Università
Ca'Foscari
Venezia

Corso di Laurea magistrale in Marketing e Comunicazione (Ordinamento ex D.M. 270/2004)

Tesi di Laurea

L'INFLUENZA DEL DESIGN SULLE SCELTE D'ACQUISTO DEL CONSUMATORE DI PROSECCO

Relatore:

Ch. Prof.ssa Isabella Procidano

Correlatore:

Ch. Prof.ssa Christine Mauracher

Laureanda:

Chiara Gallucci

Matricola 355308

Anno Accademico: 2016/2017

Ubi Consistam

INDICE

INTRODUZIONE

Il presente lavoro di tesi si propone di analizzare l'importanza del design durante il processo di scelta dei consumatori, attraverso l'utilizzo di un questionario e di metodologie delle ricerche di mercato, e quindi di costruire un modello che sia in grado di spiegare quali variabili influiscano maggiormente sulle scelte dei consumatori.

Le ricerche di marketing, sono diventate negli ultimi anni, dei processi sempre più importanti nella gestione delle aziende. Alti livelli di competizione, di innovazioni tecnologiche e le richieste sempre più imprevedibili dei consumatori hanno generato un sempre più alto livello di difficoltà delle scelte all'interno dell'ambiente in cui le aziende operano. Tale situazione è aggravata inoltre dalla sempre crescente distanza fisica tra consumatori e produttori, ciò porta l'azienda ad avere difficoltà nel comprendere i bisogni degli stessi. Da qui l'esigenza di uno strumento che fosse in grado di limitare l'incertezza e di conseguenza il rischio in cui si svolge l'attività imprenditoriale, considerata la difficoltà di poter predeterminare gli effetti di ogni decisione presa dal management.

Le ricerche di marketing sono il metodo più efficacie per diminuire tale distanza, in quanto da un lato permettono di identificare problemi e opportunità e dall'altro forniscono la soluzione al problema identificato; vengono perciò usati per la segmentazione, il posizionamento e il lancio di un nuovo prodotto (come nel caso che andremo ad analizzare successivamente).

"Le ricerche di marketing consistono nella sistematica progettazione, raccolta, analisi, e presentazione dei dati e delle informazioni rilevanti per alcune strategie di marketing a cui le aziende devono far fronte".

Da questa visione di Philip Kotler (Marketing Management, 2000, p133) si comprende come alla ricerche di merketing sia collegata l'indagine statistica, applicata seguendo il metodo scientifico.

Il processo di ricerca si compone di 5 fasi: la definizione del problema, la scelta dei metodi di raccolta dei dati, raccolta dei dati e analisi.

L'elaborato nasce dal problema di un'azienda vinicola veneta di esaminare l'importanza del design durante il processo di acquisto e di seguito di analizzare il grado di preferenza della bottiglia di prosecco attualmente sul mercato rispetto ai competitors. Nello specifico le aree di incertezza riguardano:

- La relazione tra design e scelta d'acquisto
- Il posizionamento dell'azienda rispetto ai competitors
- L'opinione dei consumatori in merito alla bottiglia attualmente in commercio
- Le caratteristiche estetiche rilevanti durante il processo di scelta
- L'individuazione di un'eventuale design sostitutivo
- Pricing

Tali quesiti vengono affrontati attraverso la costruzione di un questionario costruito ad hoc e somministrato ad un campione di 190 unità statistiche.

L'elaborato si suddivide in quattro capitoli, uno dei quali teorico e tre l'indagine statistica articolata attraverso l'elaborazione dei dati.

Il **primo capitolo** tratta la teoria dei modelli di scelta i quali permettono di conoscere le variabili che influenzano la domanda di mercato e sono interessanti per comprendere i processi che conducono alle preferenze dei consumatori. Vengono delineate successivamente le caratteristiche della regressione logistica multinomiale, (modello che ha permesso la ricerca delle variabili durante il processo di scelta per l'analisi che verrà effettuata nel capitolo quarto) e in aggiunta delinea i procedimenti da effettuare per costruire un modello e interpretare i coefficienti da questo generati.

Il **secondo capitolo** ha per oggetto l' analisi descrittiva riguardante lo studio di dati assemblati e raggruppati per esaminare le caratteristiche del soggetto in esame e per

determinare le varie tipologie di relazioni che intercorrono tra le variabili ad esso correlate. L'obiettivo è sintetizzare dati in forme comprensibili e chiare, di identificare i *casual factors* e sottolineare i fenomeni complessi, ci aiuta a tracciare interferenze attendibili dai dati osservati e infine permette di creare stime dai risultati ottenuti sul campione. Ciò permette di avere un quadro su coloro che sono gli intervistati, sulle loro abitudini e preferenze.

Il **terzo capitolo** affronta la tematica dell'analisi bivariata, cercando di analizzare ed evidenziare le relazioni più significative tra le varie domande presenti nel questionario e cercando di trarne informazioni più ricche ed interessanti. Questo tipo di analisi, partendo dall'analisi descrittiva, studia una coppia di variabili contemporaneamente per ricercare dipendenze e correlazioni in modo tale da comprendere il legame che le lega.

Il **quarto capitolo** si focalizza su una tipologia di analisi più complessa: l'analisi multivariata. Lo scopo è quello di evidenziare le relazioni più significative tra le variabili e comprendere come e quanto le variabili *predictors* influiscano sulla variabile *response* (variabile dipendente) in modo tale da poter delineare un modello in grado di prevedere il comportamento di consumatori di cui non si hanno i dati.

Per questo tipo di studio è stato necessario passare ad uno strumento più completo e adatto all'analisi statistica: Gretl. Per prima cosa il database è stato adattato in modo tale da poter essere inserito all'interno del programma, successivamente trasferito all'interno di Gretl; secondariamente è stata individuata la variabile dipendente attorno alla quale costruire il modello, ovvero la scelta della bottiglia cosi da comprendere quali siano i fattori che influiscono sulle preferenze dei consumatori. È stato poi costruito il modello attorno alla variabile Y e sono stati interpretati i vari coefficienti.

1. CAPITOLO I
I MODELLI DI SCELTA

Comprendere e misurare gli effetti delle scelte dei consumatori è uno dei più impegnativi aspetti delle ricerche di mercato. Il marketing esiste per rispondere a domande del tipo "I consumatori compreranno più sapone del marchio X se viene aumentato il contenuto di profumo?". L'econometria tradizionale non fornisce risposte a tali domande, le ricerche di mercato si sono quindi rivolte alla psicologia e ai sondaggi per rispondere a questo tipo di domande. La realtà ha costretto l'economia ad allargare la sua prospettiva in termini di analisi della domanda, gli economisti hanno iniziato ad usare indagini campionarie e ad applicare i modelli usati per l'analisi dei dati.

Le scelte si presentano in diversi modi e forme e riflettono le abitudini o le reazioni spontanee del consumatore alle variabili di marketing. Le persone fanno scelte continuamente, alcune di queste sono di interesse economico, governativo ecc. Le informazioni riguardanti le scelte possono essere raccolte da strumenti quali gli scanner dei supermercati, tuttavia tale metodo non riporta nessun tipo di dati su prodotti che non esistono; per questo possono essere d'aiuto gli approcci sperimentali (modelli di scelta).

I modelli di scelta sono largamente utilizzati nell'analisi di diverse aree oltre al marketing come i trasporti, l'ambiente, il benessere pubblico ecc. Le scelte nell'ambito del marketing differiscono rispetto a quelle in altri campi, in quanto il contesto di scelta è spesso molto complesso e, inoltre, perché i ricercatori desiderano conoscere le variabili che influenzano la domanda di mercato e i processi che conducono alle preferenze. Identificare le variabili che influenzano le scelte è impegnativo a causa anche dell'uso eterogeneo operato dai consumatori. I consumatori codificano, elaborano e reagiscono agli stimoli di marketing, ciò offre numerose opportunità

all'identificazione di variabili rilevanti e di mezzi attraverso cui combinare tali variabili per formare aspetti di considerazione, valutazione e scelta.

Il ruolo del marketing è quello di supportare il management durante il processo decisionale e nello specifico durante la decisione degli elementi da immettere sul mercato; per tale ragione il marketing necessita modelli sempre più accurati e che permettano la ricerca di variabili rilevanti durante il processo di scelta. Negli ultimi venti anni i modelli di scelta hanno subito numerose trasformazioni.

Esistono cinque fasi storiche della costruzione dei modelli:

- La prima era è definita della ***trasposizione dei metodi OR/MS[1] al marketing.*** Tali modelli erano però poco realistici e l'uso nel marketing fu limitato.
- La seconda era è caratterizzata da ***modelli adattati per rispondere a problemi di marketing.*** La mancanza di realismo sembrava essere la ragione principale del mancato utilizzo dei modelli, i quali fornivano una migliore rappresentazione della realtà ma mancavano di semplicità e usabilità. In quest'era (anni 60') i modelli descrittivi di scelta e i modelli econometrici catturarono l'attenzione dei ricercatori, tuttavia, non erano in grado di rappresentare la realtà perché si concentravano solo su alcuni aspetti, tralasciandone altri rilevanti.
- La terza era inizia nei primi anni 70'. In questo periodo l'attenzione si concentra su **modelli che forniscono una buona rappresentazione della realtà**, ma che allo stesso tempo sono facili da usare. Il focus si sposta quindi, dai problemi isolati, all'implementazione e all'implementabilità. Little(1970) esaminò il problema del mancato uso dei modelli e suggerì alcune soluzioni. Egli afferma che un manager ha bisogno di un modello di scelta basato su un set di procedure (decision calculus) attraverso cui esaminare dati e giudizi utili a generare delle decisioni. Little propone inoltre dei criteri da soddisfare per la costruzione di un buon modello di calcolo decisionale:
    - Semplicità

[1] Operation Research e Management Science: metodi usati nel secondo dopoguerra attraverso algoritmi e processi applicabili alla produzione e alla logistica. Includono teoria dei giochi, simulazioni, modelli stocastici di comportamento di scelta dei consumatori e dynamic modeling.

- Robustezza
- Facilità di controllo
- Facilità di comunicazione
- Adattabilità
- Completezza

Viene successivamente aggiunto un settimo criterio: l'evoluzione; quest'aggiunta è basata sul concetto che un modello dovrebbe svilupparsi attraverso una struttura semplice alla quale poi aggiungere dettagli.

In quest'era vengono introdotti modelli di marketing strategico e sistemi di supporto alle decisioni di marketing (MDSS), in aggiunta alcuni ricercatori si focalizzano sullo studio della relazione tra i modelli di marketing e le strategie organizzative.

- La quarta era inizia negli anni 80' ed è un'era in cui ***molti modelli vengono implementati*** a causa della maggiore reperibilità dei dati di marketing. Di particolare importanza è la "scanning revolution" che, combinata con la maggiore reperibilità dei dati, stimola l'applicazione di nuovi metodi. I modelli iniziano a focalizzarsi sul design, sul processo di scelta, sulle strategie competitive ottimali e sulla stima delle reazioni. Tali miglioramenti risultano in modelli che:
    - Soddisfano i criteri di Little
    - Sono parametrizzati su un alto numero di osservazioni
    - Considerano gli errori presenti nei dati

  La maggiore reperibilità dei dati offre anche l'opportunità ai ricercatori di costruire modelli che possono aumentare la conoscenza di marketing, infatti in quest'era vi è un boom dei sistemi computerizzati di supporto alle decisioni.
- La quinta era inizia negli anni 90' ed è caratterizzata ***dall'applicazione dei modelli all'interno delle routine.*** Si passa quindi, dal supporto delle decisioni, all'automazione. In tal modo il manager può dedicarsi a decisioni "non-routinarie" che richiedono creative thinking. Esempi di decisioni di marketing appropriate per l'automazione sono:

- Decisioni di assortimento e di allocazione sugli scaffali dei singoli punti vendita
- Offerte di prodotti personalizzati che includono prezzi e promozioni personalizzate
- Identificazione di uno specifico target al quale inviare mail di sollecitazione
- Programmi di fedeltà
- Creazione di calendari per le promozioni

1.1 IL CONCETTO

"***Un modello di scelta consiste in un set di scelte (choice set) ognuno dei quali a sua volta è formato da due o più opzioni. Ad ogni soggetto viene presentato ogni set di scelte e viene chiesto di scegliere tra una delle opzioni presentate. Il numero di opzioni è chiamato grandezza del set di scelta (choice set size).***"[2]

Un esempio di set di scelte riguarda l'analisi dei mezzi di trasporto con cui un gruppo di lavoratori si reca sul posto di lavoro. Le opzioni sono: auto, bus, bicicletta, camminare o altro. Ogni individuo deve scegliere una tra le opzioni proposte. Tale esempio illustra come negli esperimenti di scelta gli individui siano "forzati" a scegliere una delle opzioni presentate, tale esperimento è chiamato *forced choice experiment*.

In tal caso essere obbligato è semplice, in quanto gli individui considerati sono lavoratori e devono andare a lavoro in ogni caso, in aggiunta, ogni opzione era elencata, in quanto, vi era l'opzione "altro" che includeva tutte le altre alternative.

A volte però un *forced choice experiment* è usato anche in presenza di una lista di opzioni non esaustiva, in modo tale da analizzare come gli individui reagiscono alle differenti caratteristiche delle opzioni presentate. Un esempio è quello di offrire un volo economico con tempi di check in ristretti e un volo su cui ci sono meno restrizioni

[2] The Construction of Optimal Stated Choice Experiments: Theory and Methods; Deborah J. Street,Leonie Burgess

rispetto ai tempi ma che è più costoso; potrebbero esserci scelte intermedie ma non vengono proposte nel choice set.

Durante la costruzione di un choice set è importante evitare di inserire un'opzione che verrebbe scelta da tutti. In questo caso, prezzo basso e tempi di check-in con meno restrizioni, in quanto questo vanificherebbe lo scopo dell'analisi, che è sapere se gli individui danno più importanza al prezzo o alle minori restrizioni sui tempi di check-in. L'opzione che verrebbe scelta da tutti è chiamata opzione dominante (*dominating o dominant option*)[3] ed è necessario costruire un choice set evitandola ed evitando, inoltre, opzioni irrealistiche.

Esistono tuttavia delle situazioni in cui non ha senso forzare gli individui a scegliere, è per questo che i forced choice experiments includono delle opzioni chiamate "non scelte" o "nessuna delle precedenti", ovvero delle non opzioni, in modo tale da coprire tutte le situazioni. Uno scenario simile si verifica quando vi è un'opzione che ha bisogno di comparire in ogni choice set (es. quando si compara un nuovo trattamento con il vecchio), in questo caso si parla di *common base options*. In altri casi, invece, viene descritta un'opzione in modo tale da sapere se gli intervistati userebbero o no quel prodotto/servizio (*binary response experiments*); questi sono gli esperimenti più semplici.

Gli esperimenti vengono descritti da diversi attributi; ogni attributo ha due o più livelli, per esempio, nel caso del check-in vi erano due attributi: il costo e i tempi di check-in. In generale gli attributi devono avere livelli plausibili e che cambino secondo un range rilevante.

[3] Per approfondimenti consultare Huber e Zwerina (1998)

1.2 REGRESSIONE LOGISTICA MLTINOMIALE

La regressione multinomiale è un'analisi predittiva, è usata per spiegare la relazione tra una variabile dipendente nominale e più variabili indipendenti. È un modello a multi-equazioni simile alla regressione lineare multipla.

Il Modello Multinomial Logit, anche detto modello politomo, è un'estensione del modello binomiale; è un metodo di classificazione che generalizza la regressione logistica per problemi multi-classe (con due o più risultati discreti). Viene usato quando la variabile dipendente ha più di due categorie nominali, ovvero non ordinate.

É un modello in cui le probabilità dipendono da un vettore x_i di covarianti associate agli ie-simi individui.

Il modello logit è utile quando si cerca di spiegare le scelte discrete, ovvero le scelte di un'alternativa fra tante, senza un ordine naturale:

$$y_n \in \{1, 2, \ldots J\}. \qquad (1.1)$$

Il Multinomial usa solo variabili che descrivono caratteristiche e non alternative, Es:

•Scelta del tipo di trasporto tra le opzioni: bus, treno, macchina, bici. Queste sono variabili che descrivono il viaggiatore, non vi è nessuna informazione sui metodi di trasporto.

•La scelta di un'auto tra diversi modelli: 5porte, 3porte, sportiva, Suv. Vengono usate solo le informazioni relative al compratore e non sono presenti informazioni sui veicoli.

É il modello più usato in casi di variabili discrete, e serve a:

- Investigare sulla presenza di un'associazione tra le variabili d'interesse
- Misurare la forza di tale associazione
- predirre probabilità di scelta in funzione delle caratteristiche osservate
- calcolare le misure di elasticità

- analizzare e valutare le variazioni del benessere che derivano da cambiamenti esogeni negli attributi delle alternative di scelta o l'insieme di attributi disponibili
- monitorare e analizzare le scelte dei consumatori.

Il termine modello multinomial logit racchiude una varietà di modelli, il modello generale è usato quando la risposta di un'unità individuale è limitata a uno di un numero finito di valori ordinali, mentre il modello multinomial logit è usato quando il "regressore" non varia a seconda delle alternative.
É usato per predire le probabilità di diversi possibili risultati di variabili dipendenti distribuite in categorie, dato un set di variabili indipendenti.
Es. per monitorare la preferenza degli individui di recarsi a lavoro con il treno, con l'autobus, con l'auto o con la bici
Non avrebbe senso usare la regressione lineare con questo tipo di dati.

Nel modello multinomiale un individuo sceglie quindi tra *n* diverse alternative quella che gli permette di raggiungere la massima utilità, il modello binomiale è un caso speciale in cui vi sono *n=2* alternative.

1.2.1 I LOGIT

Il *logit* serve a descrivere una funzione che lega la probabilità di Y alla combinazione delle variabili indipendenti X; non è l'unica funzione in grado di modellare la probabilità di un fenomeno, ma è privilegiata essendo una trasformata del rapporto tra due probabilità complementari (*odd*).

L'idea alla base logit consiste nell'utilizzare una funzione logaritmica per limitare i valori di probabilità a (0,1), ovvero il logaritmo della probabilità di Y=1. Lo scopo dell'analisi è quello di identificare i k-1 log odds per gni categoria.
Il modello Multinomiale viene usato per analizzare la relazione tra una variabile di risposta politomica e un insieme di variabili.

Esiste una variabile per ogni categoria, tranne una, quindi se esistono K categorie avremo K-1 variabili. Ogni variabile ha un valore 1 per la propria categoria e 0 per tutte le altre categorie, c'è una categoria, la categoria di riferimento, che non ha bisogno della variabile, in quanto è definita, essendo zero tutte le altre variabili.

Il Modello Multinomiale, successivamente, stima un modello binario di regressione logistica per ogni variabile, si ottengono così k-1 modelli, ognuno dei quali indica gli effetti del fattore prognostico sulla probabilità di successo in quella categoria, rispetto alla categoria di riferimento. Ogni modello ha la sua intercetta e il suo coefficiente di regressione, il fattore di predizione (predictor) può influenzare ogni categoria in modo differente.

La funzione logit è il logaritmo naturale delle probabilità che Y sia uguale a una delle categorie. Per semplicità matematica, si assume che Y abbia solo due categorie e codice come 0 e 1. È semplicemente una funzione della media della variabile Y che viene usata invece di Y stessa; infatti durante una regressione lineare si studia la relazione e la forza di tale relazione tra due variabili, tale relazione è esplicata dalla formula:

$$Y = \beta_0 + \beta X + \varepsilon \qquad (1.2)$$

In cui la Y rappresenta la variabile dipendente che deve essere prevista, mentre la X rappresenta la variabile che serve affinché la Y possa essere predetta; β_0 anche detta α(intercetta) e β(coefficiente di regressione) sono invece i coefficienti che indicizzano la relazione e che vanno stimati per poter comprendere la forza di tale relazione, mentre ε rappresenta l'errore ed è inserita per dare più correttezza alla formula. Nella regressione multipla la situazione cambia, come anticipato precedentemente, ma il concetto alla base rimane invariato.

Tutto questo significa che quando Y è categorica, usiamo il logit di Y come la risposta nella nostra equazione di regressione invece di Y e la funzione diventa:

$$\boldsymbol{Y = Ln\left(\frac{P}{1-P}\right) = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon} \qquad (1.3)$$

Questa funzione rappresenta la formula centrale dell'analisi multivariata, in cui k è il numero di variabili indipendenti e i β_k(coefficienti di regressione) rappresentano il fatto che ognuna delle X_k (Predictors) fornisce una predizione alla variabile dipendente Y; infine P è definito come la probabilità che Y = 1. Così, per esempio, le X potrebbero essere i fattori di rischio specifici, come l'età, la pressione alta e il livello di colesterolo, e P la probabilità che un paziente sviluppi la malattia di cuore.
Se usassimo Y come variabile di riferimento e provassimo a disegnarne una linea, non sarebbe una corretta rappresentazione della relazione, inoltre, neanche l'uso di P come variabile di riferimento sarebbe scorretto, in quanto la relazione non sarebbe comunque lineare. Esistono però delle funzioni di P che funzionano per reazioni lineari con X, tra cui una di queste è il Logit.
Il Logit viene preferito perché i risultati sono relativamente semplici da interpretare, sebbene le altre opzioni siano altrettanto buone. Una volta usato tale modello è possibile ritrasformare i coefficienti di regressione, stimati in modo da interpretare gli effetti di ogni X.
IL MNL calcola un diversa variabile per ogni scelta, ogni variabile è come un punteggio di valutazione per ogni individuo, per ogni scelta; più è alto il punteggio, più è probabile che l'individuo scelga quell'alternativa.
Nel MNL le variabili indipendenti rimangono invariate, ovvero il valore della variabile è il medesimo per ogni scelta considerata, mentre i coefficienti variano.

A volte le variabili indipendenti variano a seconda della scelta, per esempio, una variabile come la religione è fissa, ma gli effetti di tale variabile variano a seconda della scelta, quindi tale variabile avrà un'importanza diversa a seconda del caso preso in analisi.

1.2.2 IL MODELLO E IL COLLEGAMENTO CON LA FUNZIONE DI UTILITA'

Il Multinomial Logic si differenzia per la determinazione dell'utilità. Si suppone che Y_i rappresenti una scelta discreta tra *J* alternative, che U_{nj} rappresenti il valore di utilità della scelta j-esima dell'individuo i-esimo e che gli individui agiscano in modo razionale per massimizzare la loro utilità.

Di seguito U_{ij} verrà trattata come variabile casuale indipendente in modo tale che:

$$U_{nj} = X_n\beta_j + \varepsilon_{nj} \qquad (1.4)$$

La variabile X_n descrive l'individuo ed è identica per ogni alternativa. Il parametro βj cambia a seconda delle alternative. Le y_n scelte osservate di un individuo n sono:

$$Yn = \begin{cases} 1 \; se \; U_{n1} \geq Uni \;\; per \; ogni \; i \\ 2 \; se \; U_{n2} \geq Uni \;\; per \; ogni \; i \\ \dots \\ J \; se \; U_{nj} \geq U_{nj} \; per \; ogni \; i \end{cases} \qquad (1.5)$$

L'errore segue in modo indipendente e identico la distribuzione del valore

$$F(\varepsilon_{nj}) = \; e^{-e^{-\varepsilon_{nj}}} \qquad (1.6)$$

La probabilità che un individuo n scelga un'alternativa j è

$$P_{nj} = P(Y_n = j|X_n) = \frac{e^{x_n\beta_j}}{\sum_{i=1}^{j} e^{x_n\beta_j}} \qquad (1.7)$$

Un aspetto interessante del multinomial logit è che gli odds-ratio (Pnj/Pni) dipendono da x_n

$$log\left(\frac{P_{nj}}{P_{nj}}\right) = x_n\left(\beta_j - \beta_i\right) \qquad (1.8)$$

1.2.3 IDENTIFICAZIONE

I parametri $\beta_j, j = 1, \dots, J$ non sono definiti: ogni vettore q aggiunto a tutti i vettori $\beta_j = \beta_j + q$ cancella le probabilità :

$$P_{nj} = \frac{e^{x_n(\beta_j+q)}}{\sum_{i=1}^{J} e^{x_n(\beta_j+q)}} = \frac{e^{q} e^{x_n\beta_j}}{e^{q}\sum_{i=1}^{J} e^{x_n\beta_j}} = \frac{e^{x_n\beta_j}}{\sum_{i=1}^{J} e^{x_n\beta_j}} \qquad (1.9)$$

Queste vengono solitamente risolte ponendo $\beta_i = 0$ per un'alternativa di riferimento i. I parametri del MNL sono di difficile interpretazione. Nè il segno, né l'intensità del parametro ha un esplicito significato. Test d'ipotesi devono, tuttavia, essere formulati molto accuratamente in termini di parametri stimati. L'effetto marginale di una variabile indipendente x_k sulla probabilità di scelta per alternative *j*

$$\frac{\partial P(y = j|x)}{\partial_{xk}} = Pj\ (\beta_{jk} - \beta^{-} k) \qquad (1.10)$$

dipende, non soltanto dal parametro β_j, ma anche dalla media di tutte le altre alternative $\beta^{-}\mathrm{k} = \frac{1}{\mathrm{J}}\sum_{\mathrm{j}} \beta_{\mathrm{jk}}$. Ulteriore ipotetica interpretazione della stima del parametro potrebbe essere ottenuta guardando ai log degli odds-ratio:

$$\frac{\partial \log(\mathrm{P_j}/\mathrm{P_i})}{\partial \mathrm{x_k}} = \beta_{\mathrm{jk}} - \beta_{\mathrm{ik}} \qquad (1.11)$$

che si riduce a

$$\frac{\partial \log(\mathrm{P_j}/\mathrm{P_i})}{\partial \mathrm{x_k}} = \beta_{\mathrm{jk}} \qquad (1.12)$$

per un confronto con le categorie di referenza i. Un parametro β_{ik} significa, tuttavia, che la probabilità relativa di scegliere j è direttamente proporzionale alla probabilità di scegliere i. Il MNL può essere inoltre usato per predirre probbilità di scelta per specifici tipi di x_{nj}

$$P\hat{}j \;=\; P\hat{}(y \;=\; j|x) \;= \frac{e^{x\beta_j}}{\sum_{i=1}^{J} e^{x\beta_j}} \qquad (1.13)$$

si può solo analizzare il cambiamento di singole caratteristiche sulla variabile predetta, cosi come tutte le informazioni sulle alternative sono racchiuse negli specifici parametri di stima delle alternative βˆ j . Oltretutto non è possibile simulare l'aggiunta o la sottrazione di una scelta alternativa.

