ", alpha)
    print("\nKarena z > z_a2 (berada pada daerah kritis) dan p < p_a, maka H0␣
 ↪ditolak")
else:
```

```
    print()
    print(-z_a2, "<= ", z, " <= ", z_a2)
    print(p, " >= ", alpha)
    print("\nKarena -z_a2 <=z <= z_a2 (tidak berada pada daerah kritis) dan p␣
↪>= p_a, maka H0 tidak dapat ditolak")
```

```
Nilai z_a/2 adalah:  1.959963984540054
Nilai -z_a/2 adalah:  -1.959963984540054
Nilai p_a adalah:  0.05

Nilai z adalah:  0.7591653095427344
Nilai p adalah:  0.4477536749931885

-1.959963984540054 <=  0.7591653095427344  <=  1.959963984540054
0.4477536749931885  >=  0.05

Karena -z_a2 <=z <= z_a2 (tidak berada pada daerah kritis) dan p >= p_a, maka H0
tidak dapat ditolak
```

Didapatkan $z = 0.7591$ dan $z_{\alpha/2} = 1.9599$. Karena $-z_a \leq z \leq z_a$ dan $p \geq \alpha$, maka z tidak terdapat pada area kritis dan H_0 tidak dapat ditolak.

Kesimpulan: Proporsi nilai total sulfur dioxide yang lebih dari 40 adalah sama dengan 50%

5 Soal 5

Melakukan test hiposis 2 sampel

[27]:
```
import pandas as pd
import scipy.stats as st
from statsmodels.stats.weightstats import ztest
from statsmodels.stats.proportion import proportions_ztest
```

[28]:
```
df = pd.read_csv("anggur.csv")
df
```

[28]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	5.90	0.4451	0.1813	2.049401	0.070574	
1	8.40	0.5768	0.2099	3.109590	0.101681	
2	7.54	0.5918	0.3248	3.673744	0.072416	
3	5.39	0.4201	0.3131	3.371815	0.072755	
4	6.51	0.5675	0.1940	4.404723	0.066379	
..	…	…	…	…	…	
995	7.96	0.6046	0.2662	1.592048	0.057555	
996	8.48	0.4080	0.2227	0.681955	0.051627	
997	6.11	0.4841	0.3720	2.377267	0.042806	
998	7.76	0.3590	0.3208	4.294486	0.098276	
999	5.87	0.5214	0.1883	2.179490	0.052923	

```
     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0              16.593818                 42.27   0.9982  3.27       0.71
1              22.555519                 16.01   0.9960  3.35       0.57
2               9.316866                 35.52   0.9990  3.31       0.64
3              18.212300                 41.97   0.9945  3.34       0.55
4               9.360591                 46.27   0.9925  3.27       0.45
..                   …                     …      …   …       …
995            14.892445                 44.61   0.9975  3.35       0.54
996            23.548965                 25.83   0.9972  3.41       0.46
997            21.624585                 48.75   0.9928  3.23       0.55
998            12.746186                 44.53   0.9952  3.30       0.66
999            16.203864                 24.37   0.9983  3.29       0.70

     alcohol  quality
0       8.64        7
1      10.03        8
2       9.23        8
3      14.07        9
4      11.49        8
..       …       …
995    10.41        8
996     9.91        8
997     9.94        7
998     9.76        8
999    10.17        7

[1000 rows x 12 columns]
```

5.1 Data kolom fixed acidity dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata kedua bagian tersebut sama?

Menggunakan data Fixed Acidity keseluruhan sebagai sampel

$H_{0}: $ Rataan kolom bagian awal sama dengan rataan kolom bagian akhir fixed acidity ($\mu_1 - \mu_2 = 0$)

$H_{1}: $ Rataan kolom bagian awal tidak sama dengan rataan kolom bagian akhir fixed acidity ($\mu_1 - \mu_2 \neq 0$)

Hasil tes diuji dilakukan dengan signifikansi $ $= 0.05$$

Uji statistik dapat dilakukan dengan two tailed test yaitu bagian kiri dengan $z < -z_{\alpha/2}$ dan bagian kanan dengan $z > z_{\alpha/2}$

Nilai z didapatkan dengan rumus

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

H_0 ditolak jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ dan $p < \alpha$ H_0 diterima jika $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$ dan $p \geq a$

[29]:
```
delta = 0
alpha = 0.05

