IE6200: Wine Quality Analysis

Group 2

Contents

1 Abstract

Wine is an alcoholic drink typically made from fermented grape juice. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide, and heat. Different varieties of grapes and strains of yeasts produce different styles of wine. It is a pleasant tasting alcoholic beverage, loved and celebrated . It will be interesting to analyze the physiochemical attributes of wine and understand their relationships and significance with wine quality.

2 Wine Quality Dataset

2.1 Description

The wine dataset used in this analysis is taken from the UCI Machine Learning repository with an objective to find out how the physiochemical properties affect the quality of wine. The data comprises of 4898 observations with 12 attributes. (Source:https://archive.ics.uci.edu/ml/datasets/wine+quality)

2.2 Attributes

Input variables (based on physiochemical tests):

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Output variable (based on sensory data):

12. quality (score between 0 and 10)

Understanding Wine Attributes and Properties

2.2.1 Acidity

Acids are one of fundamental traits in wine (the others are tannin, alcohol, sweetness and body) and contribute greatly to it's color, balance and taste. Acidity gives a wine its tart and sour taste. Fundamentally, all wines lie on the acidic side of the pH spectrum, and most range from 2.5 to about 4.5 pH (7 is neutral). There are several different types of acids found in wine, which will affect how acidic a wine tastes. The most prevalent acids found in wine are tartaric acid, malic acid, and citric acid.

Acid Types and Measures

- **fixed acidity:** Fixed acids include tartaric, malic, citric, and succinic acids which are found in grapes (except succinic). This variable is expressed in $\frac{g(tartaricacid)}{dm^3}$ in the dataset.
- **volatile acidity:** These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid though other acids like lactic, formic and butyric acids might also be present. Excess of volatile acids are undesirable and lead to unpleasant flavor. In the US, the legal limits of volatile acidity are 1.2 g/L for red table wine and 1.1 g/L for white table wine. The volatile acidity is expressed in $\frac{g(aceticacid)}{dm^3}$ in the dataset.
- **citric acid:** This is one of the fixed acids which gives a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness. It is expressed in $\frac{g}{dm^3}$ in the dataset.
- **pH:** Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. Solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.

2.2.2 Sweetness

How sweet or dry (not sweet) is the wine? During winemaking, yeast eats up sugar and makes ethanol (alcohol) as a by-product.

- Dry Wine - When the yeast is able to eat up all the sugar the result is a dry wine – higher in alcohol content and low in sugar.

- Sweet Wine - When the yeast is stopped by the winemaker (often by rapid chilling), the result is sweet wine - lower in alcohol content and high in sugar. This is why many sweet wines have less alcohol than dry wines.

Sweetness Measure:

- **residual sugar:** The sugar in wine is called "Residual Sugar" or RS. Grapes contain fruit sugars (fructose and glucose) and the residual sugar is what remains after the fermentation process stops, or is stopped. It's usually expressed in $\frac{g}{dm^3}$ in the dataset.
- **sweetness_category**: This is a derived attribute from the `residual sugar` attribute. We bucket or group sweetness measures into five qualitative buckets as follows:

Table 1: Sweetness Chart

Residual Sugar ($\frac{g}{dm^3}$)	Category
< 1	Bone Dry
1-10	Dry
10-35	Off-Dry
35-120	Sweet
120-220	Very Sweet

2.2.3 Salinity

Salinity is not a common wine descriptor. However, wine-producing countries have (widely varying) legal maximums for sodium chloride in wine. It is a concern in dry locations when frequent irrigation increases soil salinity, which increases wine salinity.

Salty Measure:

- **chlorides:** Chloride concentration in the wine is influenced by terroir and its highest levels are found in wines coming from countries where irrigation is carried out using salty water or in areas with brackish terrains. This is usually a major contributor to saltiness in wine. It's usually expressed in $\frac{g(sodiumchloride)}{dm^3}$ in the dataset.

2.2.4 Sulphites

Sulphites in wine are chemical compounds (sulphur dioxide, or SO2) that occur naturally, to a varying degree, in all types of wine. Sulfur Dioxide is naturally found in wines and is a byproduct of fermentation. But, most winemakers choose to add a little extra to prevent the growth of undesirable yeasts and microbes, as well as to protect against oxidation. It inhibits yeasts, preventing sweet wines from refermenting in the bottle. It's an antioxidant, keeping the wine fresh and untainted by oxygen. The maximum legal limit in the United States is 350 mg/l. A well made dry red wine typically has about 50 mg/l sulfites.

