STAT0032: Introduction to Statistical Data Science Project
Group C12

Authors: 20173658, 20164955, 20081671, 20040074, 909473, 20031465
Every author listed above contributed equally to the project

December 10, 2020

1 Introduction

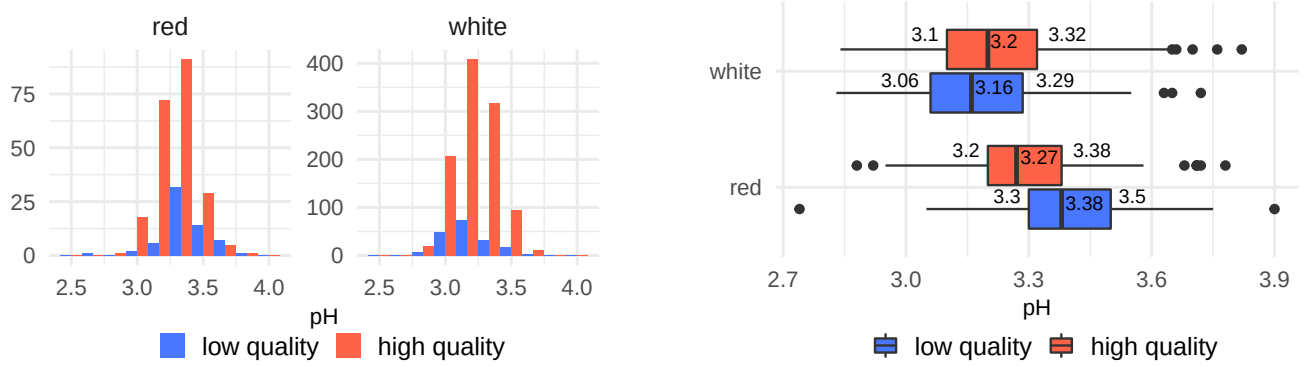
With 2020 forecast revenue over $300bn and 5% projected annual growth to 2025, the global wine industry remains both significant and dynamic despite impacts of the covid-19 pandemic and regional disparities[1]. While large retailers continue to expand aggressively, independent merchants can thrive with their product knowledge and superior customer experience. Novel data-driven approaches can complement traditional skills to help achieve greater sales and profitability. Using the Wine Quality Data Set[2], this study investigates how wine acidity, a prominent but insufficiently understood theme in client feedback, relates to quality. At a high level, acidity gives wine its tart and sour taste, and is measured by pH (*strength*) or total acidity (*amount*). We initially limit our analysis to pH in red wines (the shop specialty), later providing analysis on the differing properties of white wine.

2 Statistical Analysis

Unless otherwise specified, a significance level of 5% ($\alpha = 0.05$) is used throughout. A Bonferroni adjustment was considered, but ultimately rejected to not compromise the power in favour of a reduction in Type I error rate. A key assumption of all tests used is that observations are independent, both within and between samples. Initial exploration uncovered duplicates in the data; however as each wine sample is 'distinct' [1], we assume these simply arise from samples with the same character and the assumption of independence is therefore met.

2.1 The Data

The data comprises two subsets, collected between May 2004 and February 2007, related to samples of white and red vinho verde from the Minho region of Portugal. Wine quality is a median of at least three blind taste test evaluations, scored on a scale from 0 (very bad) to 10 (excellent). Red wine quality in the data ranges between 3 and 8. In line with the client – and to aggregate more data per group – 'low' (quality ≤ 4), 'medium' and 'high' (quality ≥ 7) quality indicators are constructed for the analysis.



(a) **(b)**

Figure 2.1: Exploration of data: (a) Count histogram of the distributions of pH by quality groups for red and white wine (notice the difference in y scale). (b) Boxplot with median, upper and lower quartile values.

2.2 The distribution of pH in red wine

To establish appropriate two-sample tests, we first examine if red wine pH sample distributions are normal and free of outlying data using both statistical and graphical methods. This is because two-sample parametric tests only have an edge in power if both conditions are true; otherwise, non-parametric tests are generally favoured [5]. The use of graphical methods provide additional information to interpret the *degree* of non-normality and presence of outliers. This lowers the risk associated with statistical methods alone, where power is typically sensitive to sample size [5]. Evaluations of pH distribution are applied to the entire red wine data ($N = 1599$), low quality only ($N = 63$) and high quality only ($N = 217$).

2.2.1 Statistical Methods: Pearson's Chi-square and Shapiro-Wilk

Pearson's Chi-square test and the Shapiro-Wilk test (modified for samples of greater than 50 [6]) are used to examine whether the samples come from normal distributions. A summary of the two tests can be found in Table 2.1, where the alternative hypothesis for both is that the distributions are arbitrary, non-normal distributions, $F(\theta)$.