1.2.4 MODELLARE I LOGIT

Il Multinomial logit è una generalizzazione del Logit binario per categorie K. Immaginiamo di voler studiare se a una persona piace il gelato. Secondo il logit binario il log della probabilità di y=1 è funzione lineare della x:

$$log\ (\ P(y=1|\ x)\ /\ P(y=0|\ x) = a+bx \qquad (1.14)$$

a è una costante e *b* è un coefficiente di regressione del vettore che misura il cambiamento dei log della probabilità di y=1 associato al cambiamento di una unità in x. Il log della probabilità varia da -∞ a +∞.

Ora immaginiamo di voler sapere il gusto preferito di gelato di un individuo, poniamo 1=nocciola 2=cioccolato 3=vaniglia (K=3)

Abbiamo quindi k-1 equazioni, ogni equazione monitora le probabilità di una scelta relativa alla categoria di riferimento, solitamente l'ultima, in questo caso vaniglia. In un modello binario avremmo solo un'equazione, mentre in questo caso abbiamo 2 equazioni:

$$log\ (\ P(y=1|x)\ /\ P(y=k|x)\ ) = a_i + b_i x \qquad (1.15)$$

questa equazione esprime le probabilità di una persona che preferisce la nocciola alla vaniglia

$$log\ (\ P(y=k\text{-}1|x)\ /\ P(y=k|x)\ ) = a_{k-1} + b_{k-1} x \qquad (1.16)$$

questa equazione esprime le probabilità di una persona che preferisce il cioccolato alla vaniglia.
Abbiamo k-1 set di coefficienti, il primo set di coefficienti esprime come le x influenzano le probabilità nocciola vs vaniglia, il secondo set di coefficienti esprime come le x influenzano le probabilità cioccolato vs vaniglia. Esponenziando le funzioni avremo:

$$P(Y=1|x)= exp(a_i+b_ix)/(1+exp(a_i+b_ix)+....+exp(a_{k-1}+b_{k-1}x) \quad (1.17)$$

$$P(Y=K-1|x)=exp(a_{k-1}+b_{k-1}x)/(1+exp(a_i+b_ix)+.....+exp(a_{k-1}+b_{k-1}x) \quad (1.18)$$

$$P(Y=K|x)=1-P(Y=1|x)-.....-P(Y=K-1|x) \quad (1.19)$$

Non ha importanza quale categoria scegliamo come riferimento perché si può sempre passare da una formula all'altra, per esempio in questo caso lo scontro mancante tra nocciola e cioccolato può essere ottenuto tramite:

$$log(P(y=1|x) / P(y=k-1|x)) = log(P(y=1|x) / P(y=k|x)) - log (P(y=k-1|x) / P(y=k|x)) \quad (1.20)$$

1.2.5 GLI ODDS RATIO E L'INTERPRETAZIONE DEI RISULTATI

Gli Odds ratio sono degli indicatori simili alle probabilità, ma hanno delle proprietà diverse che li rendono utili all'interno della statistica; sono la misura dell'associazione tra due fattori. Diversamente dalle probabilità, gli odds sono espressi dal rapporto tra le volte in cui un determinato fenomeno si è verificato e le volte in cui non si è verificato. Es: se la probabilità che un evento si verifichi è 0,7 la probabilità che tale evento non si verifichi sarà 0,3. Gli odds che questo evento si verifichi saranno 0,7/0,3= 2,3 mentre, gli odds che l'evento non si verifichi saranno 0,3/0,7 = 0,42.

Come spiegato nel paragrafo precedente, all'interno della regressione logistica gli odds rappresentano l'effetto all'interno dei predicted odds di una unità di cambiamento in X quando le altre variabili sono mantenute costanti.

L'obiettivo è misurare l'effetto di ogni X sulla Y, ma non c'è modo di esprimere attraverso un numero quanto X influenzi Y in termini di probabilità, in quanto l'effetto sulla probabilità varia a seconda del valore di X. Per tale motivo si preferisce l'uso degli odds, in quanto, essendo costanti[4], permettono di comunicare il risultato della ricerca attraverso un numero.

Gli Odds Ratios all'interno della regressione logistica possono essere interpretati come gli effetti apportati dal cambiamento di un'unità X all'interno dei predicted odds ratio tenendo le altre variabili del modello costanti. Altra importante proprietà degli odds ratio è la costanza, infatti rimangono costanti nonostante il cambiamento dei valori delle variabili indipendenti.

Un valore di 1.0 delinea un mancata associazione. I valori degli odds ratio possono essere più alti o più bassi di 1,0. La grandezza della relazione è misurata dalla distanza da 1.0. Un odds ratio minore di 1.0 indica una relazione negativa o inversa, viceversa un valore maggiore di 1.0 indica una relazione positiva.

[4] Un'importante proprietà degli odds ratio è la costanza: non varia al variare delle altre variabili indipendenti

Matematicamente gli Odds Ratio vengono calcolati come gli esponenziali dei coefficienti β risultanti dal modello in quanto partendo dall'equazione:

$$ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad (1.21)$$

Troviamo:

$$\frac{P}{1-P} = e^{\alpha+\beta x} \qquad (1.22)$$

Da cui risultano rispettivamente gli Odds che un determinato evento si verifichi e gli odds che tale evento non si verifichi.

$$Odds_{d|e} = e^{\alpha+\beta x} \qquad Odds_{d|\bar{e}} = e^{\alpha} \qquad (1.23)$$

Gli Odds ratio per definizione sono il rapporto tra gli odds, ovvero:

$$OR = \frac{e^{\alpha+\beta x}}{e^{\alpha}} = e^{\beta} \qquad (1.24)$$

Tale procedimento dimostra come gli Odds Ratio possono essere calcolati come gli esponenziali dei coefficienti β.

1.2.6 MODELLARE LE PROBABILITA'

Il modello del Multinmial logit può essere espresso anche in termini di probabilità originali π_{ij} invece dei log odds (come abbiamo visto nel parafrago 1.1). Adottando la condizione che η_{ij}=0 possiamo affermare che:

$$\pi_{ij} = exp(\eta_{ij}) / \Sigma_{ij}\, exp(\eta_{ik}) \qquad per\ j=1...J \qquad (1.25)$$

Per verificare questo risultato esponenziamo la formula $\eta = \log \pi_{ij}/ \pi_{ij}$ =a+bx per ottenere:

$$\pi_{ij} = \pi_{ij}\, exp(\eta_{ij}) \qquad (1.26)$$

La convenzione $\eta_{ij}=0$ rende questa formula valida per tutte le j. Successivamente raccogliamo le j e usiamo il fatto che $\Sigma_{ij}=1$ per ottenere:

$$\pi_{ij} = 1/\Sigma_j\, exp(\eta_{ij}) \qquad (1.27)$$

Infine usiamo il risultato sulla formula per πij; si noti che l'equazione

$$\pi_{ij} = exp(\eta_{ij}) / \Sigma^{J}_{k=1}\, exp(\eta_{ij}) \qquad (1.28)$$

automaticamente darà la precedenza a probabilità che aggiungono fino ad 1 per ogni i.

1.2.7 PROBABILITA' STIMATA DELLA REGRESSIONE

L'obiettivo della regressione logistica è quello di stimare *p* ($\hat{p}$), collegando la variabile indipendente alla distribuzione Bernulliana attraverso il logit. Partendo dalla funzione che lega le variabili indipendenti (il logit: equazione 1.3), va calcolato l'antilogit per stimare *p:*

$$\frac{p}{1-p} = e^{\beta_0+\beta_i X_i} \qquad (1.29)$$

$$p = e^{\beta_0+\beta_i X_i}\,(1-p) \qquad (1.30)$$

$$\widehat{\boldsymbol{p}} = \frac{e^{\beta_0+\beta_i X_i}}{1+e^{\beta_0+\beta_i X_i}} \qquad (1.31)$$

Per risolvere quest'ultima equazione, è necessario sostituire i valori dei coefficienti (β_i) e delle *X* (come vedremo nel quarto capitolo). Una volta calcolate le probabilità stimate, è possibile procedere con il calcolo degli effetti marginali (ovvero la differenza tra la probabilità di successo e di insuccesso) e degli Odds.

1.2.8 STIMARE IL MAXIMUM LIKELIHOOD

Il MLE (Maximumi Likelihood Estimation) è una formula per predire la probabilità che un individuo scelga una certa alternativa; la funzione per tali modelli è il prodotto delle probabilità di scelta per ogni individuo. Le probabilità di scelta sono relativamente semplici, attraverso l'uso del computer si può massimizzare la funzione quasi istantaneamente, anche per un vasto numero di scelte. Possiamo stimare i parametri di questo modello attraverso il processo MLE con le probabilità π_{ij} viste come funzione dei parametri a_i e b_i nella funzione:

$$\eta_{ij} = log\ \pi_{ij}/\pi_{ij} = a_i + bx \qquad (1.32)$$

la funzione logaritmica è log $L = N\ n=1\ J\ j=1\ dnj\ log(Pnj)$ (1.33)

in cui dnj = 1 se l'individuo n sceglie l'alternativa j e dnj = 0. Il maximum likelihood estimator $\beta\hat{}$ è costante, asintoticamente efficiente e normalmente distribuito.

1.2.9 INDIPENDENZA DELLE ALTERNATIVE IRRILEVANTI

Il multinomial è una soluzione al problema di classificazione, che assume che una combinazione lineare di caratteri osservati e di parametri, possano essere usati per determinare la probabilità di ogni particolare risultato di variabili dipendenti.
É dato per assunto che i dati siano specifici al caso, ovvero che la variabile indipendente abbia un singolo valore per ogni caso e che la variabile indipendente possa non essere predetta con precisione per ogni caso dalle variabili indipendenti.
Non è necessario che le variabili indipendenti siano statisticamente indipendenti tra di loro, tuttavia si assume che la relazione lineare sia bassa; poiché altrimenti sarebbe difficile differenziare l'impatto di ogni diversa variabile.
Il multinomial si basa sull'assunto dell'indipendenza di alternative irrilevanti (IIA) e afferma che le probabilità di preferire una classe ad un'altra non dipendono dalla presenza o assenza di altre alternative irrilevanti. Ciò permette alla scelta di *K*

alternative di essere studiata come un set di K-1 scelte binarie indipendenti, nelle quali una è usata come pivot e le altre K-1 vengono comparate a questa una alla volta.
É per tale assunto che il calcolo dell'errore in tale modello è indipendente e distribuito in modo omogeneo secondo la distribuzione del valore.
Studi psicologici hanno rilevato, tuttavia, che gli individui spesso violano questo assunto quando prendono delle decisioni, è per tale motivo che questo è uno degli argomenti più discussi e controversi tra gli studiosi della materia.
Nello specifico la discussione ricade sulla scelta del modello da utilizzare, in quanto il Multinomial Probit, un modello simile, non presenta tale assunto e presenta un calcolo diverso dell'errore quindi potrebbe sembrare più adatto durante l'analisi delle scelte. Studi hanno tuttavia negato tale ipotesi, confermando una maggiore stabilità del modello MNL a discapito del modello MPL che in alcuni casi può risultare meno accurato e meno stabile.
Ad ogni modo, comparando i due modelli, si è notato che la differenza delle predizioni è minima, quindi è preferibile usare il Modello Multinomial logit, in quanto la stabilità in statistica è fondamentale.
Un esempio è il Il Paradosso del Bus Rosso e del Bus Blu. Consideriamo la scelta di un individuo riguardante il mezzo di trasporto da utilizzare, e supponiamo che inizialmente la scelta dell'auto e del bus abbiano la stessa utilità:

$$Cn = Vauto = Vbus = V$$

$$Pauto = Pbus = ½$$

Ciò imlica che Pbus = P auto = 0.5 e Pbus / P auto =1

Supponiamo ora che sia introdotto un servizio bus che abbia servizi identici tranne per il fatto che i bus sono dipinti in blu e rosso

$$PB = PR$$

Ci aspetteremmo che PB = PR = 0.25; PC = 0.5 ma l'assunto IIA fa in modo che ciò non si verifichi, in quanto l'utilità non cambia tra il bus blu o rosso quindi:

$$Cn = Vblu = Vrosso = V$$

PB /PC = 1 (ratio does not change with new alternative)

PB /PR = 1 (by construction)

PB = PR = PC = 0.33

1.3 COSTUIRE IL MODELLO: METODO STEPWISE E TECNICHE DI SELEZIONE DELLE VARIABILI

Prima di procedere alla costruzione del modello è necessario selezionare le variabili più significative da includere nel modello multivariato. Il motivo principale per cui si usano tali strumenti di selezione automatica è la necessità di determinare il minor numero di variabili in grado di stimare accuratamente l'outcome.
Tra le strategie con cui costruire i modelli di regressione multinomiale troviamo il metodo stepwise, un algoritmo che semplifica il processo di ricerca del modello che possa spiegare un largo numero di variabili.
Questa funzione seleziona i possibili modelli, aggiungendo o togliendo una variabile al fine di migliorare il modello ad ogni step. Quando l'algoritmo non riesce a migliorare il modello, attraverso l'aggiunta o la sottrazione di una variabile, si ferma e restituisce il nuovo modello.
Il metodo stepwise può essere eseguito
-***Forward*:** si procede partendo da zero e inserendo una variabile alla volta
-***Backward:*** si inizia considerando il modello completo per poi eliminare una alla volta le variabili che hanno meno impatto sul fit.
-***Both:*** modello ibrido che incorpora caratteristiche di entrambi i metodi; il modello può quindi muoversi "avanti" o "indietro" aggiungendo o sottraendo variabili. Per analizzare la bontà del modello si usa il coefficiente AIC (Akaike's information criterion):

$$AIC = -2 * log(L) + k * gl \qquad (2.1)$$

In cui L è il "likelihood" e gl rappresenta i gradi di libertà.

L'AIC viene ricalcolato ogni volta che una variabile viene aggiunta o sottratta. I modelli vengono considerati per comparazione; il modello con l'AIC minore risulta essere quello che esprime meglio le relazioni tra le variabili.
Una volta costruito il modello che ci sembra più idoneo è utile usarlo per effettuare delle predizioni per nuovi dati.

1.3.1 BONTA' DEL MODELLO E INTERPRETAZIONE DEI COEFFICIENTI

Esistono vari coefficienti che determinano la bontà del modello, pare quindi opportuno effettuare una breve spiegazione di alcuni di essi.

- Il ***P-Value*** rappresenta il p-value del t-score ed esprime quanto statisticamente rilevante è ogni estimate. Il p-value è utilizzato per verificare se un determinata ipotesi (Ho) sia o meno verificata. Questo valore risulta rilevante non tanto per l'intercetta, quanto per le variabili predictors, p-value bassi indicano un rifiuto dell'ipotesi nulla, indicando quindi una stima dei coefficienti statisticamente rilevante (per rilevante si intende un valore minore di 0,05).
- Il ***T-Value*** rappresenta il rapporto tra il coefficiente β e l'errore standard $t\ value = \beta/SE_\beta$, ovvero quanto è grande l'estimate in confronto all'errore; più grande è l'estimate maggiore sarà il t-value e di seguito migliore sarà il modello. Il t-value misura la grandezza della distanza relativa nella variazione all'interno del campione. Più alto è il valore, maggiore è la certezza che l'ipotesi nulla sia rifiutata.
- Il P-value e il T-value sono i criteri usati per la selezione delle variabili da includere all'interno del modello e per decidere quindi quali escludere.
- L' ***Errore Residuale Standard*** rappresenta l'errore standard del modello e in generale esprime la bontà del modello nel predire i valori di Y all'interno dei dati, in media; in altre parole è "l'errore medio" del modello. Alcune volte il modello predice dei risultati "vicini" ai dati, mentre altre volte predice dei

risultati "lontani" dai dati, questo errore misura quanto il modello riesce a predire tale distanza dai dati. Più l'errore medio è alto più il modello è inaccurato, viceversa più è basso più il modello sarà accurato.

- Matematicamente è rappresentato da $\hat{\sigma}_e$:
- $$\hat{\sigma}_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{Somma\ degli\ errori\ al\ quadrato}{Errore\ dei\ gradi\ di\ libertà}} \qquad (2.2)$$

- L' ***R quadro*** o coefficiente di determinazione rappresenta la percentuale di variazione nella variabile response (Y) spiegato dalla variazione nelle variabili esplicative (predictors); in altre parole partendo dalla retta di regressione sintetizza di quanto si discosti in media la grandezza analizzata rispetto alla retta. L'R^2 è una misura della bontà nonché l'adattabilità del modello, ovvero la sua capacità di previsione; può assumere valori tra 0 e 1, lo zero esprime la mancanza di relazione tra le variabili considerate, l'uno indica una perfetta relazione tra il fenomeno e la retta. Ciò significa che più è alto il valore, migliore è il modello, tuttavia in alcuni campi un alto valore può essere rappresentato da 0,2-0,3 mentre in altri campi può essere 0,5-0,6.

  Matematicamente l'R quadro è definito come *il rapporto tra devianza totale e devianza spiegata,* oppure come *l'inverso del rapporto tra la devianza d'errore e la devianza totale*[5] ovvero il quadrato dei coefficienti di correlazione.
- L'***R quadro aggiustato***, è calcolato allo stesso modo dell'R^2 tranne per il particolare che è "aggiustato" rispetto al numero di variabili considerate nel modello. Viene interpretato allo stesso modo dell'R^2 e viene inserito in modo tale da penalizzare i modelli per l'aggiunta di variabili inutili; infatti aggiungendo una variabile al modello matematicamente l'R^2 risulta sempre maggiore (aumenta all'aumentare delle variabili) ma l'aggiunta di tale variabili potrebbe non essere utile al modello in termini esplicativi e di rilevanza statistica, mentre l' $Adjusted\ R^2$ penalizza i modelli che considerano variabili inutili (diminuisce all'aumentare delle variabili inutili), per questo sempre è minore rispetto all'R^2, se però il modello contiene delle variabili non rilevanti

[5] Regressione Multipla e Regressione Logistica: concetti introduttivi ed esempi Vincenzo Paolo Senese

allora l' $Adjusted\ R^2$ risulta molto minore rispetto all'R^2 per questa ragione vengono considerati entrambi.

- L'***F-statistic*** è una misura della bontà del modello in generale, è rilevante quando ci sono più variabili esplicative, nel caso di un modello con una sola variabile esplicativa coincide con il $T - value^2$.

  Matematicamente: $F = \frac{media\ del\ model\ al\ quadrato}{media\ dell'errre\ al\ quadrato}$

2 CAPITOLO II
L'ANALISI DEI DATI

2.1 LO SCOPO DELL'ANALISI DEI DATI

L'analisi dei dati riguarda lo studio di dati assemblati e raggruppati per la ricerca delle caratteristiche del soggetto in esame e per determinare le varie tipologie di relazioni che intercorrono tra le variabili ad esso correlate. L'obiettivo di tale analisi è quello di sintetizzare dati in forme comprensibili e chiare, identificare i *casual factors,* sottolineare i fenomeni complessi, tracciare interferenze attendibili dai dati osservati e infine creare stime dai risultati ottenuti sul campione.

L'analisi statistica si distingue in Descrittiva ed Inferenziale, la prima descrive la natura degli oggetti di studio e può esaminare i dati usando una(monovariata), due(bivariata) o più(multivariata) variabili.

L'analisi Inferenziale cerca, invece, di tracciare inferenze e conclusioni sulla popolazione, partendo dai risultati di uno studio di ricerca su un campione e utilizzando come strumenti la Stima e il Test d'ipotesi.

2.2 L'ANALISI DESCRITTIVA

La tipologia di analisi da eseguire deve tener conto di diversi fattori come, ad esempio, lo scopo dell'analisi e la natura del dato (qualitativo, quantitativo o continuo).

Un primo importante momento dell'analisi dei dati è rappresentato dalle analisi descrittive (uni-variate), metodo che studia una singola variabile per volta e che,

conseguentemente, analizza ogni aspetto del fenomeno singolarmente. Queste presentano una rappresentazione iniziale della distribuzione di ciascuna variabile, facilitando la comprensione della composizione del campione e della sua emblematicità e permettendo, infine, di indicare le tecniche statistiche da adoperare per una successiva e più esauriente analisi di decodificazione dei dati.

L'analisi monovariata è un'analisi unicamente descrittiva, che circoscrive lo studio dei metodi con cui le variabili vengono distribuite fra i casi rilevati, senza sviluppare una ricerca sulle relazioni presenti tra tali variabili. Essa, oltra ad essere una preliminare descrizione dei fenomeni oggetto di studio, costituisce un punto di partenza necessario per una successiva analisi multivariata, in quanto, solo attraverso lo studio di tali dati il ricercatore acquista quella conoscenza diretta che gli consentirà di analizzare i dati con piena consapevolezza.

Le informazioni estratte dall'indagine sono importate nel dataset, che contiene i dati organizzati in modo tale da essere pronti per essere utilizzati per le analisi, sotto forma di matrice dei dati, ovvero un insieme rettangolare di numeri, in cui ogni riga raffigura le unità d'analisi (gli individui intervistati) e ogni colonna riproduce, per ogni caratteristica, l'insieme delle misurazioni calcolate per individuo in ogni cella. Derivante dall'incrocio tra una riga e una colonna abbiamo un dato, cioè il valore assunto da una particolare variabile su un particolare caso. Ogni carattere può assumere diversi valori, detti modalità.

Per poter costruire una matrice, è necessario che le informazioni rispondano a due criteri: l'unità d'analisi deve essere sempre la stessa e su tutti i casi studiati devono essere rilevate le stesse informazioni.

Il materiale presente in matrice necessita solitamente di una codifica, ovvero una traduzione delle informazioni, questa può avvenire attraverso l'utilizzo di due strumenti: il codice (assegnare un numero ad ogni modalità) e il tracciato record (posizionare ogni variabile nella riga della matrice).

Ogni riga coincide ad un caso (ovvero l'insieme di tutte le risposte date da una medesima unità); mentre ogni colonna rappresenta una variabile (ovvero tutte le risposte date ad un medesimo quesito da tutti gli intervistati).

La rappresentazione dei dati avviene attraverso una distribuzione di frequenza, ovvero una tabella che rileva le frequenze con cui si ripetono le diverse modalità, che serve a riprodurre il numero di volte in cui un determinato fenomeno si è verificato.

2.3 ANALISI DESCRITTIVA DEL DATASET

L'obbiettivo primario di questo capitolo è sintetizzare le caratteristiche più salienti del fenomeno oggetto di studio. I dati che utilizzeremo riguardano un'indagine effettuata da un'azienda vinicola sul consumo di prosecco.

Lo scopo è analizzare le preferenze dei consumatori riguardanti il packaging e testare le potenzialità di 3 nuovi design al fine di:

- **Conoscere i consumatori**
- Determinare se e **quanto il design** della bottiglia **incide sulla scelta finale** del consumatore
- Comparare il **grado di apprezzamento della bottiglia** attualmente in commercio, rispetto ai maggiori competitors
- Analizzare la possibilità di **modificare tale design** con uno dei tre proposti e il conseguente grado di apprezzamento dei suddetti.

Le osservazioni dell'indagine sono state raccolte in un database, e successivamente elaborate attraverso la redazione di tabelle di frequenza, nelle quali ad ogni valore della variabile viene associata la frequenza con la quale esso si presenta nei dati analizzati per dare una visione generale, iniziale e sintetica dei quesiti in esame.

I dati sono stati trasferiti in un semplice file Excel, che mi ha consentito di lavorare su di essi e di estrarre le informazioni rilevanti ai fini dell'indagine. In questo capitolo presenterò le tabelle di frequenza e le relative interpretazioni dei risultati ottenuti.

Prima di procedere nell'analisi dei risultati ritengo opportuno fare un'ulteriore premessa sulla struttura delle domande che seguiranno per rendere più chiara la lettura e la comprensione dei dati.

Il sondaggio in esame consiste nell'intervista a un campione opportunamente selezionato di soggetti appartenenti alla popolazione obiettivo che si compone di 190 individui. Il sondaggio è *strutturato* (ovvero presenta un elenco di domande con un preciso ordine) e *diretto* (ovvero l'intervistato è a conoscenza dello scopo dell'indagine), è inoltre caratterizzato da una *forma chiusa*, in quanto vengono proposte, per ogni quesito (tranne che per l'età) una serie di risposte possibili, ovvero risposte chiuse.

Il questionario si compone di 39 domande che per semplicità suddivideremo in tre macro-aree:

1. nella prima parte si raccolgono le informazioni ***generali*** sugli utenti;
2. nella seconda si rilevano le valutazioni inerenti alle ***abitudini*** degli utenti;
3. nell'ultima parte si prendono informazioni sulle ***preferenze*** degli utenti.

Nella prima categoria vengono analizzate le domande di carattere generico fatte dall'intervistatore per rilevare caratteristiche oggettive della popolazione, rientrano quindi in quest'area le domande sull'età, sul sesso, sul grado di istruzione e sulla località abitativa.

Le informazioni sulle abitudini degli intervistati servono all'azienda per capire che tipo di consumatore ha davanti, in modo tale da poter sviluppare dei prodotti o servizi ad hoc o per diversificare in base alle varie fasce di interesse. In questo caso l'azienda ha deciso di analizzare la frequenza dei consumi e dell'acquisto di vino e di spumante oltre alle occasioni in cui viene consumato e alla disponibilità di prezzo.

Le informazioni sulle preferenze vengono considerate diversamente da quelle sulle abitudini in quanto, basandosi sui gusti dei consumatori, non sono riconducibili alle abitudini di consumo.

Nei paragrafi seguenti analizzerò ed elaborerò i dati relativi alle 190 unità rilevate per ognuna delle aree sopracitate, utilizzando dati percentuali in quanto di più immediata comprensione.