Sulfites Measure:

- **sulphates:** These are mineral salts containing sulfur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. They are connected to the fermentation process and affects the wine aroma and flavor. It is expressed in $\frac{g(potassiumsulphate)}{dm^3}$ in the dataset.
- **free sulfur dioxide:** This is the part of the sulphur dioxide when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. They are also known as sulfites and too much of it is undesirable and gives a pungent odor. This variable is expressed in $\frac{mg}{dm^3}$ in the dataset.
- **total sulfur dioxide:** This is the sum total of the bound and the free sulfur dioxide (SO_2). It is expressed in $\frac{mg}{dm^3}$ in the dataset. This is mainly added to kill harmful bacteria and preserve quality and freshness. There are usually legal limits for sulfur levels in wines and excess of it can even kill good yeast and give out undesirable odor.

2.2.5 Alcohol

Alcohol is formed as a result of yeast converting sugar during the fermentation process.

- Wines with higher alcohol tend to taste bolder and more oily.
- Wines with lower alcohol tend to taste lighter-bodied.
- Most wines range between 11–13% ABV.

Alcohol Measure:

- **alcohol:** It is measured in % vol or alcohol by volume (ABV).

2.2.6 Body

A wine's body describes how heavy or light a wine feels in the mouth. Wine body is broken down into three categories: light, medium or full-bodied. Body is the result of many factors – from wine variety, where it's from, vintage, alcohol level and how it's made. Body is a snapshot of the overall impression of a wine.

Body Measure:

- **density:** It is generally used as a measure of the conversion of sugar to alcohol. It is expressed in $\frac{g}{cm^3}$.

2.2.7 Classifications Attributes:

- **quality:** Wine experts grade the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.
- **quality_category:** This is a derived attribute from the `quality` attribute. We bucket or group wine quality scores into three qualitative buckets as follows:
    1. low - Wines with a quality score of 3, 4 & 5
    2. medium - Wines with a quality score of 6 & 7
    3. high - Wines with a quality scores of 8 & 9

3 Descriptive Statistics

Table 2: Descriptive Statistics

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfer_dioxide	total_sulpher_dioxide	density	pH	sulphates	alcohol	quality
sample_size	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00
missing_values	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	3.80	0.08	0.00	0.60	0.01	2.00	9.00	0.99	2.72	0.22	8.00	3.00
max	14.20	1.10	1.66	65.80	0.35	289.00	440.00	1.04	3.82	1.08	14.20	9.00
range	10.40	1.02	1.66	65.20	0.34	287.00	431.00	0.05	1.10	0.86	6.20	6.00
median	6.80	0.26	0.32	5.20	0.04	34.00	134.00	0.99	3.18	0.47	10.40	6.00
mode	6.80	0.28	0.30	1.20	0.04	29.00	111.00	0.99	3.14	0.50	9.40	6.00
mean	6.85	0.28	0.33	6.39	0.05	35.31	138.36	0.99	3.19	0.49	10.51	5.88
var	0.71	0.01	0.01	25.73	0.00	289.24	1806.09	0.00	0.02	0.01	1.51	0.78
std dev	0.84	0.10	0.12	5.07	0.02	17.01	42.50	0.00	0.15	0.11	1.23	0.89
cv	0.12	0.36	0.36	0.79	0.48	0.48	0.31	0.00	0.05	0.23	0.12	0.15
skewness	0.65	1.58	1.28	1.08	5.02	1.41	0.39	0.98	0.46	0.98	0.49	0.16
kurtosis	2.17	5.08	6.16	3.46	37.51	11.45	0.57	9.78	0.53	1.59	-0.70	0.21

3.1 PMF and CDF

- X ≡ R.V. of quality of a dry wine

Table 3: PMF and CDF

quality	count	PMF	CDF
3	13	0.004	0.004
4	128	0.035	0.039
5	966	0.268	0.307
6	1624	0.450	0.757
7	734	0.203	0.960
8	141	0.039	0.999
9	4	0.001	1.000

3.2 Joint Probability Distribution

We want to compare sweetness levels and quality scores of the wines. To do so, we construct Joint Probability Distribution