As a general distribution goodness-of-fit test for both discrete and continuous data, Pearson's Chi-square statistic serves as a benchmark for understanding the (non) normality of pH. The statistic relies on arbitrary binning for contin-

[1] For an overview of global wine trends, see https://www.marketresearch.com/Mordor-Intelligence-LLP-v4018/Global-Wine-11402064/
[2] The data was made public following [1], and features 11 chemical characteristics and a measure of quality.

uous data; while the method applied in this study ensures there are at least five observations per bin, there is no guarantee that this is optimal for power and different binning may yield different results [**RolkeFWolfgang2020Acgt**]. The Shapiro-Wilk test, on the other hand, was specifically developed to test continuous data for normality and as such where the test lacks flexibility it demonstrates superior power in comparison to other tests for normality [4].

For the Chi-Square test statistic with k bins, O_i is the observed frequency and E_i the expected frequency for bin i. Under H_0 the test statistic is a χ^2_{k-3} distribution. For the Shapiro-Wilk test statistic, W, both numerator and denominator are estimates of the variance under H_0 therefore W should be close to 1 [2].

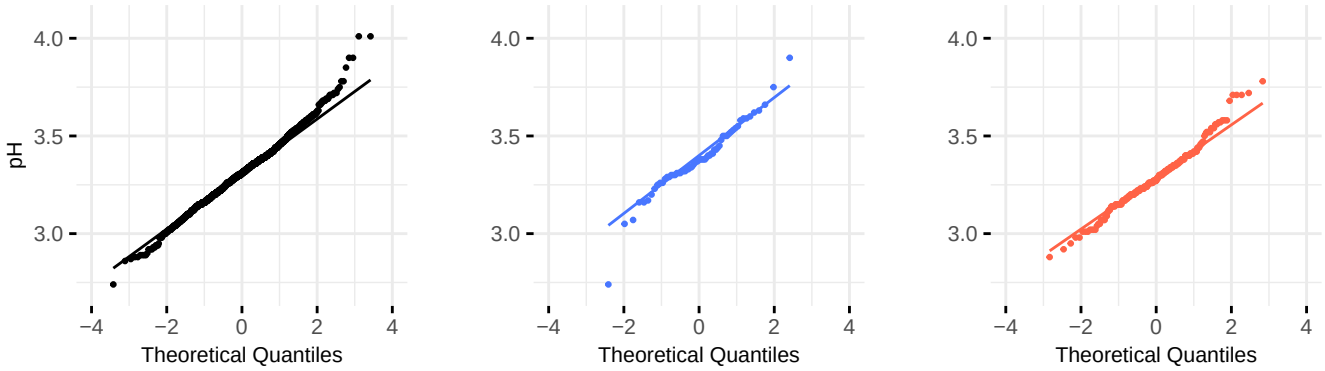
Test	**Hypothesis**	**Statistic**	H_0 **Distribution**	**p-values**		
				All	Low	High
Chi-square	$H_0 : \text{pH} \sim \mathcal{N}(\mu, \sigma^2)$	$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$	χ^2_{k-3}	2.2e-16	5.2e-2	2.2e-2
Shapiro-Wilk	$H_1 : \text{pH} \sim F(\theta)$	$W = \frac{(\sum \alpha_i x_i)^2}{\sum (x_i - \bar{x})^2}$	N/A*	1.7e-6	2.0e-2	1.2e-2

Table 2.1: Summary of normality tests on the three red wine data subsets. *See [7].

Five out of six tests produce extreme test statistics, shown by the p-values in Table 2.1, giving evidence to reject the assumption of normality. The discrepancy between the Chi-square and Shapiro-Wilk on the low quality subset warrants further investigation using graphical methods.

2.2.2 Graphical Method: Normal Q-Q Plots

Quantile-Quantile (Q-Q) plots are used to visually inspect the degree of linear relation between two CDFs [5]. When investigating normality, sample quantiles are plotted against theoretical quantiles from the standard normal.



(a) All red wine **(b)** Low quality only **(c)** High quality only

Figure 2.2: Normal Q-Q plots showing distribution of pH in the red wine data set.

The pH in the subsets are approximately normal except for the tails, where mild overdispersion is suggested (Figures 2.2a and 2.2c). Notably, all three plots contain observations that are potentially outlying and this is supported by Tukey's rule, based on the interquartile range. Figure 2.2b shows at least one significantly low observation, whilst Figure 2.2c has a small group of observations at the upper tail. The outlier values identified are kept for this study as they are within the typical range of pH for wine.

Outliers distort sample means and lead to higher sample standard deviations; thereby reducing the statistical power of parametric tests [5]. Having performed two normality tests and checked the distributions graphically, we are confident that there is enough evidence to reject normality for high, low and all-quality red wines. Consequently, non-parametric tests are also considered later when testing for differences in pH location between low and high quality red wine.