len_half_data = len(df) // 2
df_awal = df["fixed acidity"][:len_half_data]
df_akhir = df["fixed acidity"][len_half_data:]
z, p = ztest(df_awal, df_akhir, value=delta)

z_a2 = st.norm.ppf(1-(alpha/2))

print("Nilai z_a2 adalah : ", z_a2)
print("Nilai p_a adalah: ", alpha)

print("\nNilai z adalah : ", z)
print(f"Nilai p adalah : {p}\n")


if (z < -z_a2 and (p < alpha)) :
    print(z, " < ", -z_a2, "dan", p , " < ", alpha)
    print("\nKarena z < -z_a2 dan p < alpha (berada pada daerah kritis), maka␣
 ↪H0 dapat ditolak")
elif ((z > z_a2) and (p < alpha)) :
    print(z, " > ", z_a2, "dan", p , " < ", alpha)
    print("\nKarena z > z_a2 dan p < alpha (berada pada daerah kritis), maka H0␣
 ↪dapat ditolak")
else : # (-z_a2 <= z <= z_a2 and p >= alpha)
    print(-z_a2 , " <= ", z, " <= ", z_a2, "dan", p, " >= ", alpha)
    print("\nKarena -z_a2 <= z <= z_a2 dan p >= alpha (tidak berada pada daerah␣
 ↪kritis), maka H0 tidak dapat ditolak")
```

```
Nilai z_a2 adalah :  1.959963984540054
Nilai p_a adalah:  0.05

Nilai z adalah :  0.02604106999906379
Nilai p adalah : 0.9792245804254097

-1.959963984540054  <=  0.02604106999906379  <=  1.959963984540054 dan
0.9792245804254097  >=  0.05

Karena -z_a2 <= z <= z_a2 dan p >= alpha (tidak berada pada daerah kritis), maka
H0 tidak dapat ditolak
```

Didapatkan $z = 0.026$ dan $z_{\alpha/2} = 1.9599$. Karena $-z_{a/2} < z < z_{a/2}$ dan $p \geq \alpha$, maka z tidak terdapat pada area kritis dan H_0 tidak dapat ditolak.

Kesimpulan: Rataan kolom bagian awal sama dengan rataan kolom bagian akhir fixed acidity

5.2 Data kolom chlorides dibagi 2 sama rata: bagian awal dan bagian akhir kolom. Benarkah rata-rata bagian awal lebih besar daripada bagian akhir sebesar 0.001?

Menggunakan data Chlorides keseluruhan sebagai sampel

$H_{0}: $ Rataan kolom bagian awal sama dengan rataan kolom bagian akhir chlorides ditambah 0.001 ($\mu_1 - \mu_2 = 0.001$)

$H_{1}: $ Rataan kolom bagian awal tidak sama dengan rataan kolom bagian akhir chlorides ditambah 0.001 ($\mu_1 - \mu_2 \neq 0.001$)

Hasil tes diuji dilakukan dengan signifikansi $ $= 0.05$

Uji statistik dapat dilakukan dengan two tailed test yaitu bagian kiri dengan $z < -z_{\alpha/2}$ dan bagian kanan dengan $z > z_{\alpha/2}$

Nilai z didapatkan dengan rumus

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

H_0 ditolak jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ dan $p < \alpha$ H_0 diterima jika $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$ dan $p \geq a$

[30]:
```
delta = 0.001
alpha = 0.05

len_half_data = len(df) // 2
df_awal = df["chlorides"][:len_half_data]
df_akhir = df["chlorides"][len_half_data:]
z, p = ztest(df_awal, df_akhir, value=delta)

z_a2 = st.norm.ppf(1-(alpha/2))

print("Nilai z_a2 adalah : ", z_a2)
print("Nilai p_a adalah: ", alpha)

print("\nNilai z adalah : ", z)
print(f"Nilai p adalah : {p}\n")