- X ≡ R.V. of sweetness category
- Y ≡ R.V. of quality

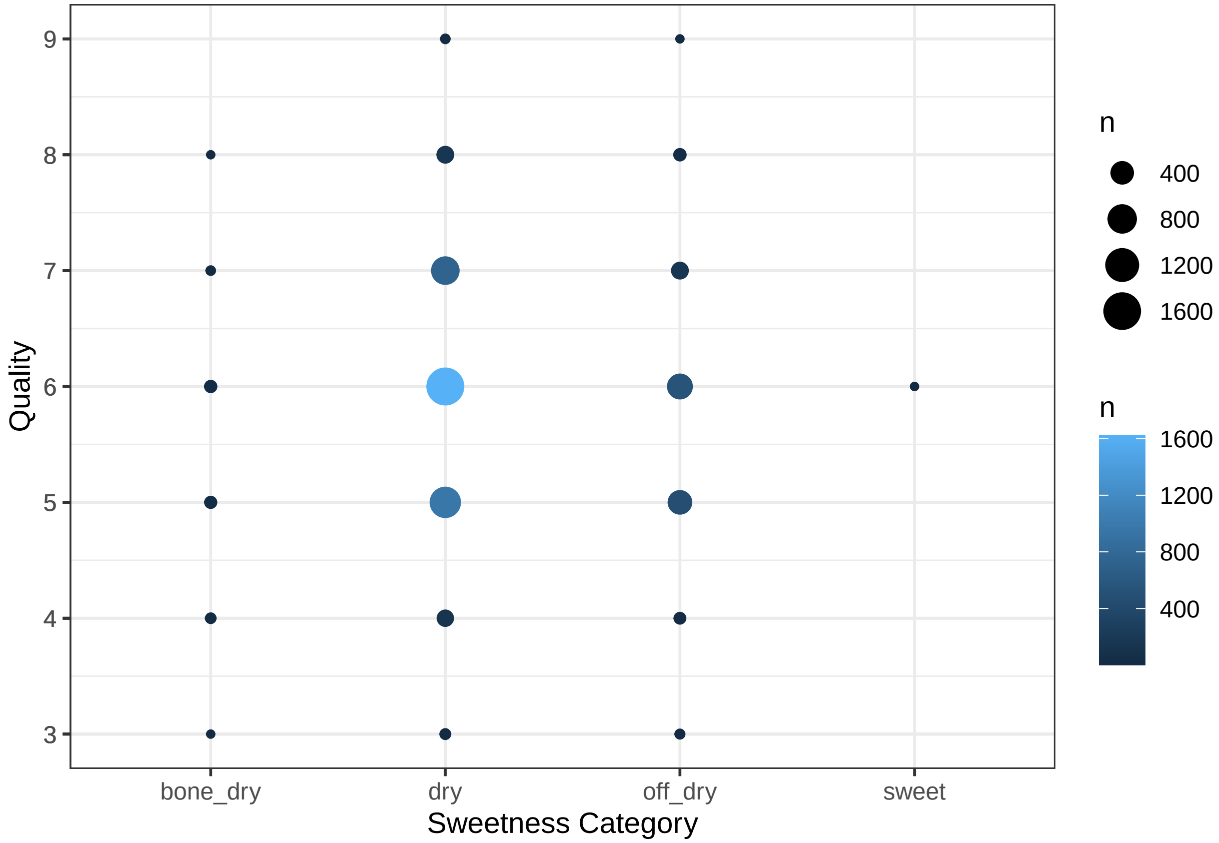


Table 4: Sweetness vs Quality

sweetness_category	3	4	5	6	7	8	9
bone_dry	0.000	0.002	0.006	0.006	0.001	0.000	0.000
dry	0.003	0.026	0.197	0.332	0.150	0.029	0.001
off_dry	0.001	0.005	0.094	0.111	0.029	0.007	0.000
sweet	0.000	0.000	0.000	0.000	0.000	0.000	0.000

- We infer that most of the wine samples in the dataset are of medium quality and dry wines.

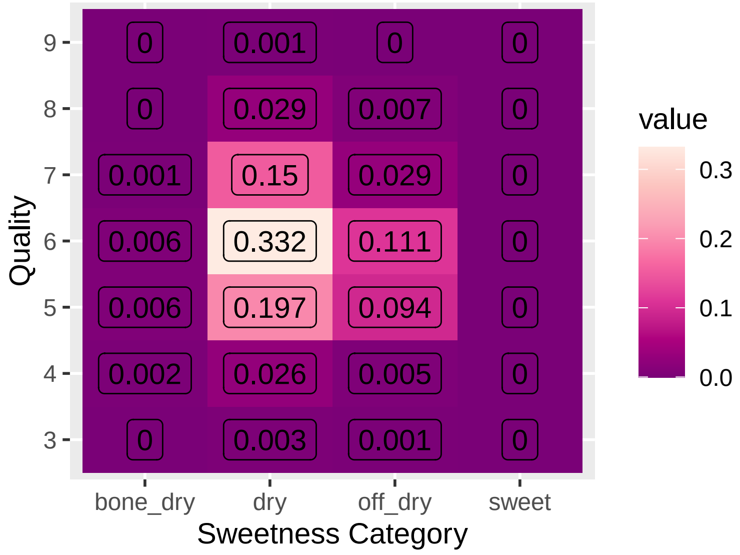
3.3 Correlation Coefficient

We want to know if there is any correlation between sweetness and quality of a wine. So, we calculate correlation between `residual_sugar` and `quality`

```
## [1] -0.09757683
```

Our data suggests that there is a weak or no correlation b/w Sweetness and quality.

