

MASTER IN DATA SCIENCE FOR MANAGEMENT

SUBJECT : ANALYSIS OF QUESTIONNAIRE

PROF. : G.COMMIS

**ETNA WINERY SURVEY**



Students:

**GIANLUIGI MAZZAGLIA**

**ANAHITA ESFANDIARYFARD**

**BIAGIO INCARDONA**

**THAMIRES DE SOUZA OLIVEIRA**

Academic year 2021-2022

**Table of Contents**

[INTRODUCTION 2](#_Toc93530119)

[1. Questionnaire Structure 2](#_Toc93530120)

[1.1. Dataset 2](#_Toc93530121)

[2. Univariate Analysis 2](#_Toc93530122)

[4.2 IRT- Item Response Theory 2](#_Toc93530123)

[4.2.1 Dataset level IRT 2](#_Toc93530124)

## INTRODUCTION

The present study, with reference to the wine, was carried out between December and January 2020/21. The purpose of the survey was to discover winery market trends and customer preferences, focusing on Etna wines. All data have been collected anonymously in line with the privacy policy. Our project was done using the statistical software SAS, in particular after a firstly description of the Questionnaire Structure we did the Univariate Analysis in order to understand the distribution of the answer. Then we handled the missing values and we performed Multivariate Analysis, in particular we focus on Item Response Theory, Factor Analysis and Latent Class Analysis.

The Questionnaire was organized in four sections, the first of which was dedicated to the wine preferences, expertise and consumers’experience ( drinking frequency, type of drinks / wine, winery experience). The second part is referred to the buying experience ( buying frequency, consumed quantities, consumption time, factors that influence the purchasing moment). The third part of the questionnaire allowed us to collect information about the knowledge an consumption of Etna DOC (where the respondents will give an answer just in case they heard about Etna wine). The fourth section allowed us to obtain information on the socio-economic characteristics of the respondent (age, gender, level of education, occupation).

# Questionnaire Structure

The Questionnaire is structured in 23 questions, that contains in total 42 elements.

*Section 1: Wine preferences, expertise and experience*

Question 1 : How much do you like the following drinks?

The questions contains 4 items and asks the respondent how much do they like drinks like wine, beer, soft drinks and cocktails. Every item is placed on a semantic scale from 1 (not at all) to 4 (a lot). This question can be useful to understand the drinking preference of the respondent.

Question 2: How much do you like the following kinds of wine?

The question contains 5 items from 1 to 4 adding the possibility to answer “Never tasted”. As before every item is placed on a semantic scale from 1 (not at all) to 4 (a lot). This question allow us to understand which kind of wine the respondent prefers within red, white, rose’, sparkling and sweet.

Question 3 : Have you ever tried a wine tasting experience?

This question asks to the respondent if he never tried a wine tasting experience, it’a a categorical variable with binary choice between yes or not. Answer this question can help us with the analysis in order to categoryze the typology of respondent, likely a person who is interested in things like this can have an high at least curiosity to improve his knowledge of wine.

Question 4: Have you ever visited a winery?

Also this question it’s a categorical variable with binary choice between yes or not. Like the previous question, it can be useful to understand if the respondent is used to going in a winery.

Question 5: Have you ever attended an in-depth wine course?

This question asks to the respondent if he never attended a wine course, it’s a categorical variable with binary choice yes or not. This allow us to better understand the respondent in terms of technical expertise.

Question 6: Which is your level of knowledge around wines?

It’s a closed item that asks about the level of knowledge around wines of the respondent, in particular it is an ordinal scale ordered from “None” (no knowledge) to “High” (professional knowledge level). This answer can also useful to understand if matches with the previous questions, a respondent who has followed a depth course it’s supposed to have an high level of knowledge around wines.

Section 2: Wine buying Experience

Question 7: How often do you buy wine on average in a month?

The question asks the respondent how often he buy wine on average in a month, it’s an ordinal variable and only one answer is allowed. The utility of this question can be to better understand if the respondent is a frequently user of wine.

Question 8: How many bottles of wine do you buy on average per month?

The question can be similar to the previous one, It’s an ordinal variable and only one answer is allowed. The difference to the previous question is to understand the quantity of bottles that the respondent buy monthly (he can choose in a range from 1-3 bottles, 4-6, to more than 12 bottles) and it can be useful to highlight if he is a consumer during meals or maybe he just want to keep some bottles to share with the others.

Question 9: How often do you buy wine in the following store?

This question contains 3 items and each of them is placed on a semantic scale from 1 (never) to 4 (every time) in order to understand if the respondent used to buy wine in a supermarket, in a winery or online, so we can better understand the typology of consumer.

Question 10: How relevant are the following features when you buy a wine?

This question contains 8 items and it’s asked to the respondent to answer a list of features that he consider important during the buying transaction. Every item is placed on a semantic scale from 1 (not at all relevant) to 4 (extremely relevant) and we find items like grape origin, grape variety, budget friendly, brand awarness, vintage, label detailed, attractive packaging, eventual promotions. This question could be extremely important to understand which of the following features can move on our respondent when he buy a wine and so we can realize the kind of person we are referring to.

Question 10: How much do you spend on a bottle of wine on average?

It’s a closed item that can be placed in an ordinal scale and ask to the respondent how much he is spend on average for a bottle of wine. The answer can be just one and this information can be useful to understand if an higher quality of wine is important for the respondent and how he can afford to spend on it.

Question 11: During the pandemic, did the frequency with which you buy wine change?