2.3.1 INFORMAZIONI GENERALI SUI CONSUMATORI

L'area d'indagine è rappresentata da due città Italiane: Padova e Milano. La maggior parte degli intervistati proviene da Padova con una percentuale del 55% mentre il 45% proviene da Milano,

Città	*n°intervistati*
Milano	86
Padova	104
tot	190

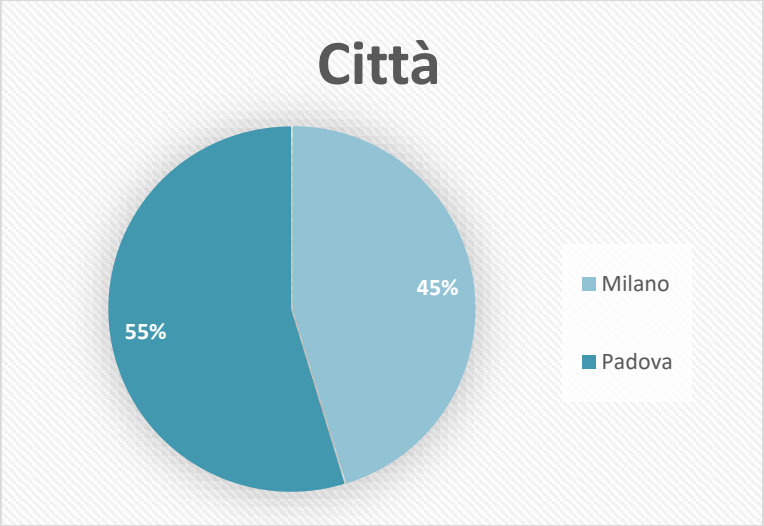

Grafico 2.1 - Dati sulla Città di appartenenza degli intervistati

Tra le unità in analisi la maggior parte risultano essere donne, con una percentuale del 56%, rispetto agli uomini con il 44%.

Sesso	*N°*
M	83
F	107

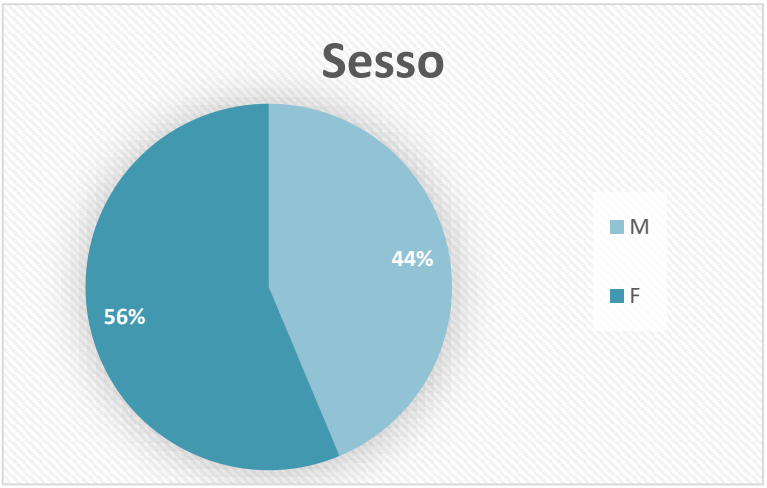

Grafico2.2- Dati sul Sesso degli intervistati.

La fascia di intervistati più rilevante è formata da persone aventi più di 60 anni (circa 26%) seguita da coloro tra i 31 e i 40 anni. L'età è un dato molto interessante in quanto permette di capire, incrociando i diversi dati, quali siano le fasce di consumatori con simili esigenze e preferenze e permette successivamente di individuarli e soddisfarli.

Età	*n°*
19-30	29
31-40	41
41-50	39
51-60	26
60 +	50

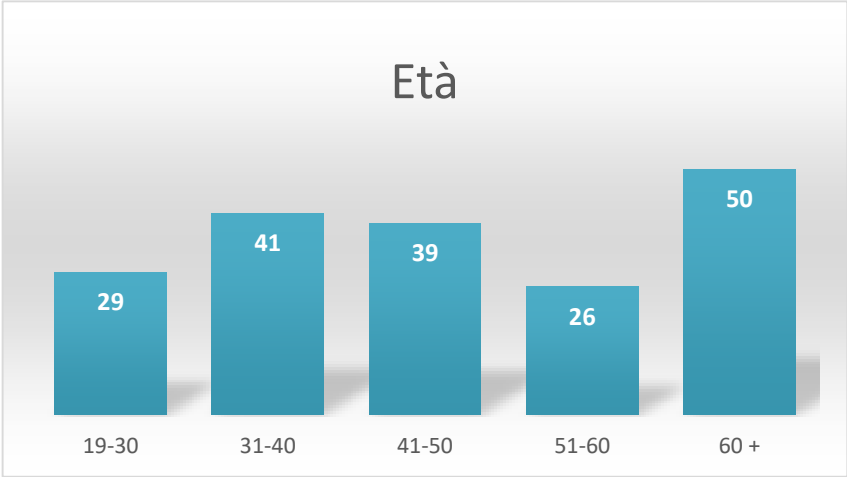

Grafico 2.3 – Dati sull' Età degli intervistati

La maggior parte degli utenti ha conseguito una laurea (41%) o un diploma superiore (44%), ciò denota un alto tasso di istruzione in circa l'85% degli intervistati, contro il 15% che ha conseguito il diploma di scuola media (12%) o elementare (3%) e ha quindi un basso livello di formazione scolastica.

Titolo di studio	*n°*
Laurea	78
diploma superiore	83
Scuola media	23
Scuola elementare/nessuno	5
NS/NR	1

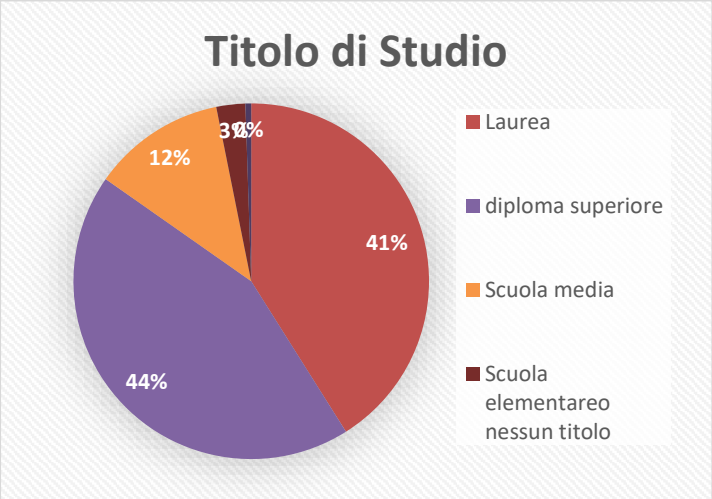

Grafico 2.4 – Dati sul Titolo di studio degli intervistati

Da questa prima analisi delle informazioni generali abbiamo appreso che:

- gli intervistati sono prevalentemente padovani e prevalentemente donne,

- la maggior parte ha un'età superiore ai 60 anni o compresa tra i 30 e i 40

- il livello di istruzione è alto.

Passiamo ora ad analizzare le caratteristiche riguardanti le abitudini.

2.3.2 INFORMAZIONI SULLE ABITUDINI DI CONSUMO DEGLI INTERVISTATI

Tra gli utenti, coloro che affermano di consumare vino una volta a settimana sono il 38%, seguiti subito dopo da chi afferma di berlo più volte a settimana con il 34,5%, infine chi afferma di consumarlo tutti i giorni è il 27,5%.

Consumo di Vino	*N°*
1 volta a settimana	73
più volte	66
tutti i giorni	51

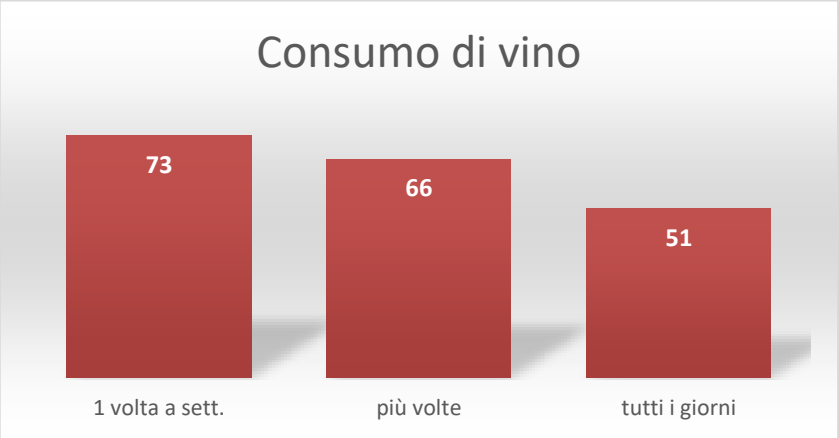

Grafico 2.5 – Dati sul Consumo di vino degli intervistati

Gli acquisti di vino e prosecco, risultano avere un andamento simile ed è quindi utile analizzarli contemporaneamente. La maggioranza dei soggetti afferma di aver acquistato vino o prosecco nell' ultimo mese (56% prosecco e 86% vino), le percentuali decrescono invece drasticamente nei periodi successivi.

Il vino ha una forte crescita nel primo periodo e una forte decrescita dal secondo periodo in avanti, mentre il prosecco pur avendo una decrescita non presenta la stessa variazione improvvisa, ciò probabilmente è attribuito alla maggiore propensione al consumo di vino piuttosto che di spumante e anche alla diversa destinazione;

Ultimo acquisto di vino	**N°**
Ultimo mese	164
da 1 a 6 mesi	17
da 6 mesi a 1 anno	5
più di un anno	2
mai acquistato	1
Ns/Nr	1

Ultimo acquisto di prosecco	**N°**
ultimo mese	107
da un mese a sei mesi	50
da sei mesi a un anno	16
più di un anno	7
Mai	2
NS/NR	8

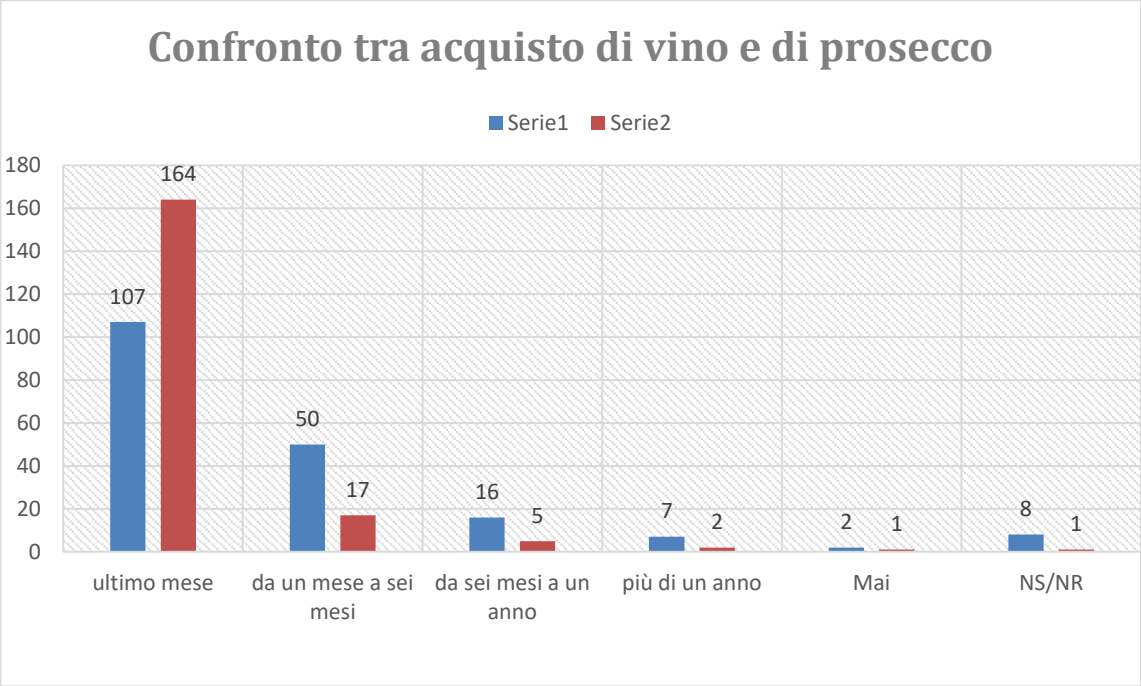

Grafico 2.6- Dati sull'ultimo acquisto di vino e di prosecco a confronto

Infatti, il principale motivo d'acquisto è quello per occasioni particolari, seguito dal consumo abituale. Si evidenzia, pertanto, la prevalenza del consumo di prosecco in occasioni particolari, piuttosto che abitualmente come invece possiamo supporre avvenga per il vino.

Motivo dell'acquisto del Prosecco	*N°*
Consumo abitudinale	54
Occasioni particolari	110
Regalo	16
Non lo acquisto	0
NS/NR	10

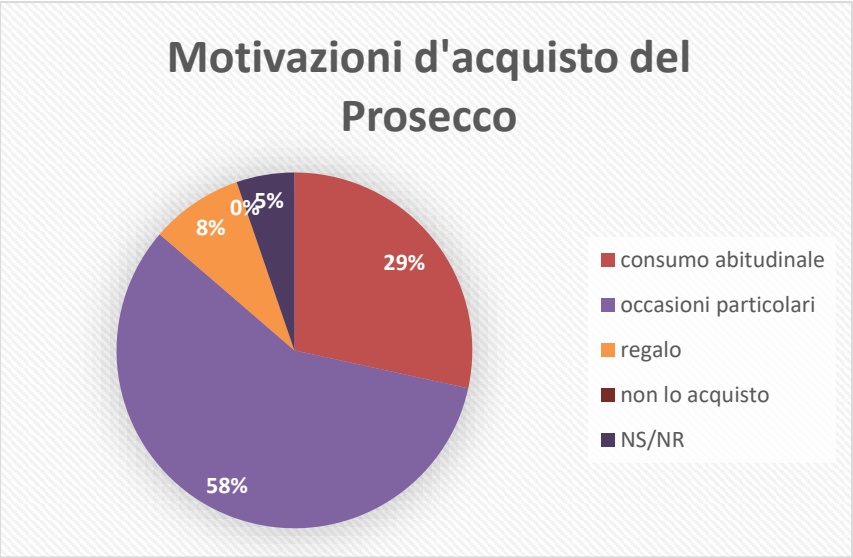

Grafico 2.7 – Dati sulla motivazione d'acquisto di Prosecco

Riguardo le abitudini sulle occasioni di consumo, la maggior parte degli intervistati afferma di consumare il prosecco principalmente nelle occasioni di festa e successivamente come aperitivo. Solo il 18% dichiara di consumare il prosecco durante i pasti. La percentuale di coloro che asseriscono di non aver mai consumato prosecco o di risposte non specificate è pressoché irrilevante.

Occasioni di consumo del prosecco	*N°*
Mai	8
Nelle occasioni di festa	113
Come aperitivo	101
Durante i pasti	51
ns\nr	2

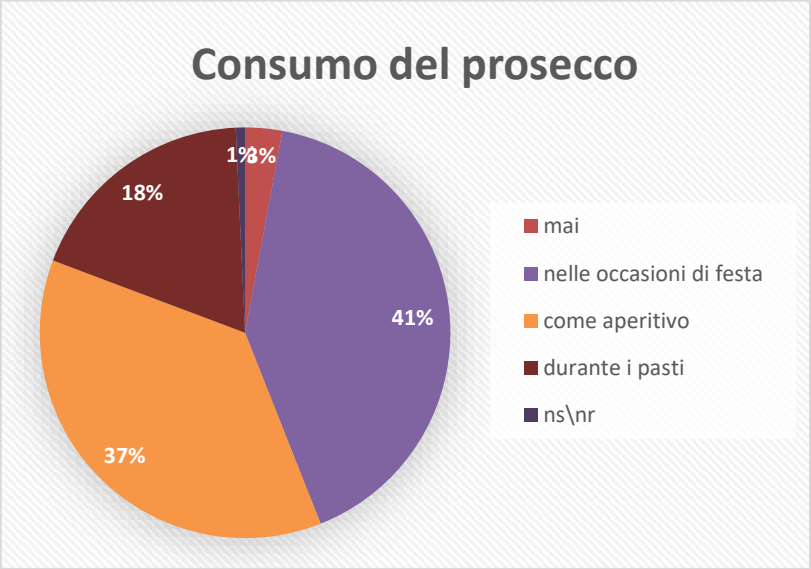

Grafico 2.8 – Dati sul Consumo di prosecco degli intervistati

Per quanto concerne il prezzo, la maggior parte degli intervistati è risultata incline a pagare un prezzo dai 5, agli oltre 6 euro per uno spumante; ciò delinea un'alta disponibilità a pagare del consumatore, in quanto il prezzo medio degli spumanti della marca "A" si aggira intorno ai 5 euro.

Prezzo	*N°*
min 4 euro	15
tra 4 e 5 euro	48
tra 5 e 6 euro	59
oltre 6	58
NS/NR	10

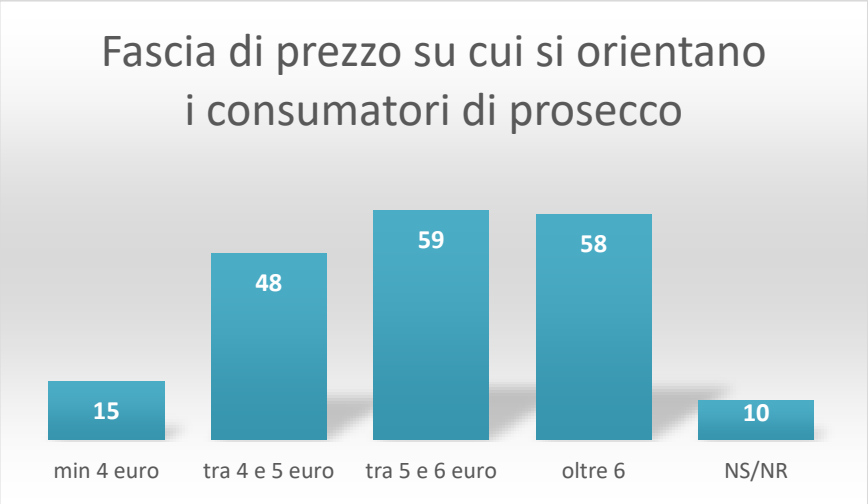

Grafico 2.9 – Dati sulla fascia di prezzo su cui si orientano gli intervistati durante l'acquisto di Prosecco

Volendo riepilogare i tratti più considerevoli dell'analisi effettuata sulle abitudini, possiamo affermare che:

- si rileva la prevalenza del consumo di vino rispetto al prosecco,

- si evidenzia come il principale motivo dell'acquisto e del consumo risulti essere quello per occasioni particolari,

- si osserva, generalmente, un'alta disponibilità a pagare.

2.3.3 INFORMAZIONI SULLE PREFERENZE DEGLI INTERVISTATI[6]

- *PREFERENZA TRA LE TRE BOTTIGLIE PROPOSTE*

Viene chiesto agli intervistati di classificare il grado di preferenza di 3 bottiglie proposte, in ordine decrescente, indicando quindi per prima la bottiglia preferita e per ultima quella meno preferita; (queste 3 bottiglie non sono sul mercato).

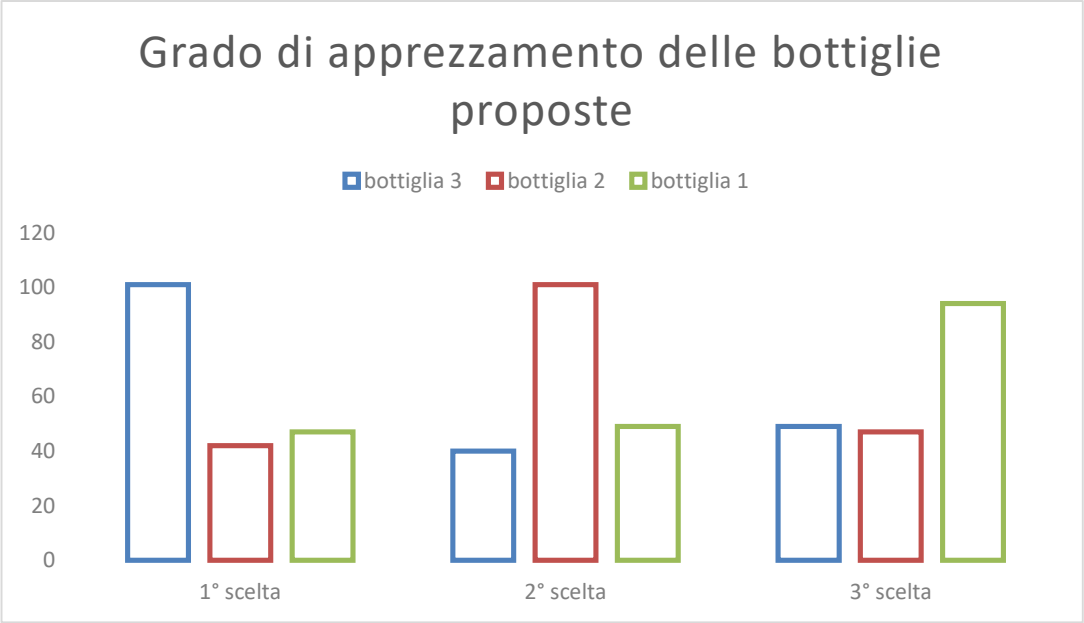

Grafico 2.10 – Dati sulla preferenza tra tre nuove bottiglie proposte. In risposta alla domanda: Quale tra queste versioni della bottiglia di prosecco "A" preferisce? (in ordine da 1 a 3)

Al primo posto tra le preferite risulta la bottiglia 3 con 101 preferenze su 190; al secondo posto si trova la bottiglia 2 sempre con 101 preferenze su 190, e al terzo posto la bottiglia 1 con 94 preferenze.

É evidente che il design della bottiglia 3 è quello più accattivante secondo gli intervistati, mentre il design della bottiglia 1 è quello che ha riscontrato minor interesse.

[6] N.B. Tutte le tabelle contenenti i dati dei vari grafici elencati in questo paragrafo si trovano in appendice A

- *CONFRONTO TRA LA BOTTIGLIA ATTUALE "A" E UNA DELLE TRE NUOVE*

Viene ora mostrata la bottiglia attualmente in commercio di A (che da ora chiameremo A Attuale) in confronto a una delle tre bottiglie, scelta a caso, menzionate nella domanda precedente (non ancora in commercio).

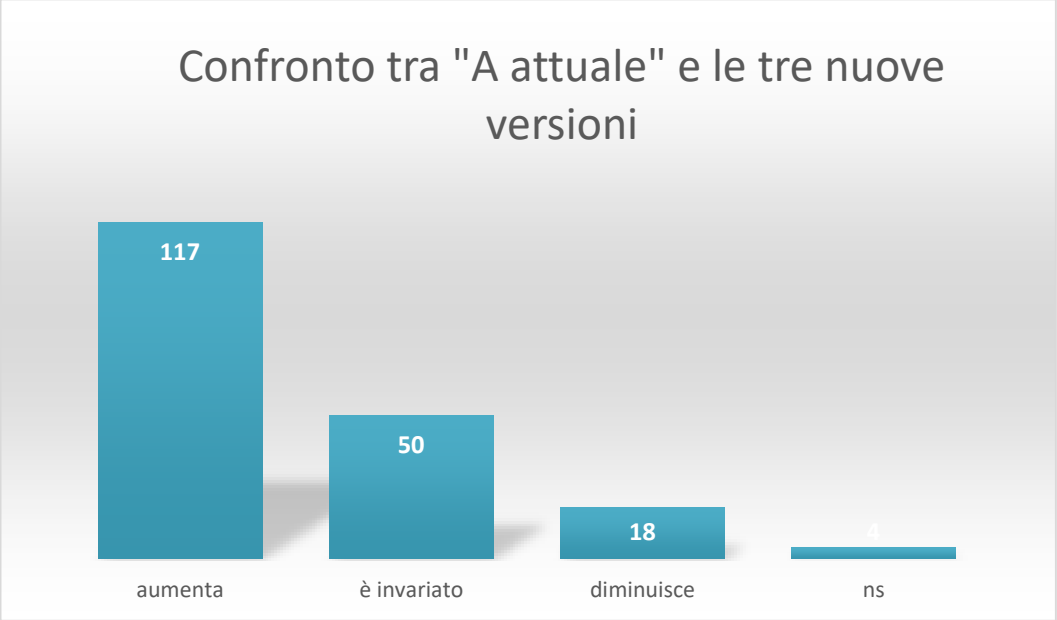

Grafico 2.11 – Dati sul confronto tra la bottiglia "A Attuale" e le tre nuove proposte, in risposta alla domanda: Guardando queste due bottiglie, questa nuova bottiglia aumenta, diminuisce o lascia invariata la sua propensione all'acquisto di prosecco a marchio "A"?

Da tale confronto emerge la propensione degli intervistati verso la nuova bottiglia proposta (ovvero una delle tre nuove proposte non ancora in commercio, scelta casualmente), in quanto più della metà degli intervistati (62%) afferma che la nuova bottiglia aumenta la propensione all'acquisto di prosecco "A", contro il 26% di "*è invariato*" e solo il 10% di "*diminuisce*".

- *PREFERENZA TRA LE BOTTIGLIE IN COMMERCIO*

Si chiede ora agli intervistati di esprimere la loro preferenza (con un massimo di 2 scelte) tra 6 bottiglie (5 sul mercato e una casuale tra le 3 considerate nel punto 1.2.1).

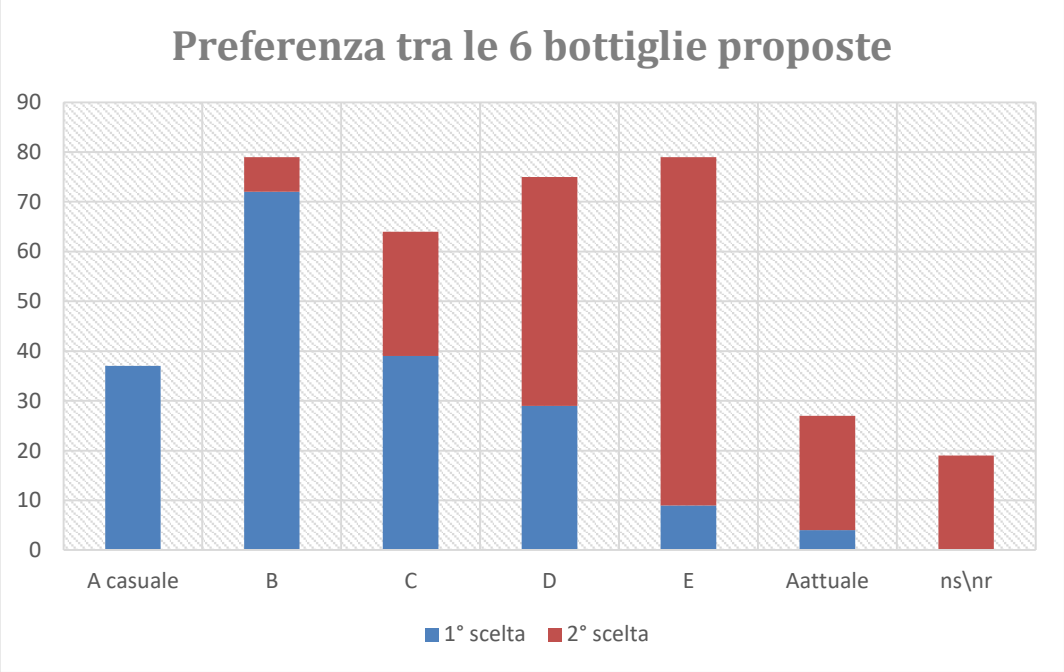

Grafico 2.12 – Dati sulla preferenza tra le sei bottiglie proposte, in risposta alla domanda: Osservi questi prodotti. Indipendentemente dalle sue abitudini di consumo o di acquisto, può indicare le due bottiglie che preferisce (massimo due)?