3.4 Heatmap

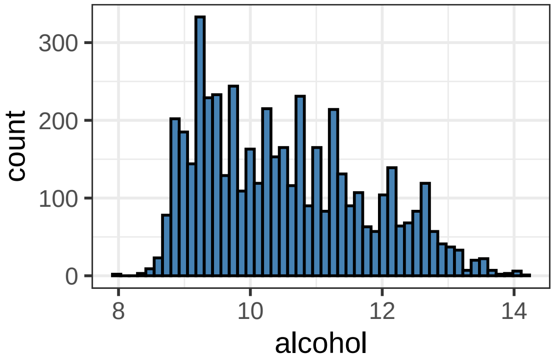


- We observe that most of the medium quality wines are dry wines(low in sweetness).

4 Goodness of Fit Test

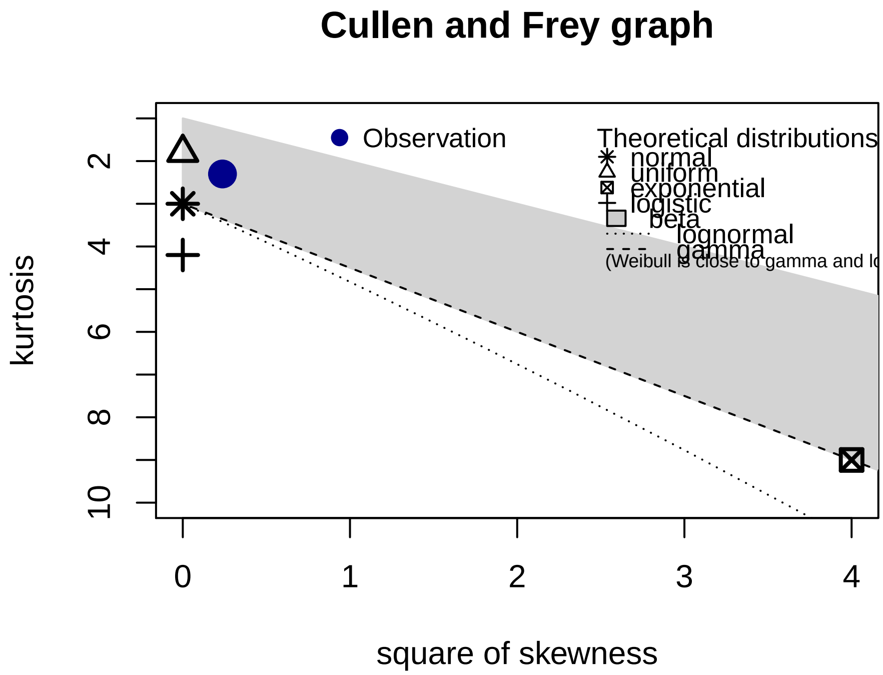
4.1 Alcohol (Continuous)

Visualizing Data



Descriptive Statistics

- Before performing any probabilistic computations, the distribution must be determined.