if (z < -z_a2 and (p < alpha)) :
    print(z, " < ", -z_a2, "dan", p , " < ", alpha)
    print("\nKarena z < -z_a2 dan p < alpha (berada pada daerah kritis), maka␣
 ↪H0 ditolak")
elif ((z > z_a2) and (p < alpha)) :
    print(z, " > ", z_a2, "dan", p , " < ", alpha)
    print("\nKarena z > z_a2 dan p < alpha (berada pada daerah kritis), maka H0␣
 ↪ditolak")
else :# (-z_a2 <= z <= z_a2 and p >= alpha)
    print(-z_a2 , " <= ", z, " <= ", z_a2, "dan", p, " >= ", alpha)
```

```
    print("\nKarena -z_a2 <= z <= z_a2 dan p >= alpha (tidak berada pada daerah␣
↪kritis), maka H0 tidak dapat ditolak")
```

```
Nilai z_a2 adalah :  1.959963984540054
Nilai p_a adalah:  0.05

Nilai z adalah :  -0.467317122852132
Nilai p adalah : 0.640273007581107

-1.959963984540054  <=  -0.467317122852132  <=  1.959963984540054 dan
0.640273007581107  >=  0.05

Karena -z_a2 <= z <= z_a2 dan p >= alpha (tidak berada pada daerah kritis), maka
H0 tidak dapat ditolak
```

Didapatkan $z = -0.4673$ dan $z_{\alpha/2} = 1.9599$. Karena $-z_{a/2} < z < z_{a/2}$ dan $p \geq \alpha$, maka z tidak terdapat pada area kritis dan H_0 tidak dapat ditolak.

Kesimpulan: Rataan kolom bagian awal sama dengan rataan kolom bagian akhir chlorides ditambah 0.001

5.3 Benarkah rata-rata sampel 25 baris pertama kolom Volatile Acidity sama dengan rata-rata 25 baris pertama kolom Sulphates ?

Menggunakan data Volatile Acidity dan Sulphates sebagian sebagai sampel

$H_{0}: $ Rataan 25 baris pertama sama kolom Volatile Acidity dengan rataan 25 baris pertama kolom Sulphates $(\mu_1 - \mu_2 = 0)$

H_1 : Rataan 25 baris pertama tidak sama kolom Volatile Acidity dengan rataan 25 baris pertama kolom Sulphates $(\mu_1 - \mu_2 \neq 0)$

Hasil tes diuji dilakukan dengan signifikansi $ = 0.05$

Uji statistik dapat dilakukan dengan two tailed test yaitu bagian kiri dengan $z < -z_{\alpha/2}$ dan bagian kanan dengan $z > z_{\alpha/2}$

Nilai z didapatkan dengan rumus

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

H_0 ditolak jika $z < -z_{\alpha/2}$ atau $z > z_{\alpha/2}$ dan $p < \alpha$ H_0 diterima jika $-z_{\alpha/2} \leq z \leq z_{\alpha/2}$ dan $p \geq a$

[31]:
```
delta = 0
alpha = 0.05

len_half_data = len(df) // 2
df_awal = df["volatile acidity"].head(25)
df_akhir = df["sulphates"].tail(25)
z, p = ztest(df_awal, df_akhir, value=delta)
```

```
z_a2 = st.norm.ppf(1-(alpha/2))

print("Nilai z_a2 adalah : ", z_a2)
print("Nilai p_a adalah: ", alpha)

print("\nNilai z adalah : ", z)
print(f"Nilai p adalah : {p}\n")

if (z < -z_a2 and (p < alpha)) :
    print(z, " < ", -z_a2, "dan", p , " < ", alpha)
    print("\nKarena z < -z_a2 dan p < alpha (berada pada daerah kritis), maka␣
 ↪H0 dapat ditolak")
elif ((z > z_a2) and (p < alpha)) :
    print(z, " > ", z_a2, "dan", p , " < ", alpha)
    print("\nKarena z > z_a2 dan p < alpha (berada pada daerah kritis), maka H0␣
 ↪dapat ditolak")
else : # (-z_a2 <= z <= z_a2 and p >= alpha)
    print(-z_a2, " <= ", z, " <= ", z_a2, "dan", p, " >= ", alpha)
    print("\nKarena -z_a2 <= z <= z_a2 dan p >= alpha (tidak berada pada daerah␣
 ↪kritis), maka H0 tidak dapat ditolak")
```