É evidente che la bottiglia preferita dagli intervistati è quella della marca "E" con più di 70 preferenze su 190. Al terzo posto troviamo la bottiglia di "A" con quasi 40 preferenze. Dato allarmante è però il livello di preferenze dato ad "A attuale" (la bottiglia attualmente in commercio della marca "A"), ciò dimostra che la bottiglia non è molto apprezzata dagli intervistati, seppur viene tenuta leggermente più in considerazione come seconda scelta.

La preferita come seconda scelta risulta essere la bottiglia "E" che viene scelta da quasi 70 persone.

- *QUANTO LE VARIABILI INFLUENZANO LA SCELTA DELLA BOTTIGLIA*

Nella seguente domanda viene chiesto agli intervistati quanto incidano sulla scelta della bottiglia diversi aspetti che l'intervistatore ritiene rilevanti, tra cui:

* Forma dell'etichetta
* Colori dell'etichetta
* Etichetta nel complesso
* Scritta prosecco
* Fascia sul collo della bottiglia
* Forma della bottiglia

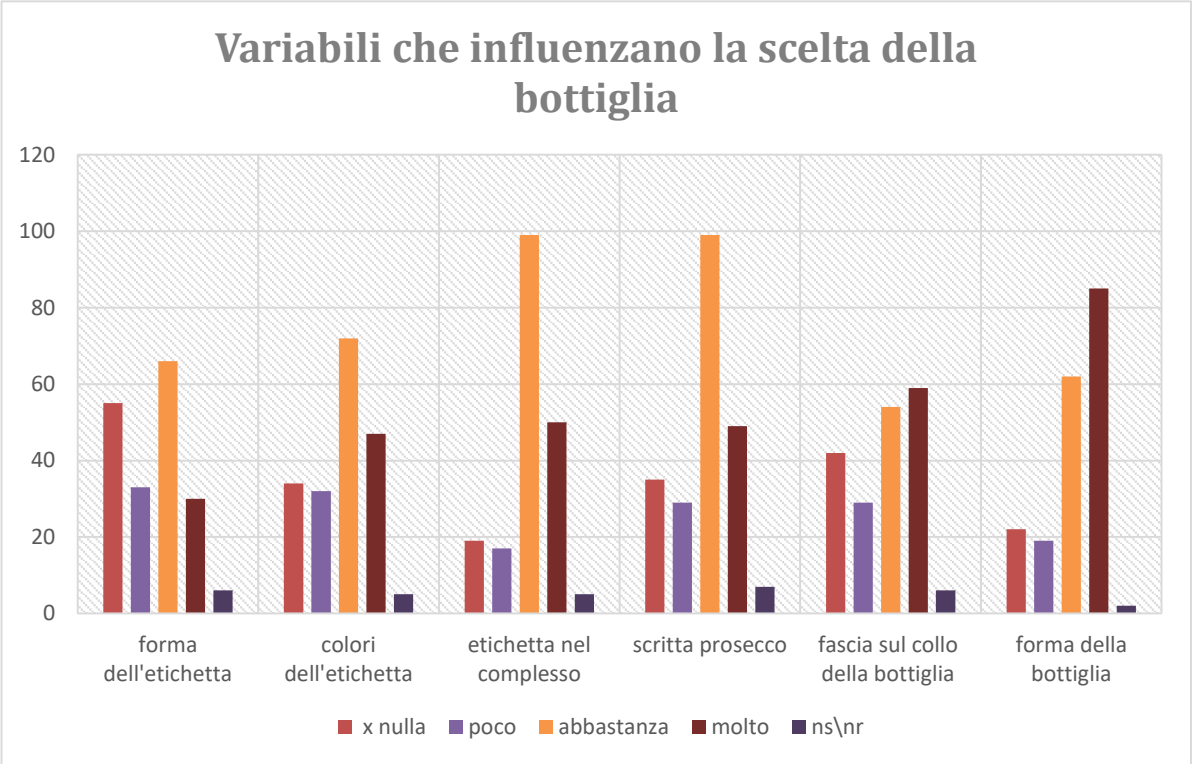

Grafico 2.13– Dati sulle variabili che incidono nella scelta d'acquisto di una bottiglia di prosecco; in risposta alla domanda: Guardando alla bottiglia che preferisce, può dire quanto ciascuno di questi aspetti ha inciso sulla sua scelta?

Si può notare come le variabili che hanno inciso maggiormente sulla scelta siano state "l'etichetta nel complesso" e la "scritta prosecco"; le variabili che hanno inciso meno sono state, invece: "fascia sul collo della bottiglia", "forma della bottiglia" e "colori dell'etichetta" seppure risultano essere variabili considerate (il numero di "abbastanza"

e “molto” risulta essere maggiore rispetto alla somma di “per nulla” e “poco”). Controversa è invece la variabile “forma dell’etichetta”, in quanto prevale il numero di “abbastanza” ma è seguito subito dopo da “per nulla”, ciò significa che quasi la metà degli intervistati considera la forma dell’etichetta durante la scelta della bottiglia, mentre l’altra metà non la ritiene una variabile rilevante.

- *QUALE BOTTIGLIA RISPECCHIA MAGGIORMENTE DETERMINATE AFFERMAZIONI*

Nella successiva domanda viene chiesto agli intervistati quale delle bottiglie in commercio rispecchia maggiormente le seguenti affermazioni:

* Esteticamente bella
* Moderna, Attuale
* Prosecco che acquisterei
* Prosecco da offrire o regalare
* Prosecco di elevata qualità

Gli intervistati avevano a disposizione 2 scelte.

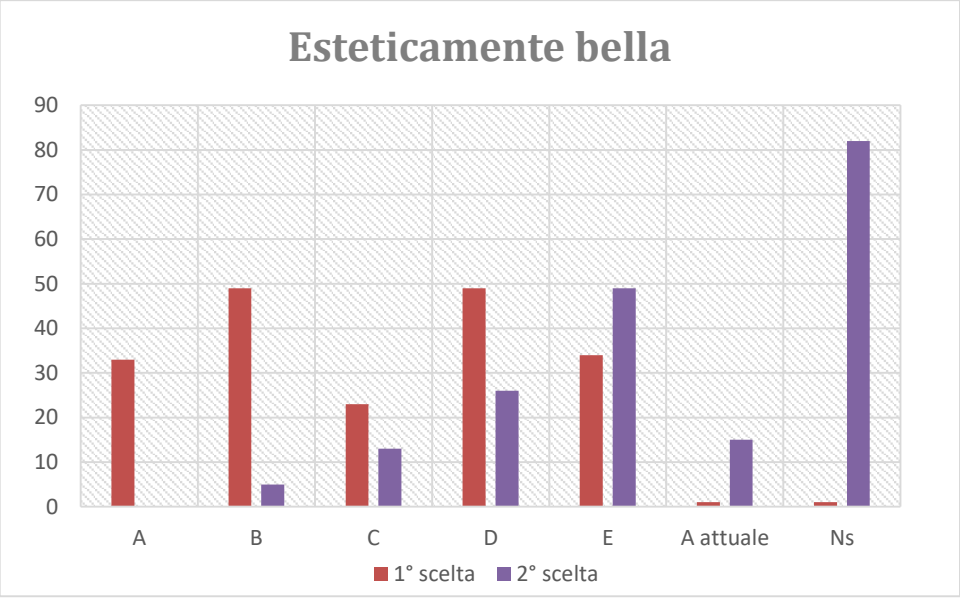

Grafico 2.14 – Dati riguardanti l’affermazione “esteticamente bella” in merito alle sei bottiglie considerate; in risposta alla domanda: Quanto queste bottiglie rispecchiano queste affermazioni (massimo due risposte)?

Prendendo in considerazione la variabile "esteticamente bella" notiamo che la preferenza ricade su B e D in egual misura (49 in entrambi i casi), mentre per la seconda scelta viene preferita a bottiglia E con 49 (dato importante considerato che 82 soggetti non hanno espresso la loro seconda preferenza).

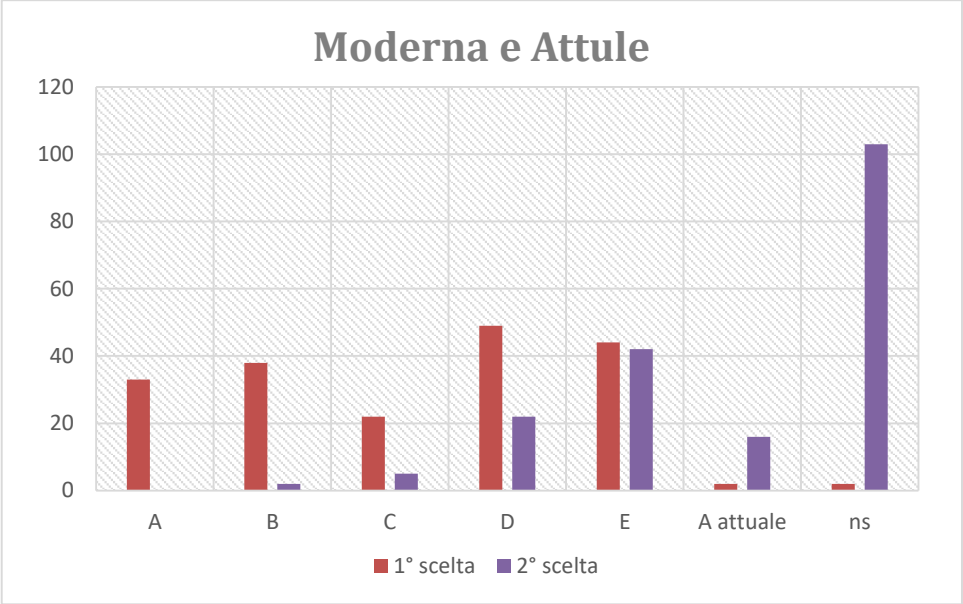

Grafico 2.15 – Dati riguardanti l'affermazione "Moderna e attuale" in merito alle sei bottiglie considerate; in risposta alla domanda: Quanto queste bottiglie rispecchiano queste affermazioni (massimo due risposte)?

Per la variabile "Moderna, Attuale" notiamo che la prima scelta ricade su D con 49, mentre la seconda su E con 42.

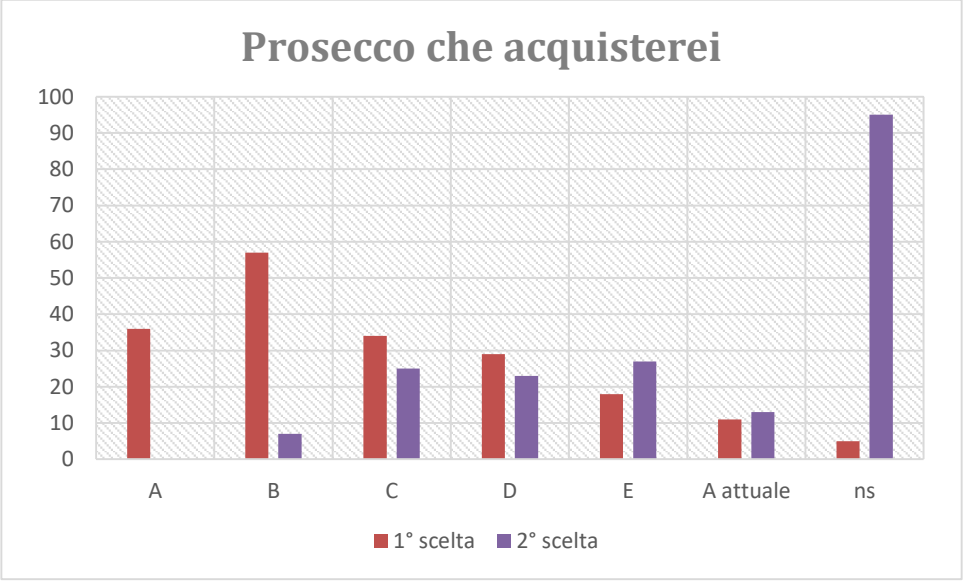

Grafico 2.16 – Dati riguardanti l'affermazione "prosecco che acquisterei" in merito alle sei bottiglie considerate; in risposta alla domanda: Quanto queste bottiglie rispecchiano queste affermazioni (massimo due risposte)?

La variabile “prosecco che acquisterei” mostra senza dubbio come la prima scelta ricada su “B” con 59 preferenze, mentre la seconda scelta ricade su E.

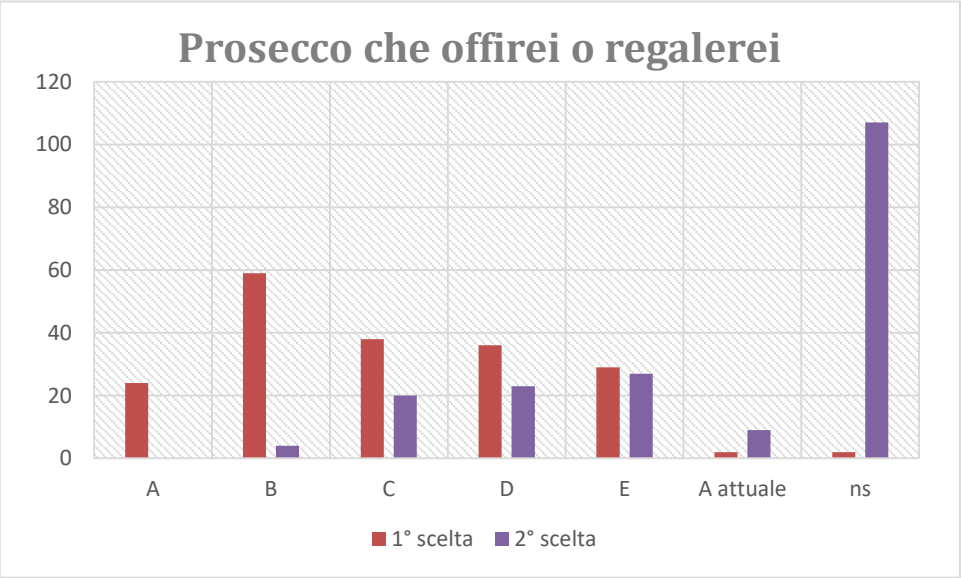

Grafico 2.17 – Dati riguardanti l’affermazione “prosecco che offrirei o regalerei” in merito alle sei bottiglie considerate; in risposta alla domanda: Quanto queste bottiglie rispecchiano queste affermazioni (massimo due risposte)?

Le stesse preferenze vengono evidenziate per “prosecco che offrirei o regalerei” con B come prima scelta con 59 preferenze ed E come seconda scelta con 27 preferenze.

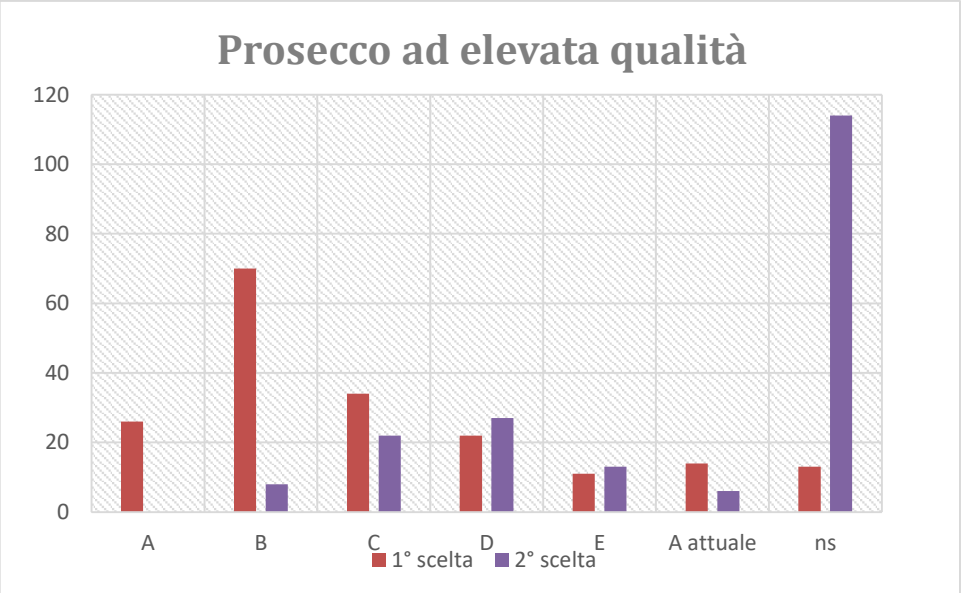

Grafico 2.18 – Dati riguardanti l’affermazione “prosecco ad elevata qualità” in merito alle sei bottiglie considerate; in risposta alla domanda: Quanto queste bottiglie rispecchiano queste affermazioni (massimo due risposte)?

Per “prosecco ad elevata qualità” la prima scelta ricade nuovamente su “B” con ben 70 preferenze, mentre la seconda scelta ricade su D con 27. La corrispondenza delle ultime 3 variabili dimostra conseguentemente una propensione all’acquisto di un prosecco ad elevata qualità da offrire agli amici. Ciò evidenzia l’importanza della variabile qualità al fine dell’acquisto.

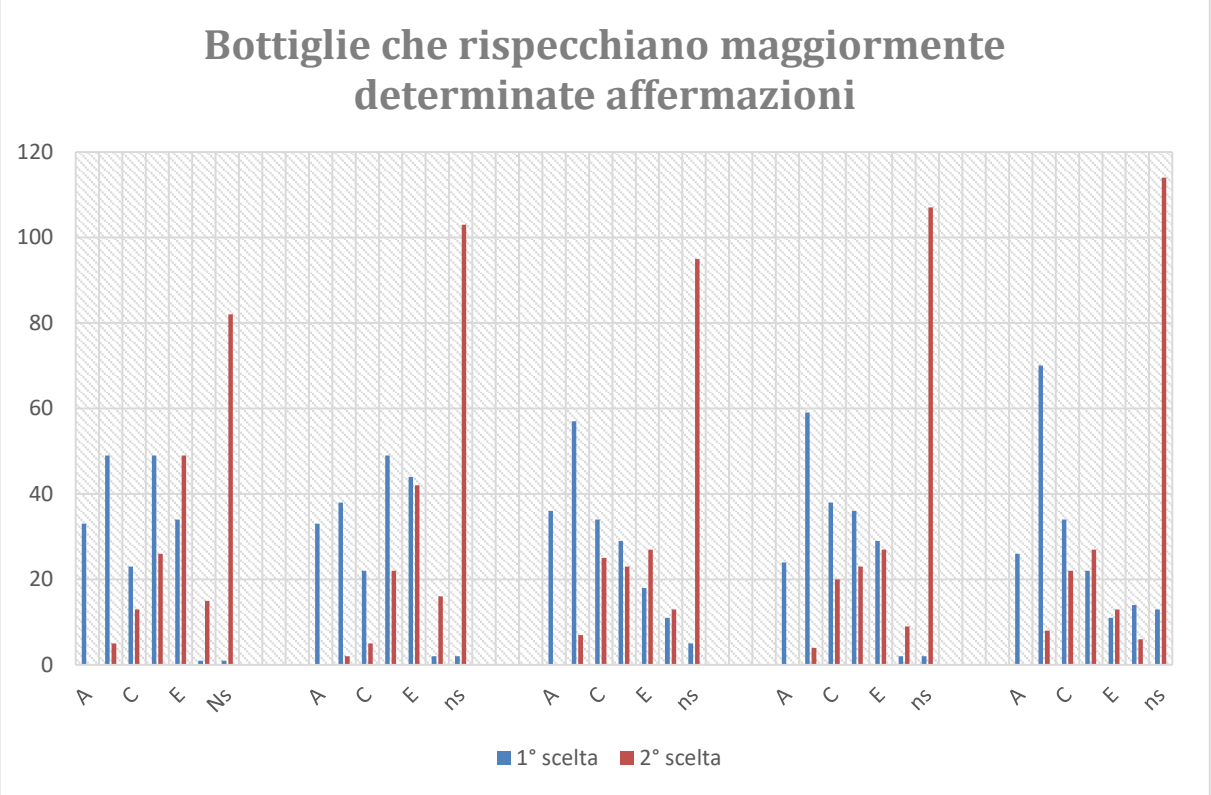

Grafico 2.19– Dati in risposta alla domanda: Quanto queste bottiglie rispecchiano queste affermazioni (massimo due risposte)? Riguardanti le affermazioni “esteticamente bella”, “Moderna e attuale”,”prosecco che acquisterei”, “Prosecco da offrire o regalare”, “Prosecco di elevata qualità”

Guardando al grafico nel suo complesse notiamo inoltre che una larga percentuale di intervistati, più della metà, ha deciso di non inserire una seconda scelta (come si evince dalla parte rossa di Ns). Inoltre è possibile sottolineare come “B” risulti quasi in tutte le domande la bottiglia avente più preferenze, tranne nel caso della variabile “Moderna e attuale”. Dati interessanti sono in fine quelli che riguardano “A”. Dai dati presi in analisi si evince come per tutte le variabili la bottiglia “A” venga preferita ad “A attuale”, ciò evidenzia che la bottiglia proposta dall’intervistatore, e non ancora in commercio, è largamente preferita rispetto a quella attualmente sugli scaffali. Inoltre in tutte le

variabili “A attuale” risulta essere quella con i più bassi punteggi, ovvero la meno preferita.

- *PREZZI ASSOCIATI A DIVERSE BOTTIGLIE*

Di seguito viene chiesto agli intervistati di associare alle diverse bottiglie già presenti in commercio (tranne A) un prezzo dai 4,50 ai 7 euro.

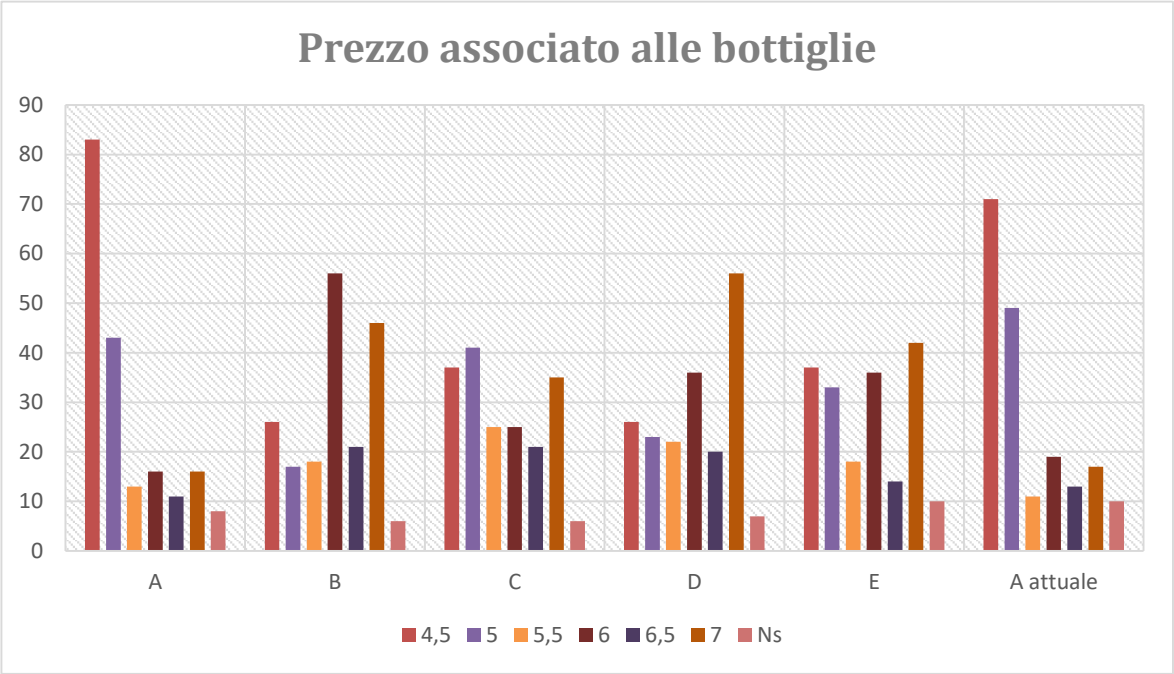

Grafico 2.20 – Dati sul prezzo associato alle sei bottiglie selezionate; in risposta alla domanda: Secondo lei quanto vale ciascuna di queste bottiglie?

Dall’analisi emerge che la bottiglia associata al prezzo più alto è la “D” seguita dalla “B” (alla quale viene associato in maggioranza sia il prezzo di 7 euro sia quello di 6 euro). Sono controverse le bottiglie “C” ed “E”, sulle quali non vi è una chiara tendenza di opinione come accade per le altre, infatti notiamo una quasi parità tra i diversi prezzi; ciò denota che metà degli intervistati ha associato un alto prezzo alle bottiglie mentre l’altra metà ha associato un basso prezzo. Per quanto riguarda la marca “A” la tendenza è invece inequivocabile, sia per “A” che per “A attuale” vengono associati i prezzi più bassi. A differenza del grafico precedente che vedeva preferita “A” ad “A attuale” in questo caso “A attuale” risulta avere un numero meno alto di associazioni ai prezzi più

bassi, rispetto ad "A"; ciò è confermato dal più alto numero di scelte del prezzo "5 euro" per "Attuale" rispetto ad "A".

- *QUANTO LA BOTTIGLIA RISPECCHIA LE AFFERMAZIONI SEGUENTI*

Il grafico seguente mostra i dati rilevati dalla domanda "quanto la bottiglia selezionata rispecchia le seguenti affermazioni".