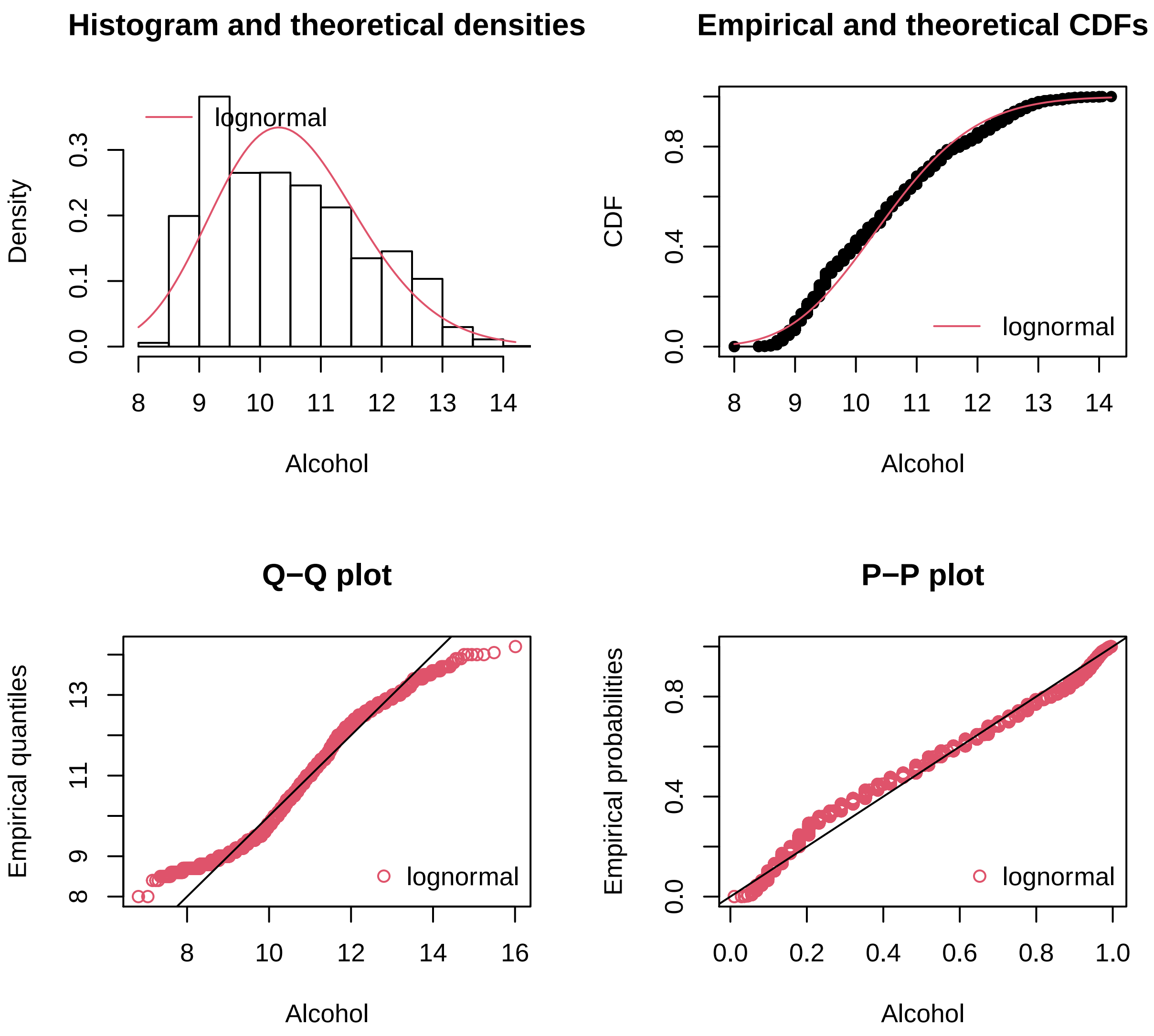
```
## summary statistics
## ------
## min:  8   max:  14.2
## median:  10.4
## mean:  10.51427
## estimated sd:  1.230621
## estimated skewness:  0.487342
## estimated kurtosis:  2.301575
```

Observation:
1. The skewness is between (-0.5,0.5) and hence it is symmetrical.
2. The Culley & Frey graph below shows that this plot is close to normal,log-normal and uniform distribution as the point observation lies near to these distributions.

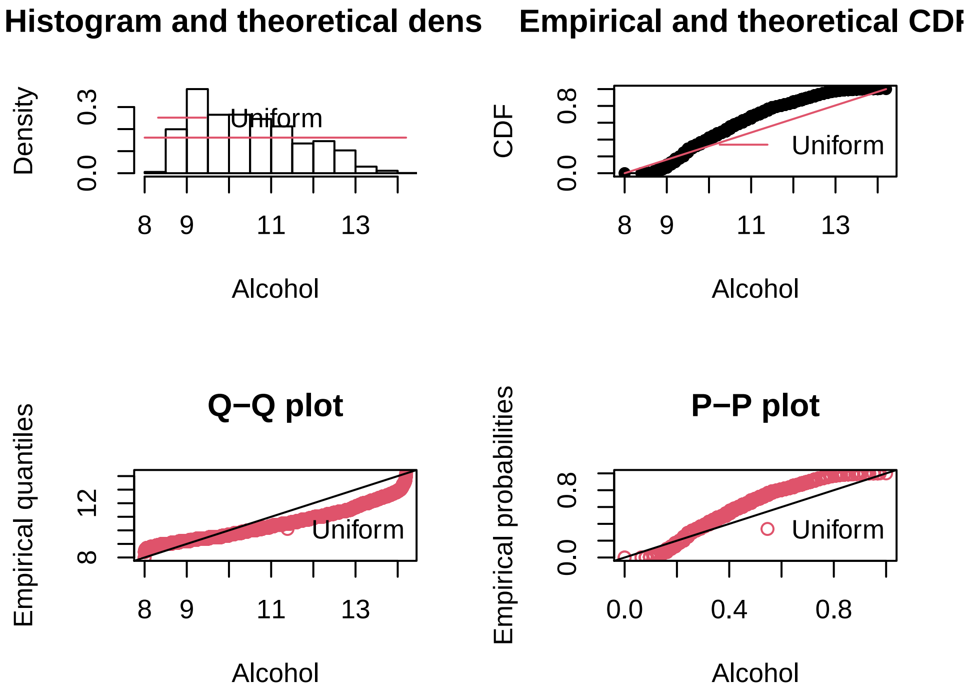
Fit

We observe that,
1. The best of fit analysis of the three distribution types below shows that log-normal distribution is the best fit.
2. Log-normal has the highest log-likelihood value, the lowest AIC and BIC values, and its theoretical P-P and Q-Q plots best match the empirical plots.
3. Hence, we do the goodness of fit test for the log-normal distribution.