2.3 Two-Sample Tests

2.3.1 Parametric: Welch's t-test

Welch's t-test is a location test on the means of two samples. Other than independence, a significant assumption is that samples are drawn from normal distributions. Unlike the two-sample t-test, it does not assume equality of variances, making it more robust when variances are unknown and sample sizes are unequal, as in the pH data.

Although the outcome of Section 2.2 is to reject the assumption that samples are drawn from normal distributions, the test may be justified by the central limit theorem [5] as sample sizes are > 50. The mean is still considered

an acceptable measure of location as the sample distributions are largely symmetric and identified outlying data from Figure 2.2 are few compared to sample sizes. It is acknowledged that the test may suffer lower power due to outlier inflation of variance estimates and be potentially inaccurate due to the non-normality.

Further details of the test may be found in Table 2.2. Under H_0 (equality of sample means), the statistic has an approximate $t-$distribution with approximate degrees of freedom:

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + s_2^2/n_2)^2/(n_2-1)}$$

where $s_{1,2},\ n_{1,2}$ are the sample standard deviations and sizes respectively.

2.3.2 Non-parametric: Mann-Whitney Test

The Mann-Whitney test is a rank-based method for testing whether two samples come from identical distributions. For two random variables with continuous distributions, it tests H_0 that the two distributions are stochastically equal. In the description of the test statistic U in Table 2.2, R_i is the rank-sum and n_i the number of observations for sample i. As $n_1,\ n_2 > 10$, the distribution of U under H_0 is approximated by the standard normal distribution [5].

The test may be more robust to the potentially non-normal and outlying data present in the pH samples compared to Welch's t-test, since it does not rely on assumptions about the underlying distributions. However, it can be difficult to determine the mechanism behind the rejection of H_0. To infer that it is specifically due to a difference in medians, additional assumption must be met: that the two samples have equal shape and scale. Analysis of the Q-Q plots in Figure 2.2 supports this assumption so the test result is evaluated as such.

2.3.3 Results

Test	Hypothesis	Statistic	H_0 Distribution	p-value
Welch's t-test	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	t_v	1.8e-4
Mann-Whitney	$H_0 : P(X_1 < X_2) = 0.5$ $H_1 : P(X_1 < X_2) \neq 0.5$	$U = \min_{i \in \{1,2\}} \left(R_i - \frac{n_i(n_i+1)}{2} \right)$	$\frac{U - E[U]}{\sqrt{Var(U)}} \sim \mathcal{N}(0,1)$	8.0e-6

Table 2.2: Summary of two-sample tests used to compare locations of low and high quality wine pH.

Both tests produce extreme test statistics that give evidence to reject the null hypothesis that the locations (mean and median) of pH are the same for low and high quality red wine. The association between pH and wine quality is that higher quality wines have a lower pH (i.e. more acidic).

2.4 Additional analysis: bootstrap confidence intervals

The tests from the previous section provide evidence that the central tendency of pH is different for low and high quality red wines. We now aim to complement the findings with related bootstrap confidence intervals [3]. We expand the analysis to the white wine data in order to assess the significance of an observation visible in figure 2.1(b): higher pH appears to be associated with *higher* wine quality in white wine, in contrast to red wine.

We sample with replacement from the data to form $D_b^* = M_{L,b}^* - M_{H,b}^*$ as estimates of $\hat{D} = \hat{M}_L - \hat{M}_H$[4] at each of $B = 100,000$ bootstrap iterations and to build intervals[5]. The normal interval assumes that the distribution of $\hat{D}$ is close to normal; to compare and corroborate the approximate coverage (roughly appropriate given Figure 2.3), we also report the pivotal and percentile intervals, which rely only on the empirical distribution of $D_{b=1\ldots B}^*$.

Type	Red wine 95% $C.I.$	White wine 95% $C.I.$
Normal	[0.06, 0.16]	[-0.07, -0.01]
Pivotal	[0.08, 0.18]	[-0.06, -0.01]
Percentile	[0.04, 0.14]	[-0.07, -0.02]

Table 2.3: Approximate (bootstrap) 95% confidence intervals.