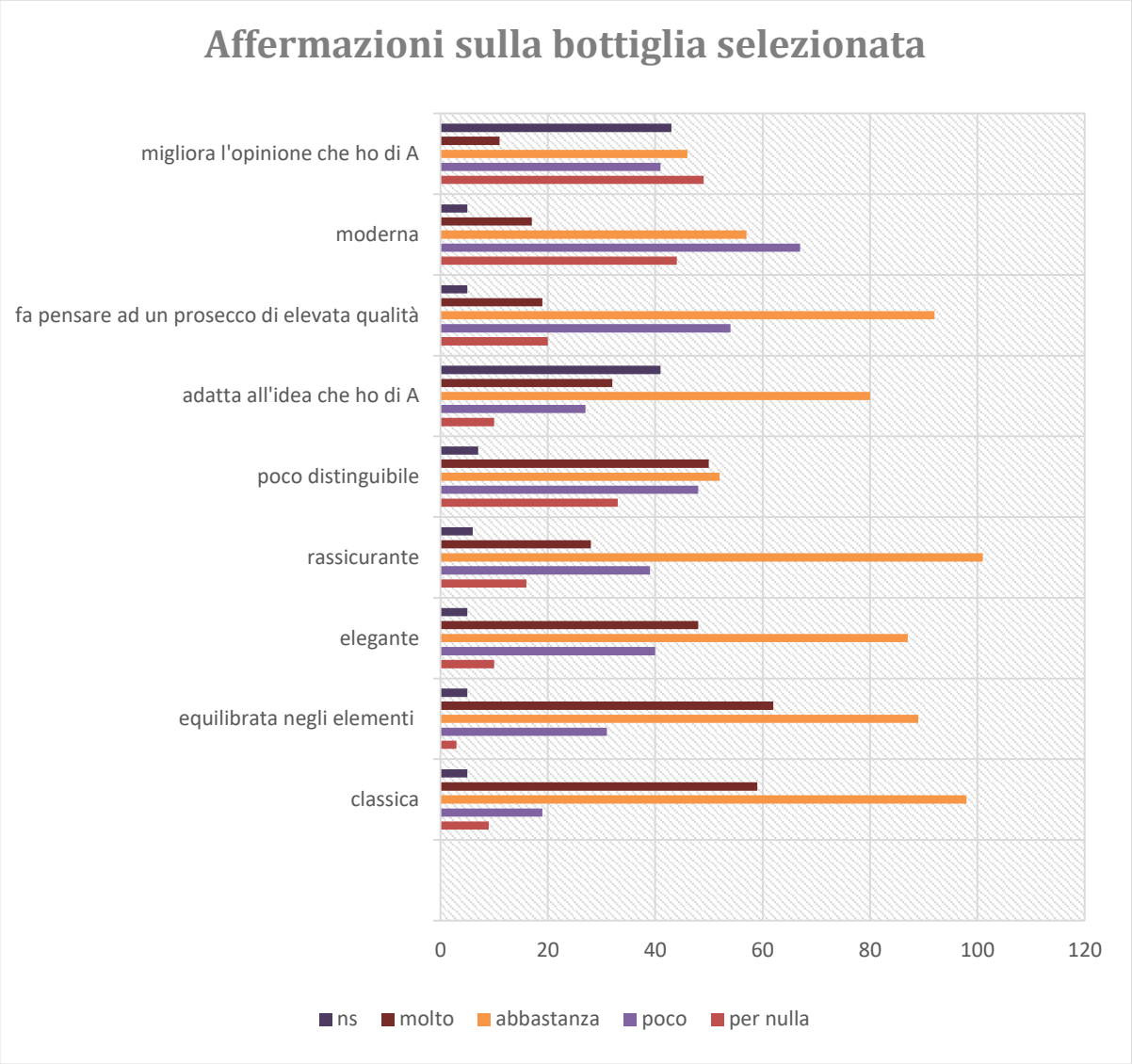

Grafico 2.21 – Dati sulle risposte alla domanda: Quanto la seguente bottiglia rispecchia le seguenti affermazioni?

Si può notare di seguito come la bottiglia selezionata sia "rassicurante", "elegante", "equilibrata", "classica", "adatta all'idea che ho di A", e "fa pensare ad un prosecco di alta qualità"; mentre risulta contrastante l'opinione riguardo alle variabili "poco

distinguibile" e "migliora l'opinione che ho di A"; infine risulta chiaro come la bottiglia selezionata non sia moderna (come dimostrano le variabili poco e per nulla rispetto alle altre). Risulta tuttavia di maggiore utilità dividere questo grafico in base alle diverse bottiglie scelte in modo tale da poter comprendere meglio l'impatto di ognuna:

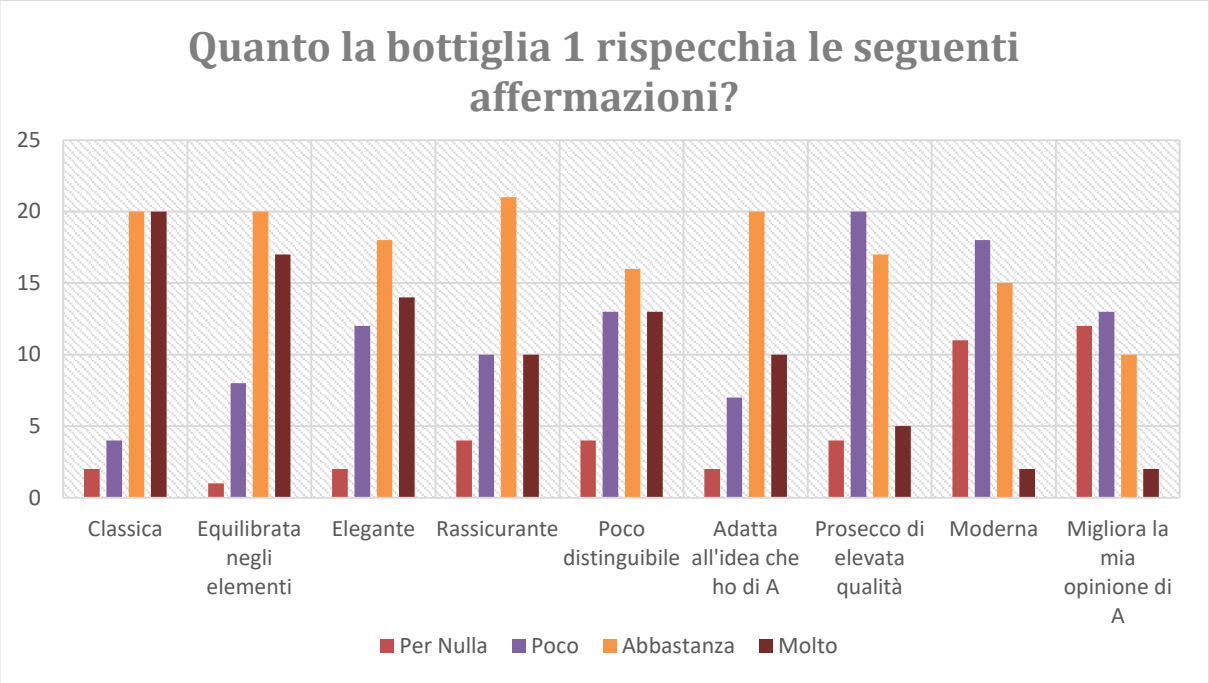

Grafico 2.22- Dati sulle risposte relative alla domanda "Quanto la seguente bottiglia rispecchia le seguenti affermazioni?"

La bottiglia viene percepita come classica, equilibrata, elegante, rassicurante e adatta all'idea che l'intervistato ha di A. Poco chiare risultano le opinioni riguardo la distinguibilità della bottiglia, infatti seppure vi sia un alto numero di "poco", vi è anche un alto numero di "molto" e "abbastanza" e in questo caso è opportuno tenerne conto. La bottiglia risulta non essere percepita come prosecco di elevata qualità e non migliora l'idea che gli intervistati hanno di A. Considerando questi dati insieme a quelli del grafico 2.10 pare chiaro come la bottiglia 1 sia la meno adatta a sostituire la bottiglia esistente; tuttavia è opportuno attendere la fine dell'analisi per essere certi di tale affermazione.

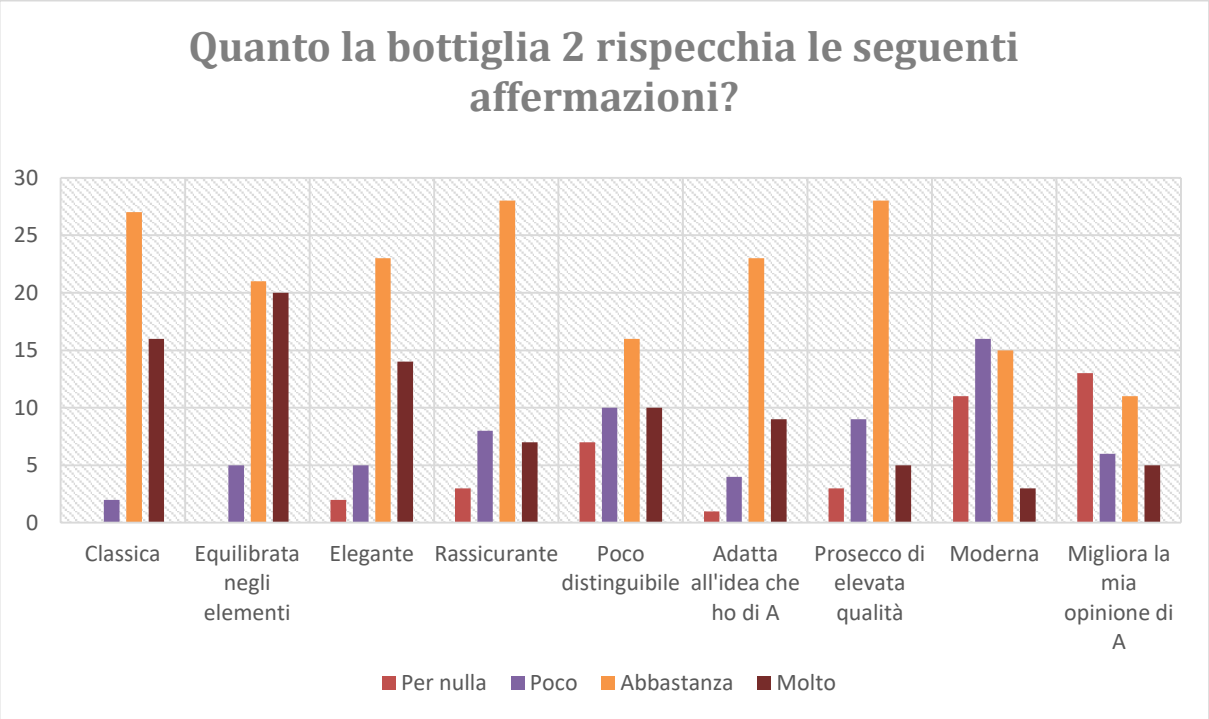

Grafico 2.23- Dati sulle risposte relative alla domanda "Quanto la seguente bottiglia rispecchia le seguenti affermazioni?"

La bottiglia 2 risulta agli occhi degli intervistati classica, equilibrata, elegante, rassicurante, adatta all'idea che gli intervistati hanno di A, di elevata qualità; tuttavia risulta anche poco distinguibile, poco moderna e non migliora l'idea che gli intervistati hanno di A. In confronto alla bottiglia 1 risulta essere più distinguibile (visto il più basso numero di "molto" e "abbastanza") e più "di elevata qualità"; infine risulta migliorare maggiormente l'opinione di A.

Considerati i valori del grafico 2.10 in concomitanza a quelli del grafico soprastante si evince che in generale la bottiglia 2 risulta più idonea a sostituire la bottiglia attualmente in commercio rispetto alla bottiglia 1.

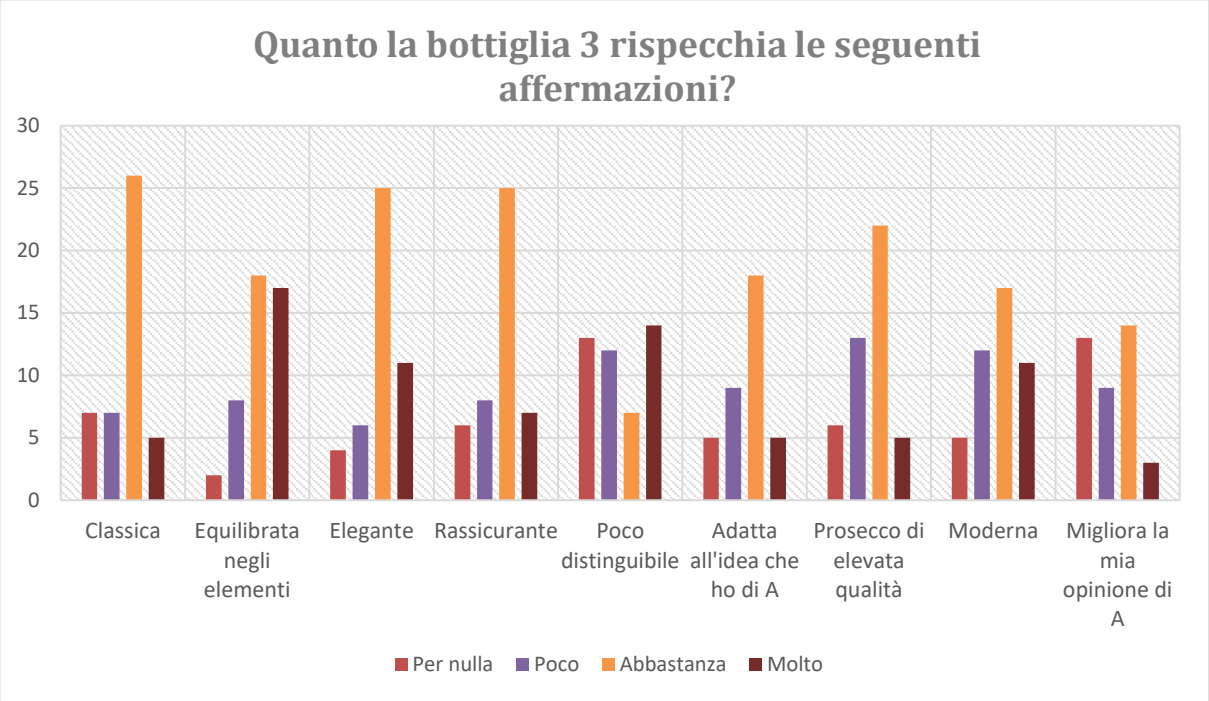

Grafico 2.24- Dati sulle risposte relative alla domanda "Quanto la seguente bottiglia rispecchia le seguenti affermazioni?"

La bottiglia 3 è considerata dagli intervistati classica, equilibrata, elegante, rassicurante, adatta all'idea che gli intervistati hanno di A e relativamente moderna. Interessante notare che rispetto alle due bottiglie precedenti risulta quella più distinguibile e che migliora maggiormente l'idea che gli intervistati hanno di A, anche se tuttavia i valori di "poco" e "per nulla" è molto alto, ciò è un dato da tenere in considerazione in quanto anche se si scegliesse la bottiglia 3 bisognerebbe essere al corrente che andrebbero comunque attuati dei cambiamenti (nel design o nella comunicazione) per migliorare la reputazione dell'azienda.
Considerando il grafico soprastante insieme al grafico 2.10 appare evidente che la bottiglia 3 risulta la più adatta a sostituire la bottiglia attualmente sul mercato, seppure è necessario effettuare tutte le altre analisi per essere certi di tale affermazione.

Al fine di attuare una buona analisi comparativa è infine opportuno considerare i dati della bottiglia attualmente sul mercato.

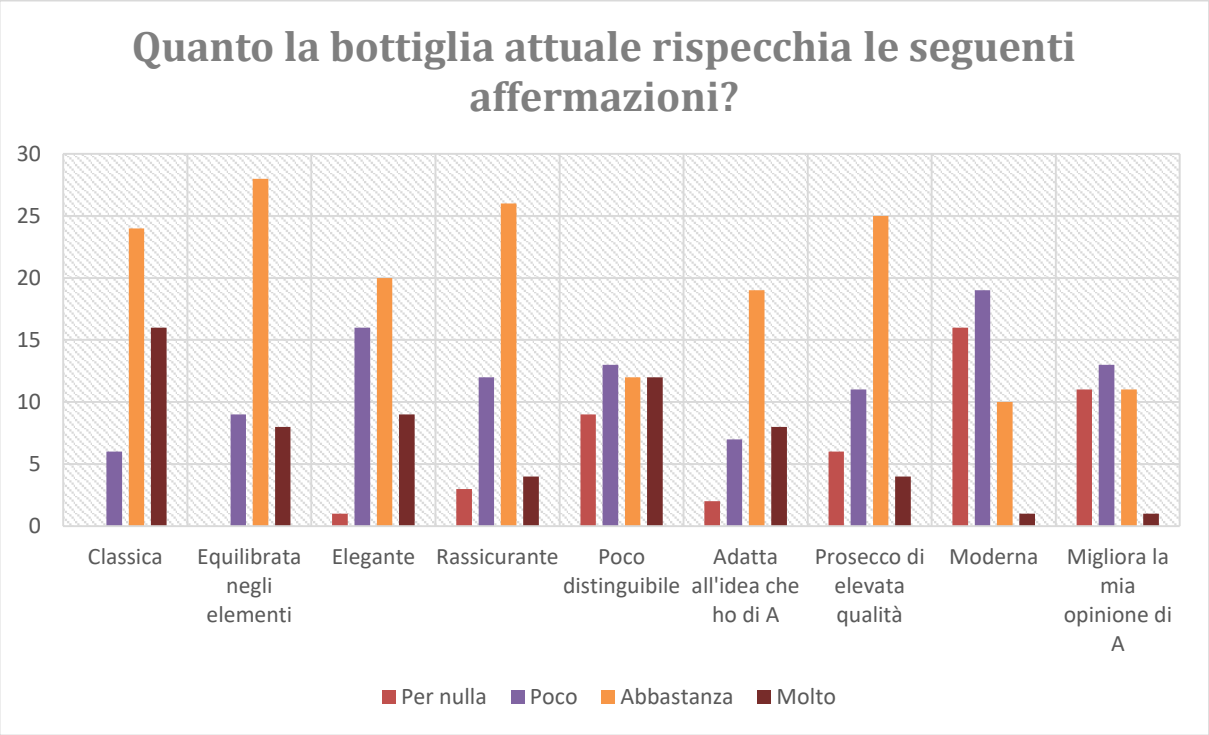

Grafico 2.25- Dati sulle risposte relative alla domanda "Quanto la seguente bottiglia rispecchia le seguenti affermazioni?"

L'analisi della tabella soprastante è interessante a causa di alcuni elementi che la distinguono dalle bottiglie elencata precedentemente. I valori della distinguibilità della bottiglia sono i più bassi tra le quattro bottiglie esaminate come anche per i valori di "migliora la mia opinione di A" che insieme alla bottiglia 1 sono i più bassi tra le 4 bottiglie; tuttavia esaminando i valori delle quattro bottiglie con attenzione si nota che le differenze non sono cosi ampie, ciò significa che l'immaginario dell'azienda A dipende dal design della bottiglia ma non viene migliorato soltanto da esso. Cosi come anticipato precedentemente l'azienda A dovrebbe soffermarsi, oltre al design, su altri aspetti in modo tale da migliorare la sua immagine.

2.4 CONCLUSIONI

Ricapitolando, ciò che evince da questa prima analisi è il profilo generale degli intervistati: chi sono, dove vivono, che preferenze hanno, quale prezzo sono disposti a pagare, quanto spesso bevono vino o prosecco ecc.

Dall'analisi descrittiva del questionario siamo quindi riusciti a conoscere meglio il consumatore, a delineare le sue caratteristiche e, nello specifico, ad apprendere che:

* Gli intervistati sono prevalentemente padovani e prevalentemente donne (tabelle…
* La maggior parte ha un'età superiore ai 60 anni o compresa tra i 30 e i 40
* Il livello di istruzione è alto
* Il consumo di vino prevale rispetto a quello di prosecco
* Il principale motivo dell'acquisto e del consumo è quello per occasioni particolari
* La disponibilità a pagare è generalmente alta.
* Tra i tre design proposti la bottiglia numero tre risulta la preferita, inoltre confrontando le tre bottiglie con quella attualmente in commercio la maggior parte degli intervistati afferma che la nuova bottiglia aumenta la propensione all'acquisto
* Il design della bottiglia di marca A attualmente sul mercato (A Attuale) risulta essere la meno preferita rispetto ai competitors e alle nuove proposte. Tale dato, insieme ai dati al punto precedente, suggerisce la necessità di un cambiamento di design. Il design che sembra più adatto a sostituire la bottiglia attualmente sul mercato sembra essere, da queste prime analisi, la bottiglia 3 in quanto tra le opzioni è quella più distinguibile, e quella che migliora maggiormente l'opinione che gli intervistati hanno dell'azienda. Tuttavia l'immagine dell'azienda tende ad aumentare di poco a seguito di un cambiamento di design, ciò implica che insieme a quest'ultimo è opportuno attuare altre azioni per migliorare l'immagine percepita dai consumatori.
* Il prezzo associato ad A e A Attuale è molto basso, ciò conferma che il design incide notevolmente sulla percezione di qualità e conseguentemente sul prezzo
* Le variabili estetiche che influenzano maggiormente la scelta della bottiglia sono la forma della bottiglia, l'etichetta nel complesso e la scritta prosecco.

3 CAPITOLO III
L'ANALISI BIVARIATA

Per avere una migliore conoscenza sui fenomeni oggetto di studio appare riduttivo considerare una sola variabile e fermarsi all'analisi descrittiva dei dati senza analizzare le possibili relazioni tra esse, infatti la sola analisi descrittiva non permette di cogliere degli aspetti critici dell'analisi come la dipendenza o l'interdipendenza tra due proprietà.

Per tale ragione in questo capitolo affronterò la tematica dell'analisi bivariata, cercando di analizzare ed evidenziare le relazioni più significative tra le varie domande presenti nel questionario e cercando di trarne informazioni più ricche ed interessanti.

Come anticipato nel capitolo precedente questo tipo di analisi, partendo dall'analisi descrittiva, studia una coppia di variabili contemporaneamente per ricercare dipendenze e correlazioni. Operativamente, per coppie di variabili si intende l'insieme costituito da due colonne del dataset.

La tecnica di studio della relazione tra le variabili e la modalità di definizione il loro grado di associazione dipendono dalla natura delle stesse. Con il termine associazione si fa riferimento al legame tra le variabili in oggetto, intendendo che alla variazione di una di esse corrispondano variazioni dell'altra; diventa quindi l'opposto di indipendenza (assoluta mancanza di legami).

È opportuno precisare che la scelta delle variabili da associare è un procedimento che dipende dalle conoscenze del ricercatore e da ciò che si vuole evidenziare, risulta quindi un processo soggettivo. Una volta chiarito il concetto di indipendenza possiamo passare alla fase successiva, ovvero testare tale indipendenza in modo tale da comprendere la relazione che lega due variabili.

3.1 L'ANALISI BIVARIATA DEL QUESTIONARIO[7]

Per questa analisi sono state scelte le coppie di variabili secondo gli accoppiamenti che sembravano più consoni e adatti.

La prima analisi si basa sulla ricerca di associazione tra le variabili sesso e consumo di vino; di seguito sono presentate le tabelle delle frequenze osservate e teoriche:

- Associazione ***Sesso-Consumo settimanale di vino***

	Quante volte alla settimana beve vino			
Sesso	**1 volta a sett.**	**più volte a sett.**	**tutti i giorni**	**Totale**
Maschi	23	34	26	83
Femmine	50	32	25	107
Totale	73	66	51	190

Tabella 3.1 Frequenze teoriche delle variabili sesso e consumo di vino settimanale

Frequenza Teorica	**1 volta a settimana**	**più volte a settimana**	**tutti i giorni**	**Totale**
Maschi	31,88947368	28,83157895	22,27894737	83
Femmine	41,11052632	37,16842105	28,72105263	107
Totale	73	66	51	190

Tabella 3.2 Frequenze teoriche delle variabili Sesso e consumo settimanale di vino

Valore del chi quadro (probabilità) = 0,028029388

$0{,}028029388 < \alpha$

[7] N.B. Nl paragrafo verranno elencate solo le tabelle più significative, per le rimanenti consultare l'Appendice C

Rifiuto Ho, verifico che il test è significativo e affermo che vi è un'associazione tra le due variabili; ovvero vi è dipendenza tra il sesso e il consumo settimanale di vino, nello specifico dalle frequenze teoriche osserviamo una ampia differenza rispetto a quelle osservate, molto evidente è il caso "una volta a settimana" che dimostra quanto le donne bevano vino meno frequentemente rispetto agli uomini.

- Associazione ***Motivo d'acquisto-Prezzo***

Successivamente è sembrato opportuno indagare sull'associazione tra prezzo e motivo d'acquisto; in questo caso si cerca di capire se il motivo d'acquisto influenza le scelte sul prezzo. Di seguito sono presentate le tabelle riguardanti le frequenze osservate e attese:

	Prezzo				
Motivo d'acquisto	***sotto i 4€***	***da 4 a 5€***	***da 5 a 6 €***	***oltre 6 €***	***Totale***
Consumo abituale	7	12	19	16	54
Occasioni particolari	6	32	37	34	109
Regalo	1	4	3	8	16
Totale	14	48	59	58	179

Tabella 3.3 Frequenze osservate rispetto alle variabili prezzo e motivo d'acquisto

Frequenza Teorica	***sotto i 4€***	***da 4 a 5€***	***da 5 a 6 €***	***oltre 6 €***	***Totale***
Consumo abituale	4,223463687	14,48044693	17,79888268	17,4972067	54
Occasioni particolari	8,525139665	29,22905028	35,9273743	35,31843575	109
Regalo	1,251396648	4,290502793	5,273743017	5,184357542	16
Totale	14	48	59	58	179

Tabella 3.4 Frequenze teoriche delle variabili Prezzo e motivo d'acquisto

Valore del chi quadro (probabilità) = 0,408685539

0,408685539 > α=0,05

Accetto Ho e affermo che non vi è dipendenza, ovvero non vi è una differenza significativa tra le differenze osservate e quelle attese.

Questo dato è molto interessante, perché non vi è una relazione tra le due variabili, ciò significa che il prezzo della bottiglia non è influenzato dall'occasione al contrario di quanto ci aspetteremmo (es pagare un prezzo molto più alto per una bottiglia da regalare piuttosto che per una da consumo abituale).

- Associazione ***Età-Occasioni di consumo di prosecco***

Altra associazione che sembra corretto fare è quella tra le fasce d'età e le varie occasioni di consumo, per comprendere se in qualche modo vi è un'associazione tra le due.