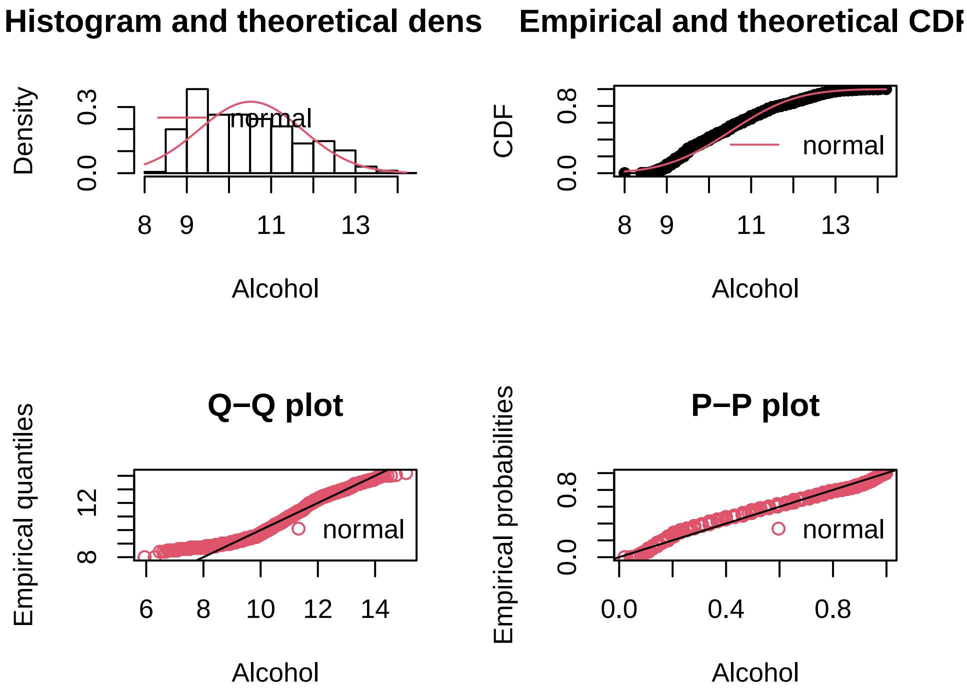
Goodness-of-Fit plot for Log-normal Distribution



Goodness-of-Fit plot for Uniform Distribution



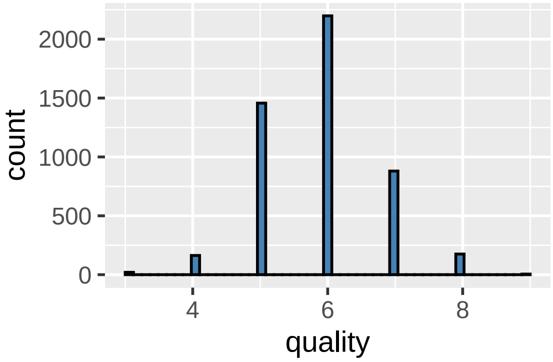
Goodness-of-Fit plot for Normal distribution



Observation: The graph shows that the log-normal distribution fits better for the continuous alcohol data.

4.2 Quality (Discrete)

Visualizing Data



Fit

```
## Chi-squared statistic:  5156.906 5156.904 22484.93
## Degree of freedom of the Chi-squared distribution:  3 4 4
## Chi-squared p-value:  0 0 0
## Chi-squared table:
##      obscounts theo 1-mle-nbinom theo 2-mle-pois theo 3-mle-geom
## <= 4       183         1477.8472       1477.8787       2665.1941
## <= 5      1457          802.0572        802.0628        324.6344
## <= 6      2198          785.7428        785.7420        277.4348
## <= 7       880          659.7944        659.7886        237.0977
## <= 8       175          484.7803        484.7722        202.6254
## > 8          5          687.7781        687.7556       1191.0136
##
## Goodness-of-fit criteria
##                                1-mle-nbinom 2-mle-pois 3-mle-geom
## Akaike's Information Criterion     18423.01   18421.01   27938.37
## Bayesian Information Criterion     18436.00   18427.50   27944.87
```

Observation: The Poisson the lowest AIC and BIC values. Hence, we go with the poisson distribution.