```
Nilai z_a2 adalah :  1.959963984540054
Nilai p_a adalah:  0.05

Nilai z adalah :  -3.9977861838398008
Nilai p adalah : 6.393766557154183e-05

-3.9977861838398008  <  -1.959963984540054 dan 6.393766557154183e-05  <  0.05

Karena z < -z_a2 dan p < alpha (berada pada daerah kritis), maka H0 dapat
ditolak
```

Didapatkan $z = -3.9978$ dan $z_{\alpha/2} = 1.9599$. Karena $z < -z_{a/2}$ dan $p < \alpha$, maka z terdapat pada area kritis bagian kiri dan H_0 dapat ditolak.

Kesimpulan: Rataan 25 baris pertama tidak sama kolom Volatile Acidity dengan rataan 25 baris pertama kolom Sulphates

5.4 Bagian awal kolom residual sugar memiliki variansi yang sama dengan bagian akhirnya?

Menggunakan data Residual Sugar keseluruhan sebagai sampel

$H_{0}: $ Variansi bagian awal kolom residual sugar sama dengan bagian akhirnya ($\sigma_1^2 = \sigma_2^2$)

H_1 : Variansi bagian awal kolom residual sugar tidak sama dengan bagian akhirnya ($\sigma_1^2 \neq \sigma_2^2$)

Hasil tes diuji dilakukan dengan signifikansi $ $= 0.05$

Uji statistik dapat dilakukan dengan two tailed test yaitu bagian kiri dengan $f < f_{1-\alpha/2}(v_1, v_2)$ dan bagian kanan dengan $f > f_{\alpha/2}(v_1, v_2)$

Nilai f didapatkan dengan rumus

$$f = \frac{s_1^2}{s_2^2}$$

Nilai v_1 dan v_2 diperolah dari

$$v_1 = n_1 - 1$$

$$v_2 = n_2 - 1$$

H_0 ditolak jika $f > f_{\alpha/2}(v_1, v_2)$ atau $f < f_{1-\alpha/2}(v_1, v_2)$ dan $p < \alpha$ H_0 diterima jika $f_{1-\alpha/2}(v_1, v_2) \leq f \leq f_{\alpha/2}(v_1, v_2)$ dan $p \geq a$

[32]:
```
alpha = 0.05
len_half_data = len(df) // 2
df_awal = df["residual sugar"][:len_half_data]
df_akhir = df["residual sugar"][len_half_data:]
v1 = len(df_awal) - 1
v2 = len(df_akhir) - 1

f = df_awal.var() / df_akhir.var()
f_kiri = st.f.ppf(1 - (alpha / 2), v1, v2)
f_kanan = st.f.ppf((alpha / 2), v1, v2)

p = 1 - st.f.cdf(f, v1, v2)


print("Nilai p_a adalah:", alpha)
print("Nilai f_kiri adalah: ", f_kiri)
print("Nilai f_kanan adalah: ", f_kanan)

print("\nNilai f adalah: ", f)
print(f"Nilai p adalah: {p}\n")

if (f > f_kanan and p < alpha) :
    print(f, " > ", f_kanan, "dan", p , " < ", alpha)
    print("\nKarena f > f_down dan p < alpha (berada pada daerah kritis), maka␣
 ↪H0 dapat ditolak")
elif (f < f_kiri and p < alpha) :
    print(f, " < ", f_kiri, "dan", p , " < ", alpha)
    print("\nKarena f < f_up dan p < alpha (berada pada daerah kritis), maka H0␣
 ↪dapat ditolak")
else : # (f_kiri <= f <= f_kanan and p >= alpha)
    print(f_kiri, " <= ", f, " <= ", f_kanan, "dan", p, " >= ", alpha)
    print("\nKarena f_kiri <= f <= f_kanan dan p >= alpha (tidak berada pada␣
 ↪daerah kritis), maka H0 tidak dapat ditolak")
```