The question contains 1 item and asks to the respondent if during the pandemic the frequency with which you buy wine is changed. Every item is placed on a semantic scale with 5 points and a middle point 0. This question can be useful to understand if the pandemic period has changed the habits of the consumer.

Question 12: For what reason have you bought wine in the last 3 months?

The question asks to the respondent the reason why he bought wine in the last 3 months. The question is a multi select Multiple choice question and allows to choice more than one option (within Home consumption, to buy a gift, for a special event, to try a new wine) and also allow using an opened answer to insert “Other” if the option doesn’t belong to the alternative. The utility of this question allow us to define a reason why the respondent used to buy wine.

Section 3: Etna DOC wine

Question 13: Have you ever heard about Etna DOC wine before?

This question is mandatory and according to the answer that the respondent will give, the next question will be different. In particular the question asks if the respondent has never heard about Etna DOC. It is a categorical variable with binary choice between yes or not. This question can be useful to understand if the respondent knows the Etna DOC or not. This item performs a branching and who answers no is skipped to the final section.

Question 14: Have you ever bought Etna wine?

It is a closed item and it asks the respondent if he never bought Etna wine. It’s a categorical variable with 3 choice (Yes, No, I don’t know) and it is useful to understand if the respondent has never testef Etna wine.

Question 15: How much do you like Etna wines more than other wines?

This question asks to the respondent how much he likes Etna wine more than other wines. Every item is placed on a semantic scale from 1 (not at all) to 4 (a lot). This question allow us to understand if the respondent appreciates the taste of Etna wine.

Question 16: How much do you agree with the following statements about Etna wine?

This question contains 4 closed items and asks to the respondent how much he agrees with certain statements. Every item is placed on a Likert scale with 4 modes from 1 (not at all in agreement) to 4 (completely in agreement). The utility of this question is to understand respondents’ opinion about Etna wine.

Question 17: How likely are you to recommend Etna wine to your family and friends?

This question asks to the respondent how likely they would recommend Etna wine to family and friends. It is on a Likert scale with 4 modes from 1 (not likely at all) to 4 (extremely likely). This question is useful to understand if respondents consider Etna wine as one of the top quality wine and they also provide to recommend it to closers.

Section 4: Socio – demographic data

Question 18: Please enter your gender.

This is a closed item and the question asks to the respondent his gender. It is a categorical variable with 3 choices (man, woman, prefer not to say).

Question 19: Please enter your age:

This question is an open item on an interval scale and the answer is required. It asks the respondent his/her age and give us a discrete variable in a set that is finite as a lower bound, so the respondent must have an age not less than 18, while the upper bound is infinite.

Question 20: Please enter your level of education.

This question allow us to understand the level of education of the respondent. We have here an ordinal scale (from primary and Medium school to University degree). This question can be useful to understand if could be a correlation between people who have an higher level of education and the knowledge of Etna wine and if in general the choices during shopping time can be different.

Question 21: Where are you from?

This question is a closed item. It asks the respondent the place where he/she comes from. It is a categorical variable with multiple modes. The respondent can be Sicilian, Italian, from an EU country or not EU.

Question 22: What is your occupation?

It is a closed item and asks to the respondent what is her/his occupation. It is a categorical variable with multiple modes. If the occupation is not in the list the respondent can also add his own job applying an open item in the space next to the voice “Other”.

# Dataset

# Univariate Analysis

# Missing values Handling

As we observe from the univariate analysis the dataset contains missing values. Which is due to branching of “buying experience” question and hearing about “etna doc” wine question.

# 4.2 IRT- Item Response Theory

Item response theory (IRT) is a paradigm for the analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other factors. IRT is a theory of testing based on the relationship between individuals’ performances on a test item and the test takers’ levels of performance on an overall measure of the ability that item was designed to measure. IRT is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters. The first parameter affecting the item reponses of an item is the respondent’s level on the latent train that item is trying to evaluate, other factors can be the difficulty level (how much of skills/satisfaction is needed to respond in a correct/positive way), the discrimination power (how good is the item in discriminating between individuals with different levels of the latent trait) and the easiness of guessing of a certain item.

Item Response Theory is based on 4 main assumptions:

1. Monotonicity: if the trait level increases, the probability of a correct/positive reponse increases as well
2. Unidimensionality: there is one dominant latent trait that is measured and this trait is the driving force for the responses observed for each item
3. Local Independence: the responses for an item don’t affect the responses for other items

There are 3 main steps in applying an IRT analysis: model selection, estimation of the parameters and estimation of the abilities of the respondent.

Initially IRT will be applied to the whole dataset in order to have an overall look on the items’ difficulty, then it will be applied onto the subgrops of items extrapolated from the original questionnaire.

## 4.2.1 Dataset level IRT

In this first step, IRT will be applied to all the items together, except for the socio-demographic items (GENDER, AGE, EDUCATION, LOCATION, JOB) since for these items doesn’t make any sense to try studying their difficulties (in this context).

In the following SAS procedure, all the default options will be used. The Marginal Maximum Likelihood estimation method will be used to estimate the parameters of the Graded Response Model (which, for binary items, became the two Parameter Logistic Model). This model expresses the probability of responding in a certain way as a function of the properties of both the item and the individual’s latent trait.



Let’s start by seeing if there is only one dominant latent trait that drives the responses observed for each factor. To do that we can use the following scree plot of the eigenvalues: Chart, line chart

Description automatically generated

As we can see from the images above the first eigen value is much higher than the others and it explains almost the 25% of the total variabilty, hence the model hypothesis is satisfied and so a unidimensional model is reasonable in this example.