5 Inferential Statistics

5.1 Hypothesis Testing

Assumptions:

1. We are a wine manufacturing company and we manufacture wine in batches(one batch contains ~500 barrels of wine).
2. After manufacturing each batch, we could choose ~10 samples, and test if its properties are meeting the industry standards and also if they are in the legally allowed levels.
3. On average, if these samples fail any of the tests we reject the entire batch i.e, discard the batch else we ship it for packaging.
4. For the purpose of this project we choose only 1 sample from each batch and make decisions based on the inferences we draw from it.

5.1.1 Testing for Industry Standards

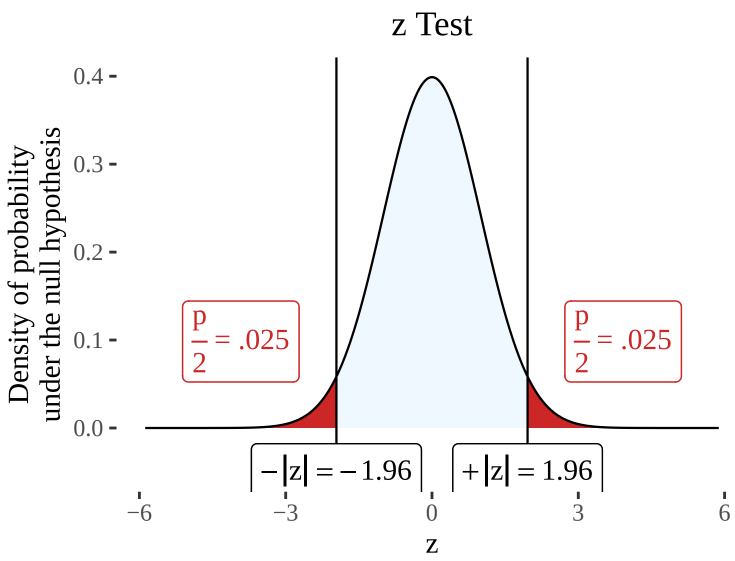
Scenario-1:

1. We manufactured a batch of wines(1 Batch = ~500 bottles of wine).
2. Now we want to check if we successfully manufactured dry wine or not.
3. For testing, we take a sample of 100 wines from the batch and check if their mean residual sugar content is equal to 5.5 g/dm3 or not
4. Why 5.5 g/dm? For Dry wines, residual sugar content should be in the range of [1 gm/dm3 - 10 gm/dm3]. Avg. is (1+10)/2 = 5.5 gm/dm3. We want to test with a confidence level of 95%.

Hypothesis

$H_0 : \mu = 5.5\ g/dm^3$
$H_1 : \mu \neq 5.5\ g/dm^3$

```
plotztest(1.96)
```



We perform **one-sample two-tailed z-test (known variance)** and observe that z_{calc} = 1.113 and critical values: [$z_{-\alpha/2}$, $z_{\alpha/2}$] = [-1.96,1.96]

Conclusion

1. As z_{calc} does not lie in rejection region and also 2*p-value(= 0.266) > alpha(=0.05), we fail to reject H_0.
2. With 95% confidence our sample data shows that our batch of barrels contains dry wine.

5.1.2 Testing for Legal Limits

Scenario-2:

1. We want to check if all of our batches contain volatile acids with in the legal limits or not. The US legal limit is 1.11 g/dm3.
2. For testing, we take a sample of 50 wines from each of the batches.(Stratified Random Sampling). We get a total of 500 wines as sample.

3. We check if their mean volatile acidity is exceeding the legal limit of 1.1 gm/dm3 or not. We want to test with a confidence level of 95%

Hypothesis

- $H_0 : \mu \leq 1.1$ g/L
- $H_1 : \mu > 1.1$ g/L

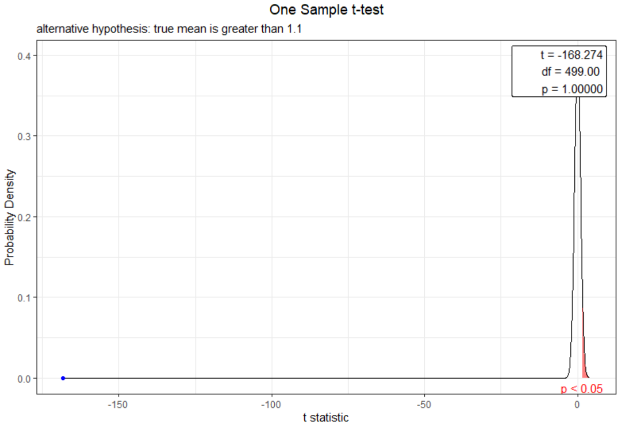


Figure 1: T-test

We perform one-sample right-tailed t-test (unknown variance) and observe that 1. 95% C.I.: (0.2714, inf) and p-value = 0.999~1.
2. Also, the sample statistic lies in the 95% Confidence Interval.