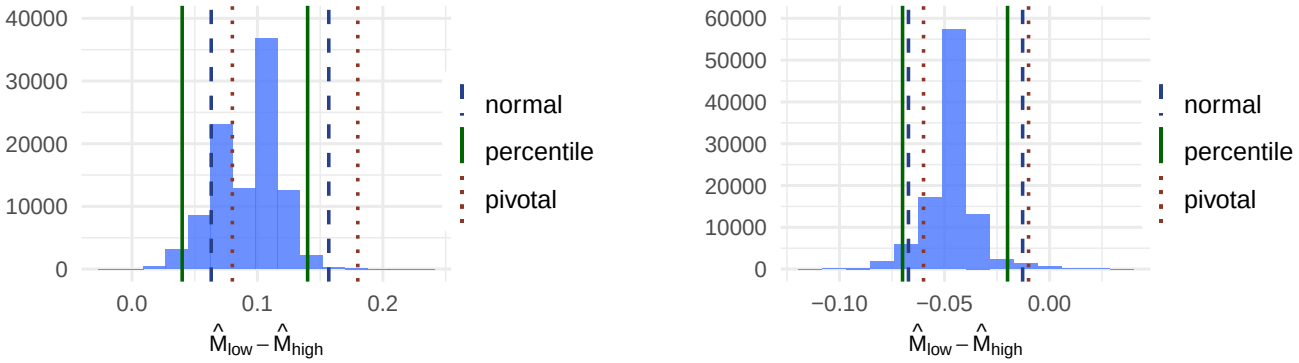
Table 2.3 shows that none of the intervals contain 0 and are generally in close agreement, as visualised alongside the bootstrap median difference estimates in Figure 2.3. For red wine this provides further confirmation that $M_L > M_H$ ($\hat{D} = 0.11, \hat{se}_B = 0.024$). For white wine, the median difference ($\hat{D} = -0.04, \hat{se}_B = 0.014$) is trapped in the

[3] We focus on sample medians in order to mitigate possible adverse effects from the outliers mentioned in Section 2.2.

[4] $\hat{M}_L, \hat{M}_H$ denote the sample median pH in low and high quality wines, respectively.

[5] Following Wasserman 2004, ch. 8., https://doi.org/10.1007/978-0-387-21736-9

widest (hybrid) interval $\hat{D} \in [-0.07, -0.01]$. The relative proximity of 0 in white wine casts doubt on the practical significance of the difference between quality groupings. On the other hand, our findings suggest at a minimum that the relationship between acidity and quality differs in nature across the two wine types. This could explain the conflicting feedback on what clients like in terms of acidity level.



(a) **(b)**

Figure 2.3: Distribution of the median difference estimates for (a) red wine and (b) white wine.

3 Conclusion

Evidence of a positive association between acidity and quality in red wine is found through statistical testing (a *negative* association between pH and quality). Non-parametric tests are used in addition to parametric, as goodness-of-fit methods highlight mild non-normality and presence of outliers (that cannot be removed) in the sample distributions. The analysis is extended through construction of approximate confidence intervals for both red and white wine using the Bootstrap. The opposite relationship is observed for white wine and this may be a factor contributing to mixed feedback in customer acidity preferences.

As these results are derived from an observational study, causation is not assumed because there could be unobserved confounding factors. The data is also specific to wine from a single geographic region; it is not advised to extrapolate these findings to wines from other countries, or even other regions of Portugal. One proposal might be to run a controlled experiment to confirm the associations found here as causation. Such a controlled experiment will allow for a more up-to-date set of data expanded to incorporate wines from other regions or countries, and make use of tasters with a range of wine experience. Experimental studies, e.g. [3], have found that expertise may impact perception; this data, based on ratings solely by wine experts, may not be representative of a more diverse customer base. Finally, in the absence of sales data we assume that wine quality is highly and positively correlated with wine sales so as to draw insight, though as other variables such as price and availability are likely to influence customer decision making there is no guarantee that this assumption is not spurious.

References

[1] Paulo Cortez et al. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decision Support Systems* 47.4 (2009), pp. 547–553. DOI: https://doi.org/10.1016/j.dss.2009.05.016.

[2] A. Ralph Henderson. “Testing experimental data for univariate normality”. In: *Clinica Chimica Acta* 366.1 (2006), pp. 112–129. ISSN: 0009-8981. DOI: https://doi.org/10.1016/j.cca.2005.11.007.

[3] Rebeckah Koone et al. “The role of acidity, sweetness, tannin and consumer knowledge on wine and food match perceptions”. In: *Journal of Wine Research* 25.3 (2014), pp. 158–174. DOI: 10.1080/09571264.2014.899491.

[4] Nornadiah Mohd Razali and Bee Yap. “Power comparisons of some selected normality tests”. In: *Statistics Faculty of Computer and Mathematical Sciences* (July 2010), pp. 126–138.

[5] John A. Rice. *Mathematical statistics and data analysis / John A. Rice.* eng. 3rd ed., international ed. Belmont, Calif.: Thomson/Brooks Cole, 2007. ISBN: 9780495118688.

[6] Patrick Royston. “Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 44.4 (1995), pp. 547–551. ISSN: 00359254, 14679876. DOI: https://doi.org/10.2307/2986146.

[7] S. S. Shapiro and M. B. Wilk. “An Analysis of Variance Test for Normality (Complete Samples)”. In: *Biometrika* 52.3/4 (1965), pp. 591–611. ISSN: 00063444. DOI: https://doi.org/10.2307/2333709.