	Età					
Occasioni	**19-30**	**31-40**	**41-50**	**51-60**	**60+**	**Totale**
Mai		2	4		2	8
Festa	20	24	23	13	32	112
Aperitivo	8	13	9	11	11	52
Pasti	2	2	4	3	5	16
Totale	30	41	40	27	50	188

Tabella 3.5 Frequenze osservate delle variabili Età e Occasioni di consumo di prosecco

F. Teorica	19-30	31-40	41-50	51-60	60+	Totale
Mai	1,276595745	1,744680851	1,70212766	1,148936	2,12766	8
Festa	17,87234043	24,42553191	23,82978723	16,08511	29,78723	112
Aperitivo	8,29787234	11,34042553	11,06382979	7,468085	13,82979	52
Pasti	2,553191489	3,489361702	3,404255319	2,297872	4,255319	16
Totale	30	41	40	27	50	188

Tabella 3.6 Frequenze teoriche delle variabili età e occasioni di consumo di prosecco

Valore del chi quadro (probabilità) = 0,762445

0,762445 > α=0,05

Accetto Ho in quanto il valore del chi quadro è maggiore rispetto al valore di α.

Vediamo che quindi, come nel caso precedente non vi è associazione tra età e occasioni di consumo, come invece ci si aspetterebbe.

- Associazione tra ***Bottiglia preferita*** (rispetto ai competitors)-***Prezzo associato***

Questo è un caso particolare perché le unità non sono più gli individui, bensì le scelte ,in quanto sono state accorpate le due colonne del dataset corrispondenti alle due scelte per la bottiglia preferita, per tale ragione N risulta il doppio rispetto ai casi precedentemente trattati.

Di seguito sono riportate le tabelle delle frequenze:

	Prezzo						
Bottiglia Preferita	***4,5***	***5***	***5,5***	***6***	***6,5***	***7***	***Totale***
A	13	3	2	7	2	9	36
A Attuale	4	9	3	1	4	5	26
B	5	4	6	21	15	28	79
C	6	12	11	10	9	16	64
D	9	6	8	11	9	28	71
E	9	9	6	19	8	23	74
Totale	46	43	36	69	47	109	350

Tabella 3.7 Frequenze osservate rispetto alle variabili Bottiglia preferita e Prezzo associato

Frequenza Teorica	***4,5***	***5***	***5,5***	***6***	***6,5***	***7***	***Totale***
A	4,731429	4,422857	3,702857	7,097143	4,834286	11,21143	36
A Attuale	3,417143	3,194286	2,674286	5,125714	3,491429	8,097143	26
B	10,38286	9,705714	8,125714	15,57429	10,60857	24,60286	79
C	8,411429	7,862857	6,582857	12,61714	8,594286	19,93143	64
D	9,331429	8,722857	7,302857	13,99714	9,534286	22,11143	71
E	9,725714	9,091429	7,611429	14,58857	9,937143	23,04571	74
Totale	46	43	36	69	47	109	350

Tabella 3.8 Frequenze teoriche delle variabili Bottiglie preferite e prezzo associato

Valore del chi quadro (probabilità) = 0,000322979

0,000322979 < α=0,05

Rifiuto Ho in quanto il valore del chi quadro è nettamente minore rispetto ad α (il test è altamente significativo in quanto il valore è largamente minore α), ciò dimostra

un'alta associazione tra le due variabili, ovvero una dipendenza tra la scelta della bottiglia e il prezzo, in quanto vi è una forte differenza tra le frequenze osservate e quelle teoriche.

Tale considerazione risulta molto evidente se andiamo ad analizzare i totali marginali del prezzo osservato. Risalta subito all'occhio la prevalenza di scelte, circa il 64%, che riguardano un prezzo alto (da 6€ in su) a differenza delle scelte riguardanti un prezzo basso (dai 4,5€ ai 6€), circa il 36%. In aggiunta dato interessante è quello del prezzo più alto (7€) che da solo costituisce circa il 31% delle scelte.

Dopo tale analisi si può affermare con sicurezza che alla bottiglia scelta come preferita è stato spesso associato un alto prezzo, ovvero la bottiglia considerata più pregiata e di alto valore e che quindi le due variabili sono dipendenti.

- Associazione tra ***Bottiglia preferita*** (rispetto ai competitors) - ***Caratteristiche della bottiglia***

In questo caso l'analisi si svolgerà in base alle caratteristiche della bottiglia, anche in questo caso sono state accorpate le due colonne del dataset riguardanti le due scelte sia per la bottiglia preferita che per le caratteristiche in quanto le domande permettevano una seconda scelta.

Diversamente da quanto mostrato in precedenza il procedimento attuato in questo caso sarà quello completo mostrato al paragrafo 3.2.

Saranno presentate cinque coppie di tabelle, una per ogni caratteristica:

- esteticamente bella
- moderna e attuale
- prosecco da acquistare
- prosecco da offrire
- prosecco di elevata qualità

Dopo aver capito che il consumatore sceglie come preferita la bottiglia che considera più "pregiata" sembra ovvio cercare di analizzare quali sono le variabili associate a tali bottiglie, ovvero quali sono le bottiglie considerate "esteticamente belle", "moderne e attuali", "da acquistare", "da offrire" e "di elevata qualità" e incrociare tali scelte con le bottiglie preferite in modo da ricercare quali variabili hanno portato alla scelta.

* **La prima caratteristica considerata è "Esteticamente bella"**; di seguito le tabelle sulle frequenze osservate e teoriche:

	Esteticamente bella						
Preferenze	**A**	**A Attuale**	**B**	**C**	**D**	**E**	**TOT**
A	26		4		3	4	37
A Attuale	1	10		1	1	4	17
B	3	1	43	9	10	12	78
C	1		3	21	19	14	58
D	1	1	3	2	34	10	51
E	1	4		3	7	37	52
TOT	33	16	53	36	74	81	293

Tabella 3.9 Frequenze osservate rispetto alle variabili Bottiglia preferita e la caratteristica esteticamente bella

Frequenze teoriche	A	A Attuale	B	C	D	E	TOT
A	4,167235	2,020478	6,692833	4,546075	9,34471	10,22867	37
A Attuale	1,914676	0,928328	3,075085	2,088737	4,293515	4,699659	17
B	8,784983	4,259386	14,10922	9,583618	19,69966	21,56314	78
C	6,532423	3,167235	10,49147	7,12628	14,64846	16,03413	58
D	5,744027	2,784983	9,225256	6,266212	12,88055	14,09898	51
E	5,856655	2,83959	9,406143	6,389078	13,13311	14,37543	52
TOT	33	16	53	36	74	81	293

Tabella 3.10 Frequenze teoriche rispetto alle variabili Bottiglia preferita e la caratteristica esteticamente bella

$Vc=46,93 < x^2= 443,93$

Rifiuto Ho, ciò significa che le due variabili sono associate e che la caratteristica "esteticamente bella" ha influito sulla scelta della bottiglia preferita, ovvero è probabile che coloro che hanno scelto come preferita una certa bottiglia lo abbiano fatto in base all'esteticità e alla bellezza della stessa, questo concetto implica che il packaging è tenuto in considerazione durante il processo di scelta.

* **La seconda caratteristica è "Moderna e attuale"**

	Moderna e attuale						
Preferenze	**A**	**A ATTUALE**	**B**	**C**	**D**	**E**	**Totale**
A	**15**		**4**	**4**	**9**	**4**	**36**
A Attuale	**1**	**5**			**5**	**4**	**15**
B	**10**	**1**	**22**	**8**	**21**	**15**	**77**
C	**3**	**3**	**8**	**12**	**6**	**23**	**55**
D	**3**	**3**	**5**	**2**	**17**	**17**	**47**
E	**1**	**5**	**1**	**1**	**13**	**20**	**41**
Totale	**33**	**17**	**40**	**27**	**71**	**83**	**271**

Tabella 3.11 Frequenze osservate rispetto alle variabili Bottiglia preferita e la caratteristica moderna e attuale

Frequenze Teoriche	A	A ATTUALE	B	C	D	E	Totale
A	4,383763838	2,2583026	5,3137	3,5867	9,4317	11,026	**36**
A Attuale	1,826568266	0,9409594	2,214	1,4945	3,9299	4,5941	**15**
B	9,376383764	4,8302583	11,365	7,6716	20,173	23,583	**77**
C	6,697416974	3,4501845	8,1181	5,4797	14,41	16,845	**55**
D	5,723247232	2,9483395	6,9373	4,6827	12,314	14,395	**47**
E	4,992619926	2,5719557	6,0517	4,0849	10,742	12,557	**41**
Totale	**33**	**17**	**40**	**27**	**71**	**83**	**271**

Tabella 3.12 Frequenze teoriche rispetto alle variabili Bottiglia preferita e la caratteristica Moderna e attuale

$Vc=46,93 < x^2= 110,5633108$

Rifiuto Ho, ciò significa che le due variabili sono associate e che la caratteristica "moderna e attuale" ha influito sulla scelta della bottiglia preferita, ovvero è probabile che coloro che hanno scelto come preferita una certa bottiglia lo abbiano fatto in base alla modernità della stessa.

* **La terza caratteristica in esame è "Prosecco che acquisterei"**

	Prosecco che acquisterei						
Preferenze	**A**	**A ATTUALE**	**B**	**C**	**D**	**E**	**Totale**
A	**22**	**1**	**4**	**3**	**4**	**2**	**36**
A Attuale	**2**	**8**	**1**	**1**		**2**	**14**
B	**6**	**1**	**48**	**11**	**6**	**5**	**77**
C	**3**	**5**	**5**	**30**	**9**	**2**	**54**
D	**2**	**3**	**3**	**6**	**26**	**6**	**46**
E	**1**	**6**	**2**	**5**	**6**	**28**	**48**
Totale	**36**	**24**	**63**	**56**	**51**	**45**	**275**

Tabella 3.13 Frequenze osservate rispetto alle variabili Bottiglia preferita e la caratteristica Prosecco che acquisterei

Frequenze teoriche	A	A ATTUALE	B	C	D	E	Totale
A	4,712727273	3,141818182	8,24727	7,33091	6,67636	5,8909	**36**
A Attuale	1,832727273	1,221818182	3,20727	2,85091	2,59636	2,2909	**14**
B	10,08	6,72	17,64	15,68	14,28	12,6	**77**
C	7,069090909	4,712727273	12,3709	10,9964	10,0145	8,8364	**54**
D	6,021818182	4,014545455	10,5382	9,36727	8,53091	7,5273	**46**
E	6,283636364	4,189090909	10,9964	9,77455	8,90182	7,8545	**48**
Totale	**36**	**24**	**63**	**56**	**51**	**45**	**275**

Tabella 3.14 Frequenze teoriche rispetto alle variabili Bottiglia preferita e la caratteristica Prosecco che acquisterei

$Vc=46,93 < x^2= 343,9$

Rifiuto Ho, ciò significa che le due variabili sono associate e che la caratteristica "prosecco che acquisterei" ha influito sulla scelta della bottiglia preferita, ovvero è probabile che gli intervistati abbiano scelto come preferita una bottiglia che avrebbero acquistato.

* **La quarta caratteristica considerata è "Prosecco che offrirei"**

	Prosecco che offrirei						
Preferenze	**A**	**A ATTUALE**	**B**	**C**	**D**	**E**	**Totale**
A	**15**	**2**	**6**	**3**	**3**	**8**	**37**
A Attuale	**1**	**3**	**1**	**1**	**3**	**3**	**12**
B	**5**		**40**	**16**	**11**	**6**	**78**
C	**1**	**1**	**8**	**21**	**12**	**8**	**51**
D	**2**	**2**	**6**	**7**	**19**	**7**	**43**
E		**3**	**2**	**9**	**11**	**23**	**48**
Totale	**24**	**11**	**63**	**57**	**59**	**55**	**269**

Tabella 3.15 Frequenze osservate rispetto alle variabili Bottiglia preferita e la caratteristica Prosecco che offrirei

Frequenze teoriche	A	A ATTUALE	B	C	D	E	Totale
A	3,3011	1,513	8,6654	7,84	8,1152	7,565	**37**
A Attuale	1,0706	0,4907	2,8104	2,543	2,632	2,454	**12**
B	6,9591	3,1896	18,268	16,53	17,108	15,95	**78**
C	4,5502	2,0855	11,944	10,81	11,186	10,43	**51**
D	3,8364	1,7584	10,071	9,112	9,4312	8,792	**43**
E	4,2825	1,9628	11,242	10,17	10,528	9,814	**48**
Totale	**24**	**11**	**63**	**57**	**59**	**55**	**269**

Tabella 3.16 Frequenze teoriche rispetto alle variabili Bottiglia preferita e la caratteristica Prosecco che offrirei

$Vc=46,93 < x^2= 160,08$

Rifiuto Ho, ciò significa che le due variabili sono associate e che la caratteristica "prosecco che offrirei" ha influito sulla scelta della bottiglia preferita, ovvero è probabile che gli intervistati abbiano scelto come preferita una bottiglia che avrebbero offerto.

* **L'ultima caratteristica studiata è Prosecco di alta qualità**

	Prosecco di alta qualità						
Preferenze	**A**	**A ATTUALE**	**B**	**C**	**D**	**E**	**Totale**
A	**17**	**3**	**8**	**3**	**2**	**1**	**34**
A Attuale	**1**	**3**	**2**		**2**		**8**
B	**4**	**5**	**47**	**11**	**6**	**2**	**75**
C	**2**	**3**	**9**	**28**	**4**	**3**	**49**
D	**1**	**4**	**6**	**7**	**23**	**3**	**44**
E	**1**	**2**	**5**	**7**	**9**	**15**	**39**
Totale	**26**	**20**	**77**	**56**	**46**	**24**	**249**

Tabella 3.17 Frequenze osservate rispetto alle variabili Bottiglia preferita e la caratteristica Prosecco di elevata qualità

Frequenze Teoriche	A	A ATTUALE	B	C	D	E	Totale
A	**17**	**3**	**8**	**3**	**2**	**1**	**34**
A Attuale	**1**	**3**	**2**		**2**		**8**
B	**4**	**5**	**47**	**11**	**6**	**2**	**75**
C	**2**	**3**	**9**	**28**	**4**	**3**	**49**
D	**1**	**4**	**6**	**7**	**23**	**3**	**44**
E	**1**	**2**	**5**	**7**	**9**	**15**	**39**
Totale	**26**	**20**	**77**	**56**	**46**	**24**	**249**

Tabella 3.18 Frequenze teoriche rispetto alle variabili Bottiglia preferita e la caratteristica Prosecco di elevata qualità

$Vc=46,93 < x^2= 218,25$

Rifiuto Ho, ciò significa che le due variabili sono associate e che la caratteristica "prosecco di elevata qualità" ha influito sulla scelta della bottiglia preferita, ovvero è

probabile che gli intervistati abbiano scelto come preferita una bottiglia che considerano di elevata qualità.

3.2 CONCLUSIONI

Questo capitolo si è concentrato sullo studio delle relazioni tra due variabili e delle loro relazioni, ciò ha aiutato a delineare le caratteristiche degli individui e a comprendere, in parte, quali sono i fattori che determinano specifiche scelte e inoltre quali siano le variabili che non presentano un grado di interdipendenza, come invece ci si aspetterebbe.

In breve ciò che pare evidente dallo studio di questo capitolo è che:

* Vi è dipendenza tra il consumo di vino e il sesso e nello specifico risulta che le donne consumano prosecco meno frequentemente degli uomini (tabelle 2.3.1.1 e 2.3.1.2), ciò conferma ciò che pare evidente anche senza l'utilizzo di tale analisi
* Non vi è associazione tra la propensione alla spesa e motivo d'acquisto (tabelle 2.3.1.3 e 2.3.1.4), come invece avremmo potuto supporre (es: un prezzo più alto per feste e regali e un prezzo più basso per consumo abituale)
* Non vi è dipendenza tra le occasioni di consumo e l'età degli intervistati (tabelle 2.3.1.5 e 2.3.1.6). Come nel punto precedente anche in questo caso sembrerebbe scontata un'associazione di tale tipo invece l'analisi rigetta tale supposizione.
* Vi è dipendenza tra la bottiglia scelta come preferita tra le 6 alternative presentate (cinque in commercio e una tra le nuove proposte) e il prezzo associato alla rispettiva bottiglia, ciò dimostra che la scelta della bottiglia preferita è ricaduta sulla bottiglia considerata più pregiata e di alto valore (tabelle 2.3.1.7 e 2.3.1.8)
* La scelta delle bottiglie preferite è stata influenzata da 5 caratteristiche (esteticamente bella, moderna e attuale, prosecco da acquistare, prosecco da offrire, prosecco di elevata qualità) come mostrano le tabelle dalla 2.3.1.9 alla 2.3.1.18.

* Infine che consumatori in generale sono attenti all'estetica della bottiglia ma anche alla qualità del contenuto e che generalmente per tali prodotti sono disposti a pagare un alto prezzo.

4 CAPITOLO IV
ANALISI MULTIVARIATA

Terminata l'analisi descrittiva e bivariata possiamo concentrarci su una tipologia di analisi più complessa: l'analisi multivariata, ovvero l'analisi di più variabili considerate contemporaneamente. Lo scopo di questo capitolo è quello di evidenziare le relazioni più significative tra le variabili e comprendere come e quanto queste (variabili *predictors*) influiscano sulla variabile *response* (variabile dipendente) in modo tale da poter delineare un modello in grado di prevedere il comportamento di consumatori di cui non si hanno i dati.

Per questo tipo di studio è stato necessario passare ad uno strumento più completo e adatto all'analisi statistica: Gretl. Per prima cosa il database è stato adattato in modo tale da poter essere inserito all'interno del programma e successivamente trasferito all'interno di Gretl; secondariamente è stata individuata la variabile dipendente attorno alla quale costruire il modello.

Lo scopo è analizzare la relazione tra la scelta della bottiglia e le altre variabili, per cui consideriamo la domanda relativa alla scelta della bottiglia preferita (tra le nuove proposte) come variabile response (Y) in modo tale da comprendere quali siano i fattori che influiscono sulle preferenze dei consumatori. Una volta definita la variabile Y è possibile iniziare a costruire il modello.

4.1 LA COSTRUZIONE DEL MODELLO

Come spiegato nel primo capitolo, il primo passo dell'analisi riguarda la selezione delle variabili da considerare all'interno del modello, in modo tale da poter successivamente studiare la loro relazione. Definita la variabile dipendente, procediamo alla selezione delle variabili esplicative, che nel nostro caso risultano essere le informazioni generali sui consumatori e le informazioni sulle loro abitudini (vengono quindi escluse dall'analisi le domande sulle preferenze):

Forma dell'etichetta
Colori dell'etichetta
Etichetta nel complesso
Scritta prosecco
Fascia sul collo della bottiglia
Forma della bottiglia
Consumo settimanale di vino
Ultimo acquisto di vino
Occasioni di consumo di prosecco
Motivo d'acquisto di Prosecco
Ultimo acquisto di prosecco
Disponibilità a pagare
Sesso
Età
Titolo di studio
Città

A questo punto ci troviamo ad avere un modello di base contenente 16 variabili predictors; è necessario quindi operare una selezione delle variabili più significative in modo tale da semplificare il modello.

Per tale obiettivo è utile procedere con una Backwards Stepwise, inserendo tutte le variabili all'interno del modello ed eliminando una alla volta quelle con un p-value non

statisticamente rilevante (maggiore di 0,05) fino ad arrivare al modello migliore. Completato tale procedimento il miglior modello risulta essere composto dalle seguenti variabili:

Consumo settimanale di vino
Disponibilità a pagare
Città
Età
Istruzione
Ultimo acquisto di vino
Forma dell'etichetta
Etichetta nel complesso
Scritta Prosecco
Fascia sul collo della bottiglia
Forma della bottiglia

4.2 IL MODELLO FINALE

Selezionate le variabili da analizzare simultaneamente è possibile procedere con l'analisi vera e propria. Tale analisi come anticipato nel primo capitolo, considerata la natura dei dati, è effettuata attraverso l'uso del multinomial logit model. Tale procedura può essere implementata da diversi strumenti, in questo caso per praticità si è preferito utilizzare GRETL.

L'output, dopo un'iniziale stepwise manuale (attraverso l'eliminazione delle variabili meno significative) è espresso nella tabella sottostante. Alla tabella è stata successivamente aggiunta la colonna degli odds ratio calcolati separatamente in un foglio Excel:

Modello 47: Logit multinomiale, usando le osservazioni 1-179
Variabile dipendente: d3a
Errori standard QML

	Coefficiente	***Errore Std.***	***z***	***p-value***		***Odds Ratio (expβ)***
Bottiglia Preferita 2						
Prezzo (<4,00 Euro)	**−2,321**	0,942	−2,463	0,0138	**	**0,098**
Prezzo (4<p<5)	−1,513	0,762	−1,984	0,0473	**	0,22
Prezzo (5<p<6)	−1,493	0,784	−1,905	0,0568	*	0,225
Città (Padova)	**1,687**	0,653	2,585	0,0097	***	**5,403**
Consumo di vino (più volte a settimana)	1,851	0,712	2,598	0,0094	***	6,366
Consumo di vino (tutti i giorni)	**2,694**	0,917	2,938	0,0033	***	**14,79**
Ultimo acquisto di vino (molto tempo fa)	1,958	0,942	2,078	0,0377	**	**7,085**
Ultimo acquisto di vino (mai)	**−16,629**	1,466	−11,34	<0,0001	***	6E-08
Istruzione (diploma superiore)	1,923	0,895	2,147	0,0318	**	6,841
Istruzione (basso livello)	**2,379**	0,994	2,393	0,0167	**	**10,79**
Età(19-30)	−0,674	0,769	−0,8765	0,3807		0,51
Età(31-40)	−0,999	0,791	−1,262	0,2070		0,368
Età(41-50)	**0,719**	0,903	0,7966	0,4257		**2,052**
Forma dell'etichetta	−0,373	0,277	−1,349	0,1774		0,689

Etichetta nel complesso	0,357	0,361	0,9894	0,3224		1,429
Scritta Prosecco	**−0,741**	0,277	−2,670	0,0076	***	0,477
Fascia sul collo della bottiglia	**−0,817**	0,297	−2,742	0,0061	***	0,442
Forma della bottiglia	**0,562**	0,277	2,024	0,0429	**	1,754
Bottiglia preferita 3						
Prezzo (<4,00 Euro)	**−2,173**	0,921	−2,359	0,0183	**	**0,114**
Prezzo (4<p<5)	−0,806	0,685	−1,177	0,2394		0,447
Prezzo (5<p<6)	−0,695	0,627	−1,107	0,2682		0,499
Città (Padova)	**1,505**	0,536	2,808	0,0050	***	**4,504**
Consumo di vino (più volte a settimana)	1,385	0,600	2,307	0,0210	**	3,995
Consumo di vino (tutti i giorni)	**2,455**	0,871	2,817	0,0048	***	**11,65**
Ultimo acquisto di vino (molto tempo fa)	0,833	0,882	0,9454	0,3445		2,3
Ultimo acquisto di vino (mai)	**−16,927**	1,381	−12,26	<0,0001	***	4E-08
Istruzione (diploma superiore)	1,591	0,851	1,868	0,0617	*	4,909
Istruzione (basso livello)	**2,550**	0,979	2,605	0,0092	***	**12,81**
Età(19-30)	−1,475	0,671	−2,197	0,0280	**	0,229
Età(31-40)	−1,256	0,733	−1,713	0,0867	*	0,285

Età(41-50)	−0,467	0,840	−0,5565	0,5778		0,627
Forma dell'etichetta	−0,713	0,255	−2,798	0,0051	***	0,49
Etichetta nel complesso	0,720	0,333	2,160	0,0308	**	2,054
Scritta Prosecco	−0,855	0,270	−3,165	0,0016	***	0,425
Fascia sul collo della bottiglia	−0,766	0,247	−3,096	0,0020	***	0,465
Forma della bottiglia	1,008	0,266	3,783	0,0002	***	2,74

Media var. dipendente	2,307263	SQM var. dipendente	0,828226
Log-verosimiglianza	−132,0020	Criterio di Akaike	336,0040
Criterio di Schwarz	450,7498	Hannan-Quinn	382,5325

Numero dei casi 'previsti correttamente' = 117 (65,4%)
Test del rapporto di verosimiglianza: Chi-quadro(36) = 129,299 [0,0000]

4.2.1 INTERPRETAZIONE DEI COEFFICIENTI E DEGLI ODDS RATIO

Per poter estrapolare informazioni utili dai dati appena elencati è essenziale decodificare i coefficienti e gli odds ratio. Per praticità verrà fornita un'interpretazione dei valori maggiormente significativi e statisticamente rilevanti.

I ***Coefficienti*** esprimono una rilevanza pratica del valore tra le variabili predictors e la variabile response: Il valore ***Constant*** β_0, ovvero il valore medio di Y quando $X = 0$ in questo caso non è presente all'interno del modello perché non statisticamente rilevante, mentre i valori seguenti rappresentano i valori di β_i. L'informazione da estrapolare in questo caso è il segno, in quanto questi esprimono l'impatto che le rispettive variabili predictors hanno sulla variabile dipendente Y. Un segno positivo

esprime un impatto positivo sulla variabile Y, viceversa un segno negativo esprime un impatto negativo sulla Y e di conseguenza sui log odds. Per approfondire l'argomento e avere dati più precisi è utile analizzare gli Odds Ratio (expβ). Per decodificare gli ***Odds Ratio*** dobbiamo, a differenza dei coefficienti, analizzare se questi siano maggiori di uno, uguali ad uno o minori di uno.

Va precisato inoltre che i coefficienti vanno analizzati rispetto alle categorie di riferimento per ogni variabile (eccetto per le ultime cinque variabili).