Conclusion

1. As p-value > 0.05 we failed to reject the null hypothesis.
2. We infer that all of our batches contain the amount of volatile acidity within the legal limits.

5.1.3 Testing for Quality Assurance(Health perspective)

Scenario-3:

1. We want to check if we are producing healthy wine across all batches.
2. For a healthy wine, the pH level should be in [2.9-3.9]. We take a target standard deviation of 0.1.
3. For testing, we take a sample of 50 wines from each of the batches(Stratified Random Sampling). We get a total of 500 wines as sample.
4. We check if the variance of pH is equal to 0.01 or not. We want to test with a confidence level of 95%.

Hypothesis

- $H_0 : \sigma^2 = 1.1$
- $H_1 : \sigma^2 \neq 1.1$

We perform one-Sample two-tailed Variance test - Chi-Square test and observe that
1. The 95% Confidence Interval is [0.019 0.0249].
2. The sample statistic lies in the 95% Confidence Interval.

Conclusion

1. As χ^2_{calc} does not lie in the rejection region, we failed to reject the null hypothesis.
2. We infer that we are manufacturing healthy wine in all of our batches.

5.1.4 Testing for R&D

Scenario-4:

1. We want to check if there is any variation in the sulphur content between low quality and high-quality wines.
2. For testing, we take two samples, 100 each from low quality and high quality wines.
3. We check if ratio of their sample variances is equal to 1 or not. We want to test with a confidence level of 95%.

Hypothesis

- $H_0 : \sigma_1^2 = \sigma_2^2$
- $H_1 : \sigma_1^2 \neq \sigma_2^2$

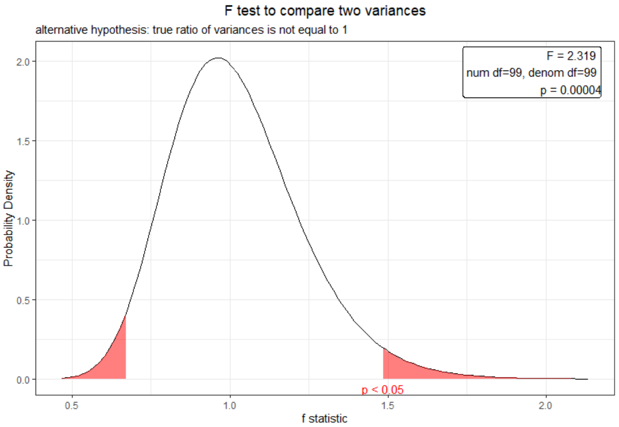


Figure 2: F-test

We perform **one-sample two-tailed Variance test - Chi-Square test** and observe that,
1. The 95% Confidence Interval is [1.211 2.674].
2. $F_{calc} = 1.79$ and p-value = 0.003781.

Conclusion

1. As 2*p-value(=0.006) < alpha(=0.05) we reject the Null Hypothesis, H_0.
2. We infer that there is high variation in the sulphur content of low quality wines.

5.1.5 Testing for Business Development

Scenario-5:

1. We want to target mainly middle-class customers. So, we want to ensure that each batch contains atleast 70% of medium quality wines.
2. For testing, we take a sample of 50 wines from each of the batches.(Stratified Random Sampling). We get a total of 500 wines as sample.

3. We check if the proportion of dry wines is less than 0.7 or not. We want to test with a high confidence level of 99%.

Hypothesis

- $H_0 : p \geq 0.7$
- $H_1 : p < 0.7$

We perform **one-sample left-tailed proportion-test** and observe that,
1. The 99% Confidence Interval is [0, 0.67].
2. $p_{calc} = 0.63$

Conclusion

1. Since, p_{calc}(calculated using z_{calc}) does not lie in the rejection region, we fail to reject H_0.
2. With very high confidence (99%) we conclude that we are ensuring that every batch of barrels contains at least 70% of medium quality wines.

6 References

1. Discover The 5 Basic Wine Characteristics | Wine Folly
2. Understanding Acidity in Wine | Wine Folly
3. Sugar in Wine Chart (Calories and Carbs) | Wine Folly
4. Chloride concentration in red wines: influence of terroir and grape type (scielo.br)
5. The Bottom Line on Sulfites in Wine | Wine Folly
6. Alcohol Content in Wine and Other Drinks (Infographic) | Wine Folly
7. Red Wines From Lightest to Boldest (Chart) | Wine Folly