```
Nilai p_a adalah: 0.05
Nilai f_kiri adalah:  1.1920574017201653
Nilai f_kanan adalah:  0.8388857772763105
```

```
Nilai f adalah:  0.9420041066941615
Nilai p adalah: 0.747589820237691

1.1920574017201653  <=  0.9420041066941615  <=  0.8388857772763105 dan
0.747589820237691  >=  0.05

Karena f_kiri <= f <= f_kanan dan p >= alpha (tidak berada pada daerah kritis),
maka H0 tidak dapat ditolak
```

Didapatkan $f = 0.9420$, $f_{kiri} = 1.1920$, dan $f_{kanan} = 0.8389$. Karena $f_{kiri} < f < f_{kanan}$ dan $p \geq \alpha$, maka f tidak terdapat pada area kritis dan H_0 tidak dapat ditolak.

Kesimpulan: Variansi bagian awal kolom residual sugar tidak sama dengan bagian akhirnya

5.5 Proporsi nilai setengah bagian awal alcohol yang lebih dari 7, adalah lebih besar daripada, proporsi nilai yang sama di setengah bagian akhir alcohol?

Menggunakan data Alcohol keseluruhan sebagai sampel

$H_{0}: $ Proporsi nilai bagian awal alcohol yang lebih dari 7 lebih daripada proporsi nilai yang sama di setengah nilai yang sama di setengah bagian akhirnya $(p_1 - p_2 = 0)$

H_1 : Proporsi nilai bagian awal alcohol yang lebih dari 7 lebih daripada proporsi nilai yang sama di setengah nilai yang sama di setengah bagian akhirnya $(p_1 - p_2 > 0)$

Hasil tes diuji dilakukan dengan signifikansi $ = 0.05$

Uji statistik dapat dilakukan dengan one tailed test ke arah kanan dengan mencari $z > z_\alpha$

Nilai z didapatkan dengan rumus

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Nilai $\hat{p}$ diperolah dari

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

dengan $\hat{q}$ adalah

$$\hat{q} = 1 - \hat{p}$$

H_0 ditolak jika $z > z_\alpha$ dan $p < \alpha$ H_0 diterima jika $z \leq z_\alpha$ dan $p \geq \alpha$

[33]:
```
delta = 0
alpha = 0.05

len_half_data = len(df) // 2
df_awal = df[:len_half_data]
df_akhir = df[len_half_data:]

z, p = proportions_ztest([len(df_awal[df_awal["alcohol"] > 7]),
                          len(df_akhir[df_akhir["alcohol"] > 7])],
```

```
                          [len(df_awal), len(df_akhir)],
                          value=delta, prop_var=delta)

z_a = st.norm.ppf(1 - alpha)

print("Nilai z_a adalah : ", z_a)
print("Nilai p_a adalah: ", alpha)

print("\nNilai z adalah : ", z)
print(f"Nilai p adalah : {p}\n")

if (z > z_a and p < alpha):
    print(z, " > ", z_a, "dan", p , " < ", alpha)
    print("\nKarena z > z_a dan p < alpha (berada pada daerah kritis), maka H0␣
 ↪dapat ditolak")
else : # (z <= z_a and p >= alpha)
    print(z, " <= ", z_a, "dan", p , " >= ", alpha)
    print("\nKarena z <= z_a dan p >= alpha (tidak berada pada daerah kritis),␣
 ↪maka H0 tidak dapat ditolak")
```

```
Nilai z_a adalah :  1.6448536269514722
Nilai p_a adalah:  0.05

Nilai z adalah :  0.0
Nilai p adalah : 1.0

0.0  <=  1.6448536269514722 dan 1.0  >=  0.05

Karena z <= z_a dan p >= alpha (tidak berada pada daerah kritis), maka H0 tidak
dapat ditolak
```

Didapatkan $z = 0$ dan $z_\alpha = 1.645$. Karena $z < z_\alpha$ dan $p \geq \alpha$, maka z tidak terdapat pada area kritis dan H_0 tidak dapat ditolak.

Kesimpulan: Proporsi nilai bagian awal alcohol yang lebih dari 7 lebih daripada proporsi nilai yang sama di setengah nilai yang sama di setengah bagian akhirnya