Categorie di riferimento:

- Prezzo: prezzo alto (<6 euro)
- Consumo di vino: una volta la settimana
- Città: Milano
- Ultimo acquisto di vino: recentemente (nell'ultimo mese)
- Età: fascia di età dai 51 ai 60+
- Titolo di studio: Laurea

Nello specifico quello che esprimono questi coefficienti è:

- ***Prezzo***:

  Bottiglia 2: Per la categoria "prezzo minore di 4 euro" il segno risulta negativo, ciò significa che (considerando la categoria di riferimento prezzo alto) chi ha una bassa disponibilità a pagare generalmente preferisce la bottiglia 1 alla 2. Interessante notare come anche per le altre disponibilità a pagare il segno rimanga negativo, ciò implica che anche i consumatori con una maggiore disponibilità a pagare sceglieranno più probabilmente la bottiglia 1 rispetto alla 2. Volendo approfondire tali affermazioni possiamo esaminare gli Odds ratio che nel caso del prezzo minore di 4 euro risultano molto vicini allo zero. Matematicamente questo valore si traduce come segue:

  $$OR = {Odds\ Prezzo\ basso\ (4)}\Big/{Odds\ prezzo\ alto\ (7)} = 0{,}098$$

  Ciò significa che gli odds di scegliere la bottiglia 2 piuttosto che la 1 per coloro che hanno una bassa disponibilità a pagare sono veramente bassi;

Tuttavia gli odds di scegliere la bottiglia 2 screscono al crescere della disponibilità (come dimostrano i valori degli OR successivi)

Bottiglia 3: Anche in questo caso i coefficienti sono negativi, ciò implica che (rispetto alla categoria di riferimento) chi ha una bassa, media o medio alta disponibilità a pagare tende a preferire la bottiglia 1 alla 3. Analizzando gli OR anche in questo caso sono molto vicini allo zero, indicando che la tendenza a preferire la 1 alla 3 è molto forte; tale tendenza tuttavia, come nel caso precedente, si affievolisce al crescere della disponibilità a pagare.

- ***Città***:

  Bottiglia 2: Il segno del coefficiente è positivo, ciò implica che I padovani hanno più probabilità di scegliere la bottiglia 1 piuttosto che la 2 rispetto ai milanesi. Tale probabilità risulta essere rafforzata se si esaminano gli OR, infatti gli odds che un padovano scelga la bottiglia 2 piuttosto che la 1 sono 5,4 volte maggiori rispetto ad un milanese.

  Bottiglia 3: Scenario pressoché identico si verifica anche per la bottiglia 3, preferita rispetto alla 1 dai Padovani con un odds di 4,5.

- ***Consumo di vino***:

  Bottiglia 2: cloro che consumano vino tutti i giorni (rispetto a coloro che lo consumano una volta a settimana) sono più propensi a scegliere la bottiglia numero 2 piuttosto che la uno, stesso vale per coloro che consumano vino più volte a settimana. Interessante notare come gli odds ratio aumentino all'aumentare del consumo, infatti, gli odds per coloro che consumano vino tutti i giorni sono 14,8 volte maggiori rispetto a chi lo consuma una volta a settimana, più del doppio rispetto a coloro che lo consumano più volte a settimana.

  Bottiglia 3: Situazione simile si ripresenta anche per la bottiglia 3, preferita alla 1 da parte di coloro che hanno un alto consumo di vino. In questo caso gli odds ratio per coloro che consumano vino tutti i giorni sono 11,65 volte maggiori rispetto a chi consuma vino una sola volta a settimana. Rimane valida la considerazione fatta precedentemente sull' aumento del valore degli OR

all'aumento del consumo con la differenza che l'aumento è di circa il triplo rispetto a coloro che consumano vino più volte alla settimana.

- ***Ultimo acquisto di vino***:

  Bottiglia 2: Dato molto interessante riguarda coloro che non hanno mai acquistato vino (-16,62) che dimostra la tendenza di questa fascia (a differenza della categoria di riferimento: coloro che hanno acquistato recentemente) a preferire la bottiglia numero 1, al contrario di coloro che hanno acquistato vino molto tempo fa che al contrario tendono a preferire la bottiglia 2. Volendo approfondire gli OR per coloro che non hanno mai acquistato molto tempo fa rispetto a che ha acquistato recentemente sono molto vicini allo zero, ciò dimostra una bassissima probabilità che questa fascia di consumatori preferisca la bottiglia 2. Al contrario del caso precedente gli OR per coloro che hanno acquistato molto tempo fa sono 7,08 volte maggiori rispetto a coloro che hanno acquistato vino recentemente, ciò dimostra una forte propensione alla scelta della bottiglia 2. Tale divario dimostra un drastico cambiamento nel gusto rispetto alla frequenza di acquisto.

  Bottiglia 3: Simile scenario si ripropone per la scelta numero 3: coloro che non hanno mai acquistato tendono a preferire la bottiglia numero 1 con un OR molto vicino allo zero. Al contrario, coloro che hanno acquistato molto tempo fa preferiscono invece la 3 con un OR di 2,3 maggiore rispetto a coloro che hanno acquistato vino recentemente.

- ***Istruzione:***

  Bottiglia 2: Gli intervistati aventi un diploma superiore (rispetto alla categoria di riferimento: laureati) tendono a preferire la bottiglia 2 alla 1. Tale tendenza è maggiore per coloro che hanno un basso livello di istruzione, come dimostrano i coefficienti e gli odds ratio.

  Bottiglia 3: Scenario simile si ripresenta per la bottiglia 3, preferita alla 1 dagli intervistati aventi un diploma e da coloro aventi un basso livello d'istruzione. La differenza rispetto al caso precedente sta negli OR, infatti gli odds di coloro che

hanno un basso livello di istruzione sono 14 volte maggiori rispetto a chi possiede una laurea; 3 volte maggiori rispetto a coloro aventi un diploma.

- ***Età:***

  Bottiglia *2:* Dai coefficienti dell'età si evince che rispetto alla categoria di riferimento (età dai 51 ai 60+) i due gruppi di età inferiore tendono a preferire la bottiglia 1 a differenza della fascia d'età 41-50 che al contrario tende a preferire la bottiglia numero 2. Interessante notare come all'aumento dell'età aumentino anche gli odds ratio ciò significa che gli odds di coloro che sono nella fascia d'età compresa tra i 41-51 anni saranno 2,05 volte maggiori rispetto alla categoria di riferimento (51-60+) ovvero circa il quadruplo rispetto alle due fasce d'età minori.

  Bottiglia 3: A differenza della scelta due i coefficienti delle tre classi di età risultano essere tutti negativi, ciò sottolinea la tendenza di tali classi, rispetto alla classe di riferimento (51-60+), a preferire la bottiglia 1 alla 3, l'analisi degli odds ratio ci informa inoltre che gli odds che venga preferita la bottiglia 3 sono molto bassi; ciò sottolinea una forte probabilità che la bottiglia 1 venga preferita alla tre dalle tre classi d'età elencate precedentemente rispetto alla classe d'età maggiore.

- ***Forma dell'etichetta:***

  Per entrambe le scelte chi ha dato un'alta importanza alla forma dell'etichetta ha tendenzialmente preferito quella della bottiglia numero 1

- ***Etichetta nel complesso:***

  Bottiglia 2: Chi ha dato un'alta importanza all'etichetta nel complesso ha tendenzialmente preferito la bottiglia 2 alla 1.

  Bottiglia 3: Chi ha dato un'alta importanza all'etichetta nel complesso ha generalmente preferito la bottiglia 3 alla 1, con un odds 2,052 volte maggiore rispetto a coloro che hanno dato una bassa importanza.

- ***Scritta prosecco:***

  Per entrambe le scelte coloro che hanno dato un'alta importanza alla scritta prosecco tendenzialmente hanno preferito la bottiglia 1

- ***Fascia sul collo della bottiglia:***

  Per entrambe le scelte coloro che hanno dato un'alta importanza alla fascia sul collo della bottiglia tendenzialmente hanno preferito la bottiglia 1

- ***Forma della bottiglia:***

  Bottiglia 2: Coloro che hanno dato un'alta importanza alla forma della bottiglia hanno generalmente preferito la bottiglia 2 alla 1 con un odds ratio di 1,75 rispetto a coloro che hanno dato una bassa importanza

  Bottiglia 3: Coloro che hanno dato un'alta importanza alla forma della bottiglia hanno generalmente preferito la bottiglia 3 alla 1 con un odds ratio di 2,74 rispetto a coloro che hanno dato una bassa importanza

4.2.2 IL CALCOLO DEGLI EFFETTI IMMAGINARI E DEGLI ODDS

A questo punto, dopo aver calcolato e analizzato i valori dei coefficienti possiamo procedere sostituendo i valori più significativi all'interno della funzione $\hat{p}$ in modo tale da analizzare le probabilità stimate del modello:

<u>Bottiglia 2 (x=1)</u>

$$\widehat{p_1} = \frac{e^{0+PD+c.giornaliero+ultimo\ acq.molto\ tempo\ fa+basso\ liv.d'istr.}}{1+e^{0+PD+c.giornaliero+ultimo\ acq.molto\ tempo\ fa+basso\ liv.d'istr.}} = 0{,}999$$

Ciò significa che la probabilità di scegliere la bottiglia 2 essendo di Padova, avendo un consumo di vino giornaliero, avendo effettuato una bottiglia di vino molto tempo fa e avendo un basso livello d'istruzione è del 99,9%.

Bottiglia 2 (x=0)

$$\widehat{p_0} = \frac{e^0}{1 + e^0} = 0{,}5$$

Ciò significa che la probabilità di scegliere la bottiglia 2 per coloro che sono di Milano, che consumano vino una volta a settimana, che hanno acquistato vino recentemente e che sono laureati è del 50%

Effetti Marginali

$\hat{p}_1 - \hat{p}_0 = 0{,}9998 - 0{,}5 = 0{,}4998$

La differenza tra le due probabilità è circa del 50%

Odds

$$Odds\ che\ la\ bottiglia\ 2\ venga\ scelta = \frac{\hat{p}_1}{1 - \hat{p}_1} = \frac{0{,}9998}{1 - 0{,}9998} = 6111{,}943$$

$$Odds\ che\ la\ bottiglia\ 2\ non\ venga\ scelta = \frac{\widehat{p}_1}{1 - \widehat{p}_1} = \frac{0{,}5}{1 - 0{,}5} = 1$$

<u>Bottiglia 3 (x=1)</u>

$$\widehat{p_1} = \frac{e^{0 + PD + c.giornaliero + ultimo\ acq.molto\ tempo\ fa + basso\ liv.d'istr.}}{1 + e^{0 + PD + c.giornaliero + ultimo\ acq.molto\ tempo\ fa + basso\ liv.d'istr.}} = 0{,}999$$

Ciò significa che la probabilità di scegliere la bottiglia 3 essendo di Padova, avendo un consumo di vino giornaliero, avendo effettuato una bottiglia di vino molto tempo fa e avendo un basso livello d'istruzione è del 99,9%.

Bottiglia 3 (x=0)

$$\widehat{p_0} = \frac{e^0}{1 + e^0} = 0{,}5$$

Ciò significa che la probabilità di scegliere la bottiglia 3 per coloro che sono di Milano, consumano vino una volta a settimana, che hanno acquistato vino recentemente e che sono laureati è del 50%.

Effetti Marginali

$\hat{p}_1 - \hat{p}_0 = 0{,}999 - 0{,}5 = 0{,}4993$

La differenza tra le due probabilità è circa del 50%

Odds

$$Odds\ che\ la\ bottiglia\ 3\ venga\ scelta = \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{0{,}9993}{1-0{,}9993} = 1545{,}943$$

$$Odds\ che\ la\ bottiglia\ 2\ non\ venga\ scelta = \frac{\hat{p}_1}{1-\hat{p}_1} = \frac{0{,}5}{1-0{,}5} = 1$$

Analizziamo ora alcuni valori significativi singolarmente:

<u>Consumo giornaliero di vino</u>

Bottiglia 2:

$\widehat{\boldsymbol{p}} = \frac{e^{2,69}}{1+e^{2,69}} = 0{,}936$ $\qquad \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0{,}5$ $\qquad \widehat{\boldsymbol{p}} - \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0{,}436$

$$\boldsymbol{Odds} = \frac{0{,}936}{1-0{,}936} = 14{,}9$$

Ciò significa che la probabilità che la bottiglia 2 venga scelta al posto della 1 da coloro che consumano vino giornalmente sono del 93%, con un odds di 14,9. Ovviamente se calcolassimo l'odds per coloro che bevono vino una volta a settimana e facessimo il rapporto tra i due odds otterremmo l'odds ratio ($^{14,9}/_{1} = 14{,}9$)

Bottiglia 3:

$\widehat{\boldsymbol{p}} = \frac{e^{2,45}}{1+e^{2,45}} = 0{,}920$ $\qquad \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0{,}5$ $\qquad \widehat{\boldsymbol{p}} - \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0{,}42$

$\boldsymbol{Odds} = \frac{0{,}92}{1-0{,}92} = 11{,}64$

La probabilità che la bottiglia 3 venga scelta al posto della 1 da coloro che consumano vino giornalmente sono del 92%, con un odds di 11,6.

<u>Prezzo basso</u>

Bottiglia 2:

$\widehat{\boldsymbol{p}} = \frac{e^{0,43}}{1+e^{0,43}} = 0{,}60$ $\qquad \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0{,}5$ $\qquad \widehat{\boldsymbol{p}} - \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0{,}10$

$$\boldsymbol{Odds} = \frac{0{,}60}{1-0{,}60} = 1{,}53$$

La probabilità che chi abbia una bassa disponibilità a pagare scelga la bottiglia 1 piuttosto che la 2 è del 60% rispetto a coloro con un'alta disponibilità, con un odds di 5,3.

Bottiglia 3:

$\widehat{\boldsymbol{p}} = \frac{e^{0,46}}{1+e^{0,46}} = 0,61$ $\qquad \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0,5$ $\qquad \widehat{\boldsymbol{p}} - \widehat{\boldsymbol{p}}_{\boldsymbol{0}} = 0,11$

$\boldsymbol{Odds} = \frac{0,61}{1-0,61} = 1,58$

La probabilità che una persona preferisca la bottiglia 1 alla 3 avendo una bassa disponibilità a pagare sono del 61% rispetto a coloro con un'alta disponibilità a pagare, con un odds di 1,58.

<u>Forma della bottiglia</u>

In questo, diversamente dalle analisi precedenti i valori che può assumere la variabile non sono più binari bensì variano da 1 a 4 a seconda dell'importanza che viene data alla forma della bottiglia (1-per nulla importante; 4-molto importante)

Bottiglia 2:

$\widehat{\boldsymbol{p}}(x = 4) = 0,90$ $\quad \widehat{\boldsymbol{p}}(x = 1) = 0,63$ $\quad \widehat{\boldsymbol{p}}(x = 4) - \widehat{\boldsymbol{p}}(x = 1) = 0,267$

$\boldsymbol{Odds}(\boldsymbol{x = 4}) = 9,46$ $\qquad \boldsymbol{Odds}\ (\boldsymbol{x = 1}) = 1,75$ $\quad OR = 5,39$[8]

La probabilità che una persona preferisca la bottiglia 2 alla 1 avendo assegnato la massima importanza alla forma della bottiglia è del 90%, ovvero 5,39 volte maggiore rispetto a chi invece ha dato la minima importanza alla forma della bottiglia.

Bottiglia 3:

$\widehat{\boldsymbol{p}}(x = 4) = 0,98$ $\quad \widehat{\boldsymbol{p}}(x = 1) = 0,73$ $\quad \widehat{\boldsymbol{p}}(x = 4) - \widehat{\boldsymbol{p}}(x = 1) = 0,23$

$\boldsymbol{Odds}(\boldsymbol{x = 4}) = 56,37$ $\qquad \boldsymbol{Odds}\ (\boldsymbol{x = 1}) = 2,74$ $\quad OR = 20,57$

La probabilità che una persona preferisca la bottiglia 3 alla 1 avendo assegnato il massimo punteggio alla forma della bottiglia è del 98%,

[8] N.B. Questo OR rappresenta la differenza tra gli odds di coloro che hanno dato il massimo punteggio alla forma della bottiglia e coloro che invece hanno dato il minimo, al contrario del valore nella tabella che invece esprime la differenza in un'unità di cambiamento

ovvero 20,6 volte maggiore rispetto a chi invece ha assegnato alla forma della bottiglia il minimo punteggio.

4.2.3 CONCLUSIONI

Questo capitolo si è concentrato sulla costruzione di un buon modello che possa spiegare la relazione tra la variabile dipendente e quelle indipendenti e sull'interpretazione dei coefficienti successivamente generati. A seguito dell'analisi appena effettuata è necessario sintetizzare le informazioni critiche e trarre alcune conclusioni in modo tale da convertire i valori numerici in decisioni strategiche di marketing per poter così limitare l'incertezza con cui si svolge l'attività imprenditoriale.

Ad una prima analisi è subito evidente che il processo di scelta della bottiglia è maggiormente influenzato da variabili quali il consumo di vino, l'ultimo acquisto di vino (e non di prosecco come invece ci si potrebbe aspettare) istruzione e variabili riguardanti l'estetica del packaging; mentre meno rilevanti risultano variabili che sembrerebbero centrali come l'età o il prezzo.

Le bottiglie 2 e 3 vengono preferite alla bottiglia 1 da coloro che hanno un basso livello di istruzione e un alto consumo di vino con una probabilità del 99%. Questi valori risultano fondamentali per le scelte di posizionamento e targeting.

Considerando la variabile età si evince che la bottiglia 1 viene preferita alla 2 e alla 3 dalle fasce d'età minori, al contrario delle fasce maggiori che tendono a preferire la bottiglia 2 e 3. In questo caso, come nel precedente, la scelta della bottiglia da lanciare sul mercato dipende dal target di riferimento a cui si vuole rivolgere l'azienda.

Per effettuare delle scelte riguardanti il posizionamento è necessario esaminare i valori riguardanti il prezzo, dai quali risulta chiaramente che gli aventi un'alta disponibilità a pagare tendono a preferire la bottiglia 3 al contrario della 1 che è invece preferita da coloro che hanno una bassa disponibilità a pagare. Ancora una volta la scelta della bottiglia dipende dalla volontà dell'azienda di mantenere il posizionamento attuale o di migliorarlo.

L’analisi degli aspetti estetici esprime con chiarezza le preferenze degli intervistati, permettendo all’azienda di esaminare l’importanza di ogni singolo aspetto collegato alle 3 bottiglie considerate.

Forma dell’etichetta preferita: bottiglia 1
Etichetta nel complesso preferita: bottiglia 3
Scritta prosecco preferita: bottiglia 1
Fascia sul collo preferita: bottiglia 1
Forma della bottiglia preferita: bottiglia 3

In generale l’azienda dovrebbe, a seconda delle sue esigenze, del target di riferimento e del posizionamento desiderato, operare le sue scelte in maniera strategica, inoltre, secondo queste indicazioni l’azienda dovrebbe rivedere il design raggruppando gli aspetti preferiti per creare una nuova bottiglia da lanciare sul mercato.

CONCLUSIONI

Il ruolo del marketing è quello di supportare il management durante la decisione di immettere un nuovo prodotto sul mercato, attraverso modelli sempre più accurati, che permettono l'identificazione delle variabili rilevanti durante il processo di scelta.

Il presente lavoro di tesi ha cercato di implementare le suddette procedure, attraverso l'utilizzo di metodologie statistiche, per analizzare l'importanza del design durante i processi di scelta del prosecco.

Dall' analisi descrittiva del **secondo capitolo** emerge il profilo generale degli intervistati: chi sono, dove vivono, che preferenze hanno, quale prezzo sono disposti a pagare, quanto spesso bevono vino o prosecco ecc. Siamo quindi riusciti a conoscere meglio il consumatore, a delineare le sue caratteristiche e, nello specifico, ad apprendere che: Gli intervistati sono prevalentemente padovani e prevalentemente donne, la maggior parte ha un'età superiore ai 60 anni o compresa tra i 30 e i 40, Il livello di istruzione è alto, il consumo di vino prevale rispetto a quello di prosecco, il principale motivo dell'acquisto e del consumo è quello per occasioni particolari, la disponibilità a pagare è generalmente alta.

Tra i tre design proposti la bottiglia numero tre risulta la preferita, inoltre confrontando le tre bottiglie con quella attualmente in commercio la maggior parte degli intervistati afferma che la nuova bottiglia aumenta la propensione all'acquisto

Il design della bottiglia di marca A attualmente sul mercato (A Attuale) risulta essere la meno preferita rispetto ai competitors e alle nuove proposte. Tale dato, insieme ai dati al punto precedente, suggerisce la necessità di un cambiamento di design. Il design che sembra più adatto a sostituire la bottiglia attualmente sul mercato sembra essere, da queste prime analisi, la bottiglia 3 in quanto tra le opzioni è quella più distinguibile, e quella che migliora maggiormente l'opinione che gli intervistati hanno dell'azienda.

Tuttavia l'immagine dell'azienda tende ad aumentare di poco a seguito di un cambiamento di design, ciò implica che insieme a quest'ultimo è opportuno attuare altre azioni per migliorare l'immagine percepita dai consumatori.

Il prezzo associato ad A e A Attuale è molto basso, ciò conferma che il design incide notevolmente sulla percezione di qualità e conseguentemente sul prezzo

Le variabili estetiche che influenzano maggiormente la scelta della bottiglia sono la forma della bottiglia, l'etichetta nel complesso e la scritta prosecco.

Dall'analisi più approfondita effettuata nel **terzo capitolo** è stato possibile comprendere, in parte, quali sono i fattori che determinano specifiche scelte e inoltre quali siano le variabili che non presentano un grado di interdipendenza, come invece ci si aspetterebbe.

In breve ciò che pare evidente dallo studio di questo capitolo è che:

Vi è dipendenza tra il consumo di vino e il sesso e nello specifico risulta che le donne consumano prosecco meno frequentemente degli uomini, ciò conferma ciò che pare evidente anche senza l'utilizzo di tale analisi.

Non vi è associazione tra la propensione alla spesa e motivo d'acquisto, come invece avremmo potuto supporre (es: un prezzo più alto per feste e regali e un prezzo più basso per consumo abituale)

Non vi è dipendenza tra le occasioni di consumo e l'età degli intervistati. Come nel punto precedente anche in questo caso sembrerebbe scontata un'associazione di tale tipo invece l'analisi rigetta tale supposizione.

Vi è dipendenza tra la bottiglia scelta come preferita tra le 6 alternative presentate (cinque in commercio e una tra le nuove proposte) e il prezzo associato alla rispettiva bottiglia, ciò dimostra che la scelta della bottiglia preferita è ricaduta sulla bottiglia considerata più pregiata e di alto valore

La scelta delle bottiglie preferite è stata influenzata da 5 caratteristiche (esteticamente bella, moderna e attuale, prosecco da acquistare, prosecco da offrire, prosecco di elevata qualità).

Infine che consumatori in generale sono attenti all'estetica della bottiglia ma anche alla qualità del contenuto e che generalmente per tali prodotti sono disposti a pagare un alto prezzo.

Il **quarto capitolo** si è concentrato sulla costruzione di un buon modello che possa spiegare la relazione tra la variabile dipendente e quelle indipendenti e sull'interpretazione dei coefficienti successivamente generati. È opportuno a tal punto sintetizzare le informazioni critiche e trarre alcune conclusioni in modo tale da convertire i valori numerici in decisioni strategiche di marketing per poter così limitare l'incertezza con cui si svolge l'attività imprenditoriale è necessario.

È evidente che il processo di scelta della bottiglia è maggiormente influenzato da variabili quali il consumo di vino, l'ultimo acquisto di vino (e non di prosecco come invece ci si potrebbe aspettare) istruzione e variabili riguardanti l'estetica del packaging; mentre meno rilevanti risultano variabili che sembrerebbero centrali come l'età o il prezzo.

Le bottiglie 2 e 3 vengono preferite alla bottiglia 1 da coloro che hanno un basso livello di istruzione e un alto consumo di vino con una probabilità del 99%. Questi valori confermano ciò che era già stato anticipato nei capitoli precedenti e risultano fondamentali per le scelte di posizionamento e targeting.

Dall'analisi del prezzo risulta che gli aventi un'alta disponibilità a pagare tendono a preferire la bottiglia 3 al contrario della 1 che è invece preferita da coloro che hanno una bassa disponibilità a pagare.

L'analisi degli aspetti estetici esprime con chiarezza le preferenze degli intervistati, permettendo all'azienda di esaminare l'importanza di ogni singolo aspetto collegato alle 3 bottiglie considerate grazie al quale si può "costruire" la bottiglia perfetta rivedendo il design raggruppando i vari aspetti preferiti per creare una nuova bottiglia da lanciare sul mercato; in questo caso: la forma della bottiglia e l'etichetta nel

complesso della bottiglia 3 insieme alla fascia sul collo e alla scritta prosecco della bottiglia 1.

In generale l'azienda dovrebbe operare le sue scelte strategiche partendo dalla presente analisi adattandola a seconda delle sue esigenze, del target di riferimento e del posizionamento desiderato.

Completata l'analisi è possibile rispondere ai quesiti che hanno reso l'analisi necessaria; nello specifico:

* Come è posizionata la bottiglia attualmente sul mercato rispetto ai competitors?
  La bottiglia è la meno preferita rispetto ai competitors, risulta inoltre avere i punteggi più bassi anche per le variabili "prosecco che acquisterei", "prosecco che offrirei" e "prosecco ad elevata qualità" ciò espone la necessità di un cambio repentino di design.
* Il design della bottiglia influenza l'acquisto?
  Si. Risultano infatti molto rilevanti ai fini dell'acquisto le variabili estetiche quali la forma della bottiglia, dell'etichetta, la scritta prosecco ecc.
* Con quale bottiglia andrebbe eventualmente sostituita la bottiglia attualmente sul mercato?
  È stato più volte specificato e verificato che la bottiglia più adatta risulta essere la numero 3, tuttavia per un risultato migliore sarebbe consigliabile creare un nuovo design che comprenda gli aspetti preferiti di ogni bottiglia.
* La nuova bottiglia migliorerebbe il posizionamento e l'immagine dell'azienda?
  Si. Tuttavia è necessario affiancare degli ulteriori cambiamenti per un'efficacia maggiore, quali campagne di comunicazione atte a migliorare l'immagine percepita dai consumatori.
* Quali politiche di pricing applicare?
  La bottiglia attualmente sul mercato, insieme alle 3 nuove proposte, risulta avere un bassissimo posizionamento per quanto riguarda il pricing, sicuramente non a livello dei concorrenti. La disponibilità a pagare della maggior parte dei consumatori risulta essere inferiore al prezzo medio dei prodotti venduti dall'azienda. Ciò richiede ancora una volta oltre ad un cambio di design (che

comunichi alta qualità) un'azione di comunicazione atta a migliorare l'immagine dell'azienda la reputazione e di seguito la disponibilità a pagare dei consumatori.

La presente analisi ha apportato numerosi dati interessanti per l'azienda e utili ad indirizzare le decisioni strategiche e a diminuire l'incertezza del management nello svolgimento delle sue attività. Sarebbe senza dubbio interessante, in seguito, effettuare un'ulteriore analisi per monitorare gli effettivi cambiamenti apportati dall'inserimento sul mercato del nuovo design.

