University of Bern
CAS Applied Data Science
2019/2020

Improvement of Wine Testing Procedures by Means of Statistical Analysis

Simon Büchi
Michael Horn

Storyline

- we are a testing laboratory responsible for the correct declaration of single-varietal wines which obtain state-approved labels
- so far, we are measuring 13 variables in order to distinguish between 3 wine varieties provided by local winegrowers
- for efficiency reasons, we want to reduce the default amount of measured variables as much as possible, without reducing testing accuracy
- our approach: we compare the declaration of batches from winegrowers with reference samples of known composition

M&M

- Data used: Data set «Wine» from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Wine) containing 178 observations
- Procedure:
  - split data set into 3 reference data sets (44, 51, 43 observations for variety 1, 2, 3) and 2 trial data sets (20 observations each)
  - statistical tests for Gaussian distribution, equal variance and sample differentiation

Distribution Analysis

	Gaussian distribution		Equal variance	
	D'Agostino-Pearson test Shapiro-Wilk test		Levene test Bartlett test	
measured variable	transformation		transformation	
	w/o	log	w/o	log
Alcohol	x	x	x	x
Malic acid				
Ash		x		
Alcalinity of ash		x		
Magnesium				
Total phenols		x		
Flavanoids		x		
Nonflavanoid phenols				
Proanthocyanins				
Color intensity				
Hue		x		
OD280/OD315 *	x			
Proline		x		

* of diluted wines

Legend

- statistical tests were performed on the reference data set of each of the 3 wine varieties
- x: all tests fail to reject H0 at a significance level of 5 % (p-values > 0.05)
- H0: variables follow a Gaussian distribution / variances are equal across the 3 reference data sets

Identification of Ideal Variable for Declaration Testing

measured variable	Kolmogorov-Smirnov test	Mann-Whitney U test	log transformed variable	unpaired t test **
Alcohol	x	x	Alcohol	x ***
Malic acid				
Ash			Ash	
Alcalinity of ash			Alcalinity of ash	
Magnesium				
Total phenols	x	x	Total phenols	x
Flavanoids	x	x	Flavanoids	x
Nonflavanoid phenols				
Proanthocyanins		x		
Color intensity	x	x		
Hue			Hue	
OD280/OD315 *		x		x
Proline	x	x	Proline	x

* of diluted wines ** unequal variances *** for both measured and transf. variable

Alcohol: concentration can be determined in a cheap and simple way

ALCOHOL IS THE IDEAL VARIABLE FOR DECLARATION TESTING

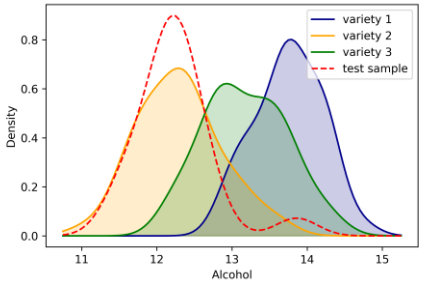
Legend

- statistical tests were performed on the reference data sets comparing the 3 varieties to each other
- x: all tests reject H0 at a significance level of 0.1 % (p-values < 0.001)
- H0: the varieties compared have the same distribution (K-S test, U test) / the same mean (unpaired t-test)

Proof of Principle Testing

- a wine batch of known variety 2 is used for proof of principle testing
- the distribution analysis reveals no Gaussian distribution, therefore the nonparametric tests are used

	Wine batch of known variety 2 compared with reference sample:		
	variety 1	variety 2	variety 3
Kolmogorov-Smirnov test	x		x
Mann-Whitney U test	x		x
	Variable: Alcohol		
	x: p-value < 0.05		



- the outcome is as expected: the variable alcohol is sufficient for declaration testing

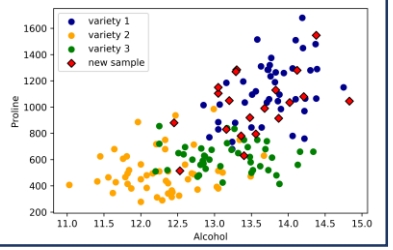
Fraud Detection

- new wine batch with declaration «Variety 1» passes in
- alcohol concentration is measured and compared to reference samples (varieties 1, 2 and 3)
- result is suspect: differentiation between new wine batch and reference sample of variety 3 is not possible

	Wine batch (declaration: variety 1) compared with reference sample:			Wine batch (declaration: variety 1) compared with reference sample:		
	variety 1	variety 2	variety 3	variety 1	variety 2	variety 3
Kolmogorov-Smirnov test		x			x	
Mann-Whitney U test		x	x		x	
unpaired t-test (unequal var.)		x	x		x	
	Variable: Alcohol			Variable: Alcohol		
	x: p-value < 0.05			x: p-value < 0.01		

- further investigations involving the collection and analysis of additional variables prove declaration fraud:

the new sample is a mixture of wines from varieties 1 and 3



Conclusion:

The analysis workflow can be improved by reducing the variables required for testing to only one variable, Alcohol.

This variable is sufficient to differentiate between the 3 wine varieties and to even serve for fraud detection.