APPENDICE

Appendice A: TABELLE SULLE INFORMAZIONI RISPETTO ALLE PREFERENZE

* Grafico 2.10

Bottiglia Preferita 1°posto	n
1	47
2	42
3	101
Tot	190

Bottiglia preferita 2° posto	n
1	49
2	101
3	40
Tot	190

Bottiglia preferita 3° posto	n
1	94
2	47
3	49
Tot	190

* Grafico 2.11

Con la nuova bottiglia la propensione all'acquisto di prosecco "A"	n
Aumenta	117
Rimane invariata	50
Diminuisce	18
Ns	4
Tot	189

* Grafico 2.12

Quali tra queste bottiglie preferisce?	1° scelta	2° scelta
A	37	1
B	72	7
C	39	25
D	29	46
E	9	70
A Attuale	4	23
Ns	0	18
Totale	190	190

* Grafico 2.13

Quanto hanno inciso questi elementi sulla scelta della bottiglia preferita?	Per nulla	Poco	Abbastanza	Molto	Ns	Tot
Forma dell'etichetta	55	33	66	31	5	190
Colori dell'etichetta	34	32	72	47	5	190
Etichetta nel complesso	19	17	99	50	5	190
Scritta prosecco	35	29	70	49	7	190
Fascia sul collo della bottiglia	42	29	54	59	6	190
Forma della bottiglia	22	19	62	85	2	190

* Grafico 2.19

Quali di queste bottiglie rispecchia queste affermazioni? (prima scelta)	A	B	C	D	E	A Attuale	Ns	Tot
Esteticamente bella	33	49	23	49	34	1	1	190
Moderna	33	39	22	49	44	2	1	190
Prosecco che acquisterei	36	57	35	29	18	11	4	190
Prosecco da offrire	25	59	38	37	29	2	0	190
Prosecco di elevata qualità	27	70	34	22	12	14	11	190

Quali di queste bottiglie rispecchia queste affermazioni? (seconda scelta)	A	B	C	D	E	A Attuale	Ns	Tot
Esteticamente bella	0	5	13	26	49	15	82	190
Moderna	0	3	5	22	42	16	102	190
Prosecco che acquisterei	0	7	26	23	27	13	94	190
Prosecco da offrire	0	4	20	24	27	9	106	190
Prosecco di elevata qualità	0	8	22	27	14	6	113	190

* Grafico 2.20

Prezzo	n
Minore di 4 euro	15
Tra 4 e 5 euro	48
Tra 5 e 6 euro	59
Oltre 6	58
NS/NR	10
Tot	190

* Grafico 2.22

	Classica	Equilibrata negli elementi	Elegante	Rassicurante	Poco distinguibile	Adatta all'idea che ho di A	Prosecco di elevata qualità	Moderna	Migliora la mia opinione di A
Per Nulla	2	1	2	4	4	2	4	11	12
Poco	4	8	12	10	13	7	20	18	13
Abbastanza	20	20	18	21	16	20	17	15	10
Molto	20	17	14	10	13	10	5	2	2
NS	0	0	0	1	0	7	0	0	9
Tot	46	46	46	46	46	46	46	46	46

* Grafico 2.23

	Classica	Equilibrata negli elementi	Elegante	Rassicurante	Poco distinguibile	Adatta all'idea che ho di A	Prosecco di elevata qualità	Moderna	Migliora la mia opinione di A
Per nulla	0	0	2	3	7	1	3	11	13
Poco	2	5	5	8	10	4	9	16	6
Abbastanza	27	21	23	28	16	23	28	15	11
Molto	16	20	14	7	10	9	5	3	5
NS	1	0	2	0	3	9	1	1	11
Tot	46	46	46	46	46	46	46	46	46

* Grafico 2.24

	Classica	Equilibrata negli elementi	Elegante	Rassicurante	Poco distinguibile	Adatta all'idea che ho di A	Prosecco di elevata qualità	Moderna	Migliora la mia opinione di A
Per nulla	7	2	4	6	13	5	6	5	13
Poco	7	8	6	8	12	9	13	12	9
Abbastanza	26	18	25	25	7	18	22	17	14
Molto	5	17	11	7	14	5	5	11	3
Ns	1	1	0	0	0	9	0	1	7
tot	46	46	46	46	46	46	46	46	46

* Grafico 2.25

	Classica	Equilibrata negli elementi	Elegante	Rassicurante	Poco distinguibile	Adatta all'idea che ho di A	Prosecco di elevata qualità	Moderna	Migliora la mia opinione di A
Per nulla	0	0	1	3	9	2	6	16	11
Poco	6	9	16	12	13	7	11	19	13
Abbastanza	24	28	20	26	12	19	25	10	11
Molto	16	8	9	4	12	8	4	1	1
Ns	0	1	0	1	0	10	0	0	10
Tot	46	46	46	46	46	46	46	46	46

Appendice B: LA TABELLA A DOPPIA ENTRATA

Per comprendere un fenomeno è interessante sapere se lo stato assunto da un soggetto su una variabile può contribuire a determinare il suo stato su un'altra variabile.

La tabella a doppia entrata, o tavola di contingenza è la modalità con cui è rappresentata una distribuzione doppia di frequenza di due variabili x e y, siano esse entrambe qualitative, una quantitativa e una qualitativa oppure entrambe quantitative. Se x e y sono categoriali oppure una qualitativa e una categoriale ordinata avremo una tabella di contingenza, se invece x e y sono cardinali avremo una tabella di correlazione.

La tabella a doppia entrata viene usata per descrivere la distribuzione congiunta di un certo numero di dati in funzione a due variabili, ovvero per indicare che i valori su una variabile sono condizionati dai valori presenti sull'altra.

Tale tabella oltre ad appurare come il campione si distribuisce in funzione delle due variabili è utile per accertare che il numero dei casi che hanno un certo valore su una variabile siano in relazione con i valori ottenuti su un'altra variabile.

LA COSTRUZIONE DELLA STRUTTURA DELLA TABELLA

Per costruire una tabella è necessario che siano verificati alcuni criteri:

- Numero limitato di modalità di ciascuna variabile
- Numero di casi sufficientemente ampio

Una volta appurati i precedenti criteri si può procedere alla costruzione della struttura della tabella.

Generalmente una tabella di contingenza formata da un campione N di osservazioni distribuite in due variabili qualitative avente *r* categorie e l'altra avente *c* categorie è conosciuta come tabella *r* x *c*.

La struttura è formata dalle frequenze osservate nelle *i*-esime righe e nelle *j*-esime colonne, per cui i valori all'interno della tabella sono l'insieme delle frequenze

congiunte *nij* ovvero il numero di unità che presentano contemporaneamente la modalità *i*-esima del primo carattere e *j*-esima del secondo carattere; la frequenza osservata nella *ij*-esima cella è rappresentata con *nij* .

Il totale delle righe è indicato da *ni* ed è rappresentato nell'ultima colonna, mentre il totale delle colonne è indicato da *nj* ed è rappresentato nell'ultima riga; insieme formano i totali marginali, ovvero la somma delle relative frequenze di cella e in termini di frequenze sono il risultato di:

$$n_i = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij} \qquad (3.1.)$$

$$n_j = n_{j1} + n_{j2} + \cdots + n_{jr} = \sum_{i=1}^{r} n_{ij} \qquad (3.2)$$

Conseguentemente possiamo notare che:

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \qquad (3.3)$$

Con N si intende il totale delle osservazioni considerato all'interno dell'analisi.

Una volta costruita la tabella a doppia entrata possiamo cominciare a porci delle domande, solitamente il quesito centrale è "le variabili sono dipendenti/indipendenti?". Per rispondere a questa domanda è necessario chiarire il concetto di indipendenza tra variabili.

"*Due variabili sono indipendenti quando non vi è nessuna consistente o prevedibile associazione tra di loro. In questo caso la distribuzione di frequenza di una variabile non è legata alle categorie della seconda variabile*."[9]

[9] Statistics for the Behavioral Sciences di Frederick J. Gravetter,Larry B. Wallnau pag.619

Per esempio, se consideriamo le variabili sesso e febbre e la proporzione dei maschi e delle femmine che si ammalano è uguale allora significa che la malattia è indipendente dal sesso; se la proporzione differisce allora la variabile febbre risulta essere associata maggiormente ad un sesso piuttosto che all'altro.

Ovviamente ciò (dipendenza come uguaglianza tra le proporzioni) è valido in una tabella 2x2, ovvero se le due variabili hanno due categorie (in questo caso maschio/femmina, con febbre/senza febbre) ma non è valido per una tabella generica rxc, ovvero se le due variabili hanno più di due categorie. È necessario quindi capire cosa implica il concetto d'indipendenza nel caso generico rxc.

Innanzitutto supponiamo che nella popolazione da cui è prelevato il campione la probabilità di un'osservazione appartenente alla categoria i-esima della variabile riga e alla categoria j-esima della variabile colonna sia rappresentata da pij; ne consegue che la frequenza attesa nella cella ij-esima Fij risultante dal campione di N individui è data da:

$$F_{ij} = N\, p_{ij} \qquad (3.4)$$

Pi rappresenta la probabilità di avere un'osservazione appartenente alla i-esima categoria della variabile riga (senza riferimenti alla variabile colonna) mentre pj rappresenta la corrispondente probabilità per la categoria j-esima della variabile colonna.
L'indipendenza tra due variabili all'interno della popolazione implica quindi:

$$p_{ij} = p_i * p_j \qquad (3.5)$$

in termini di frequenze attese l'indipendenza implica:

$$F_{ij} = N\, p_i\, p_j \qquad (3.6)$$

Ma in che modo ciò può essere d'aiuto considerando che l'indipendenza tra due variabili è stata definita solamente in termini di valori di probabilità della popolazione sconosciuta?
La risposta risiede nel concetto che tali probabilità possono essere stimate dalle frequenze osservate ed è facile mostrare come le stime $\hat{p}_i$ e $\hat{p}_j$ delle probabilità pi e pj si basino sui valori osservati dei totali marginali; quindi:

$$\hat{p}_i = \frac{n_i}{N} \qquad e \qquad \hat{p}_j = \frac{n_j}{N} \qquad (3.7)$$

L'uso delle stime pi e pj nell'equazione (2.1.1.7) permette la stima delle frequenze attese nella ij-esima cella della tabella se le due variabili sono indipendenti attraverso la formula:

$$\boldsymbol{E_{ij}} = N\,\hat{p}_i\hat{p}_j = N\left(\frac{n_i}{N} * \frac{n_j}{N}\right) = \frac{n_i * n_j}{N} \qquad (3.8)$$

Le frequenze attese o ***frequenze teoriche*** sono calcolate per ogni cella della tabella di contingenza moltiplicando rispettivamente i totali marginali e dividendoli per N.
Se le due variabili sono dipendenti le frequenze stimate usando la formula (2.1.1.8) e le frequenze osservate dovrebbero differire di un ammontare dovuto unicamente a fattori casuali, se tuttavia le variabili non sono indipendenti la differenza attesa sarà maggiore. Conseguentemente sembra opportuno basare ogni test d'indipendenza tra due variabili formando una tabella di contingenza bidimensionale sull'ampiezza delle differenze le frequenze nij ed Eij.
N.B. Molto importante a questo punto non confondere Eij con Fij.

TEST D'INDIPENDENZA

Una volta chiarito il concetto di indipendenza possiamo passare alla fase successiva, ovvero testare tale indipendenza in modo tale da comprendere la relazione che lega due variabili; per fare ciò è necessario investigare sull'ipotesi:

$$p_{ij} = p_i * p_j \qquad (3.9)$$

In generale definiremo tale ipotesi come ***Ipotesi nulla*** (Ho).

Il test del chi quadro compara il numero di casi che ricadono all'interno di ogni cella con la frequenza teorica (quella che si avrebbe in caso di mancata associazione tra le due variabili). Il test si basa quindi sulla differenza tra le frequenze osservate (nij) e i valori stimati (Eij) usando il valore X2 come mostrato dalla formula:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \qquad (3.10)$$

X^2 dipende quindi dalla differenza tra n_{ij} e E_{ij} , se le due variabili sono indipendenti la differenza sarà minima e di conseguenza X^2 sarà minore nel caso in cui Ho è verificata piuttosto che in quello in cui è rifiutata.

Ciò che è necessario ora è quindi un metodo per decidere quali valori di X^2dovrebbero condurre all'accettazione dell'ipotesi Ho e quali al rifiuto. Tale metodo si basa sulla ***distribuzione del chi quadro*** ovvero una distribuzione teorica formata da un insieme di valori comprendenti tutti i possibili casi in cui Ho è verificata. Tale distribuzione si basa su due fattori:

- Tutti i valori all'interno della distribuzione sono positivi
- I valori attesi del chi quadrato sono bassi quando Ho è verificata

Valori con "bassa" probabilità portano al rifiuto dell'ipotesi gli altri invece portano all'accettazione. Per bassa probabilità si intende generalmente un valore di *0.05* o *0.01* conosciuto come ***valore critico***.

Il test d'ipotesi d'indipendenza può essere effettuato comparando il valore del X^2 con i valori presenti nelle tabelle della distribuzione del chi quadro (appendice).

Se il valore all'interno della distribuzione del chi quadro (trovato dall'incrocio tra i gradi di libertà e i valori delle probabilità 0.05 o 0.01) risulta essere maggiore rispetto al valore di X^2 allora Ho è rifiutata e ciò implica è un'associazione significativa tra le due variabili, nel caso contrario si accetta Ho (indipendenza tra le variabili).

Vi è tuttavia un altro fattore che gioca un ruolo nella distribuzione del chi quadro, ovvero il numero di categorie; più categorie abbiamo più il valore del chi quadro tenderà ad alzarsi. Le categorie vengono considerate nella distribuzione del chi quadro sotto forma di ***gradi di libertà (gl)*** che per la bontà del test sono definiti come:

$$df = (r-1)(c-1) \qquad (3.11)$$

Dove per *c* ed *r* si intende rispettivamente il numero di categorie colonna e categorie riga. I gradi di libertà misurano letteralmente il numero di libere scelte che esistono quando si determina l'ipotesi nulla o le frequenze teoriche.

I gradi di libertà risultano essere il numero di categorie meno uno perché per esempio nel caso di tre categorie si è liberi di scegliere le proporzioni per le prime due ma non per la terza; nello specifico supponiamo di scegliere il 25% per la prima categoria e 50% per la seconda la terza sarà necessariamente il 25% in modo da considerare il 100% della popolazione.

Riassumendo Il test d'indipendenza viene implementato attraverso i seguenti step:

1. Selezionare un campione casuale dalla popolazione e raccogliere i dati in riferimento a due variabili all'interno di una tabella a doppia entrata
2. Determinare l'ipotesi nulla Ho
3. Calcolare le frequenze teoriche (2.1.1.8)
4. Calcolare la differenza tra le frequenze osservate e quelle teoriche per ogni cella della tabella di contingenza
5. Elevare al quadrato le differenze in modo tale da avere tutti valori positivi

6. Dividere la differenza al quadrato per la frequenza teorica (in modo da capire quanto significativa è la discrepanza tra i valori attesi e quelli effettivamente osservati
7. Sommare i valori di ogni cella per ottenere il valore del chi quadro
8. Calcolare i gradi di libertà
9. Localizzare la regione critica incrociando la probabilità α=0,05 e i gradi di libertà all'interno delle tabelle del chi quadro
10. Confrontare il valore trovato al punto precedente con il chi quadro (6). Se il chi quadro è maggiore del valore critico rifiuteremo l'ipotesi Ho quindi avremo dipendenza, altrimenti accetteremo Ho e confermeremo che vi è indipendenza tra le due variabili.

Appendice C:

La seguente tabella espone il procedimento “manuale” effettuato per il calcolo del chi quadro.

Frequenza	Età					
	Gruppo1	Gruppo 2	Gruppo 3	Gruppo 4	Gruppo 5	Totale complessivo
Bottiglie "nuove"						
1	14	21	16	12	10	73
2	12	17	17	8	12	66
3	4	4	7	7	29	51
Totale complessivo	30	42	40	27	51	190

Frequenza teorica	gruppo 1	gruppo 2	gruppo 3	gruppo 4	gruppo 5	tot
1	11,52631579	16,136842	15,368421	10,373684	19,594737	73
2	10,42105263	14,589474	13,894737	9,3789474	17,715789	66
3	8,052631579	11,273684	10,736842	7,2473684	13,689474	51
TOT	30	42	40	27	51	190

DIFFERENZE (fe-ft)	gruppo 1	gruppo 2	gruppo 3	gruppo 4	gruppo 5	tot
1	2,473684	4,86315	0,63157	1,62631	9,59473	0
2	1,5789473	2,41052	3,10526	1,37894	5,71578	0
3	4,0526315	7,27368	3,73684	0,24736	15,3105	0
TOT	0	0	0	8,882E-16	0	

Differenza al quadrato	gruppo 1	gruppo 2	gruppo 3	gruppo 4	gruppo 5	tot
1	6,119113573	23,650305	0,398892	2,644903	92,058975	
2	2,493074792	5,810631	9,642593	1,901495	32,670249	
3	16,42382271	52,906482	13,963989	0,061911	234,41222	
TOT						

Diff. Al quad - F teorica	gruppo 1	gruppo 2	gruppo 3	gruppo 4	gruppo 5	tot
1	0,530882	1,4656092	0,0259553	0,2549627	4,6981481	6,975557346

2	0,23923445	0,398276	0,6939793	0,2027409	1,8441317	3,378362288
3	2,039559684	4,6929186	1,3005676	0,0084432	17,123538	25,16502658
TOT	2,809676133	6,5568038	2,0205022	0,4661468	23,665817	35,51894621

L'associazione seguente: Sesso-Prezzo non è risultata rilevante (Accetto Ho, non c'è dipendenza)

Frequenze	**Prezzo**					
Sesso	**<4**	**4<P<5**	**5<P<6**	**>6**	**ns**	**Totale complessivo**
M	7	19	24	29	2	83
F	8	29	35	29	6	107
Totale complessivo	**15**	**48**	**59**	**58**	**8**	**190**

frequenza teorica	**<4**	**4<P<5**	**5<P<6**	**>6**	**ns**	**Totale complessivo**
M	6,552631	20,968	25,77	25,336	3,49473	83
F	8,447368	27,031	33,22	32,663	4,50526	107
Totale complessivo	15	48	59	58	8	190

diff. tra le freq.	**<4**	**4<P<5**	**5<P<6**	**>6**	**ns**	**Totale complessivo**
M	0,447368	-1,9684	-1,773	3,6631	-1,4947	
F	-0,447368	1,9684	1,773	-3,663	1,4947	
Totale complessivo	0	0		0	0	

diff. tra le freq.	**<4**	**4<P<5**	**5<P<6**	**>6**	**ns**	**Totale complessivo**
M	0,200138	3,8746	3,1459	13,418	2,2342	24,142326
F	0,200138	3,8746	3,1459	13,418	2,2342	24,142326
Totale complessivo	0,400277	7,7493	6,2919	26,837	4,4684	48,284653

diff. tra le freq.	**<4**	**4<P<5**	**5<P<6**	**>6**	**ns**	**Totale complessivo**
M	0,030543	0,1847	0,1220	0,5296	0,63931	2,9583163
F	0,023692	0,1433	0,0946	0,4108	0,49591	2,2947687
Totale complessivo	0,054235	0,3281	0,2167	0,9404	1,13523	5,2530850

La seguente associazione: Sesso-Motivo d'acquisto non risulta rilevante (Accetto Ho, non vi è dipendenza)

Frequenze	**Motivo d'acqisto**				
Sesso	**Consumo abituale**	**Occasioni**	**Regalo**	**ns**	**Totale complessivo**
M	23	49	7	4	83

F	31	61	9	6	107
Totale complessivo	**54**	**110**	**16**	**10**	**190**

freq. Teorica	Consumo abituale	Occasioni	Regalo	ns	Totale complessivo
M	23,5894737	48,0526	6,989474	4,3684	83
F	30,4105263	61,9474	9,010526	5,6316	107
Totale complessivo	54	110	16	10	190

Diff tra le freq.	Consumo abituale	Occasioni	Regalo	ns	Totale complessivo
M	-0,5894737	0,94737	0,010526	-0,3684	0
F	0,58947368	-0,9474	-0,01053	0,3684	0
Totale complessivo	0	0	0	0	0

Diff tra le freq.	Consumo abituale	Occasioni	Regalo	ns	Totale complessivo
M	0,34747922	0,89751	0,000111	0,1357	1,380831025
F	0,34747922	0,89751	0,000111	0,1357	1,380831025
Totale complessivo	0,69495845	1,79501	0,000222	0,2715	2,76166205

Diff tra le freq.	Consumo abituale	Occasioni	Regalo	ns	Totale complessivo
M	0,01473027	0,01868	1,59E-05	0,0311	0,064495355
F	0,01142628	0,01449	1,23E-05	0,0241	0,050029107
Totale complessivo	0,02615655	0,03317	2,81E-05	0,0552	0,114524462

La seguente associazione: Età-Motivo d'acquisto non è statisticamente rilevante (Accetto Ho, non vi è dipendenza)

Frequenza	Motivo d'acquisto				
Età	**Consumo abituale**	**Occasioni**	**Regalo**	**ns**	**Totale complessivo**
Gruppo1	**8**	**20**	**1**	**1**	**30**
Gruppo2	**9**	**24**	**5**	**4**	**42**
Gruppo3	**14**	**20**	**4**	**2**	**40**
Gruppo4	**9**	**16**	**2**		**27**
Gruppo5	**14**	**30**	**4**	**3**	**51**
Totale complessivo	**54**	**110**	**16**	**10**	**190**

frequenza teorica	Consumo abituale	Occasioni	Regalo	ns	Totale complessivo
Gruppo1	8,52632	17,3684	2,526	1,579	30
Gruppo2	11,9368	24,3158	3,537	2,211	42
Gruppo3	11,3684	23,1579	3,368	2,105	40
Gruppo4	7,67368	15,6316	2,274	1,421	27
Gruppo5	14,4947	29,5263	4,295	2,684	51

Totale complessivo	54	110	16	10	190

differenza tra freq.	**Consumo abituale**	**Occasioni**	**Regalo**	**ns**	**Totale complessivo**
Gruppo1	-0,52632	2,63158	-1,53	-0,58	0
Gruppo2	-2,93684	-0,31579	1,463	1,789	0
Gruppo3	2,63158	-3,15789	0,632	-0,11	
Gruppo4	1,32632	0,36842	-0,27	-1,42	0
Gruppo5	-0,49474	0,47368	-0,29	0,316	
Totale complessivo		0		0	0

differenza al quad	**Consumo abituale**	**Occasioni**	**Regalo**	**ns**	**Totale complessivo**
Gruppo1	0,27701	6,92521	2,33	0,335	9,86703601
Gruppo2	8,62504	0,09972	2,141	3,202	14,0678116
Gruppo3	6,92521	9,9723	0,399	0,011	17,3074792
Gruppo4	1,75911	0,13573	0,075	2,019	3,98914127
Gruppo5	0,24476	0,22438	0,087	0,1	0,65573407
Totale complessivo	17,8311	17,3573	5,031	5,668	45,8872022

differenza al quad	**Consumo abituale**	**Occasioni**	**Regalo**	**ns**	**Totale complessivo**
Gruppo1	0,03249	0,39872	0,922	0,212	1,56564254
Gruppo2	0,72256	0,0041	0,605	1,449	2,78057359
Gruppo3	0,60916	0,43062	0,118	0,005	1,16346801
Gruppo4	0,22924	0,00868	0,033	1,421	1,69191919
Gruppo5	0,01689	0,0076	0,02	0,037	0,0818644
Totale complessivo	1,61033	0,84973	1,699	3,124	7,28346773

BIBLIOGRAFIA

Antithrombotic Trialists's Collaboration (2002). "Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients".

Barbaranelli, C. (2003). "Analisi dei dati". Milano: Led.

Baron, R.M., Kenny, D.A. (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations", 51(6), 1173-1182.

Berry, W.D., Feldman, S. (1985). "Multiple Regression in Practice" (Sage University Paper Series on Quantitative Applicatiosns in the Social Science). Newbury Park, CA: Sage.

Bland In, Altman DG (2000). "The Odds Ratio".

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". Journal of Econometrics, 31: 307–327.

Caudek, C., Luccio, R. (2001). "Statistica per psicologi". Bari: Gius. Laterza& Figli Spa.

Daniel McFadden (2007). "Consumer behaviour and the measurement of well-beign"

Davies HTO (1998). "When can odds ratios mislead?"

Deeks JJ (1998). "When can odds ratios mislead?"

Deeks J (1998). "What is an odds ratio?"

Dick R. Wittink (2000). "Building models for marketing decisions".

Durbin, J. e Watson, G. (1950). "Testing for serial correlation in least squares regression". Biometrika, (37): 409–428.

Engle, R. F. (1982). “Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation”. Econometrica, 50(4): 987–1007.

Giulio Palomba. “L’R2 ? No, grazie!” Universit`a Politecnica delle Marche Dipartimento di Scienze Economiche e Sociali (DISES)

Greg Allenby, Sandeep R. Chandukala, Thomas Otter, Jaehwan Kim (2008). “Choice models in marketing”.

Holcomb WL (2001). “Use and misuse of Odds Ratios”.

Yoram Wind. “Marketing research and modelling: progress and prospects”.

Joseph M. Hilbe. “Logistic Regression Models”

Keppel, G., Saufley, W.H., Tokunaga, H. (2001). “Disegno sperimentale e analisi dei dati in psicologia”. Napoli: Edises.

Klaus Zwerina (1997). “Discrete choice experiments in marketing”.

Leonie Burgess, Deborah J. Street (2007). “The construction of optimal stated choice experiment”.

Marisa Giorgetti,Davide Massaro(2007). “Ricerca e percorsi di analisi dei dati con SPSS”.

Menard, S. (2001). “Applied Logistic Regression Analysis” (II Ed.) (Sage University Paper Series on Quantitative Applicatiosns in the Social Science). Thousand Oaks, CA: Sage.

Manera, M. e Galeotti, M. (2005). “Microeconometria”. Carocci.

Miceli, R. (2001). “Percorsi di ricerca e analisi dei dati”. Torino: Bollati Boringhieri editore S.r.l.

Palomba, G. (2010). “Elementi di statistica per l’econometria”. CLUA libri, Ancona. 2a edizione”.

R Development Core Team (2006). R: “A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. (URL http://www.R-project.org).

Sackett DL (1996). “Down with odds ratios!”

Vincenzo Paolo Senese. “Regressione Multipla e Regressione Logistica: concetti introduttivi ed esempi”.

Zhang J, Yu K (1998). “What’s the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes”.