Radboud University Nijmegen

Distiller

Uncovering group differences in progressive profiling

Robin Rutten
Student number: 0712620

Supervisors:

Prof.dr.ir. Th.P. van der Weide
Dr. A. van Rooij
Dion Meijer, GX Software

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in

Computer Science

Thesis number: 668

Nijmegen, May 2013

Abstract

Marketeers profile customers to gain deeper understanding of who their customers are and what are their needs, habits and interests. Progressive profiling is a method of gathering information about a prospect incrementally over time. GX Software's online engagement product BlueConic progressively profiles website visitors on a large scale.

Usually a profile is not considered independently, but as being part of a larger like-minded group or a market segment. The idea is that when we consider two independent segments, there are some characteristics at which these segments differ greatly. Stated differently: there may be some properties that distinguish or typify those segments.

In this thesis we examine how we can automatically find these distinctive features in two independent segments. The progressive nature of the data, which brings a lot of uncertainty, leads us to a statistical approach that considers both the significance, or certainty, of the difference and its effect size. We have built a prototype application which is capable of performing this analysis on actual datasets, and furthermore presents the result to the marketeer in a clear and understandable manner. This application is named *Distiller*, as it extracts the essential properties that distinguish two segments from each other.

Contents

Chapter 1

Introduction

1.1 Context

Customer profiling is the process of gathering data about a company's customers. Companies gather customer data for several purposes. It provides a certain insight in the characteristics and intentions or needs of their customers and can, among others, be used to target particular customers or groups of customers. The underlying idea here is that based on previous interactions, so called *touch point*s of the customer, or interactions of similar customers with the organization, content can be delivered that is most relevant to a particular customer's needs or interests.

For an online business this customer typically is the online visitor, who interacts with the company through its websites. But it may also include other channels, such as the organization's mobile application, social network page or e-mail. Wiedmann et al. [25] refer to **online profiling** as

> the collection of information about Internet surfing behaviour across many different websites for the purpose of formulating a profile of users' habits and interests.

This definition however neglects the fact that online profiling may also include a broad range of non-behavioral information about the visitor. Types of data that online business typically like to gather include

Demographics like age, gender, household income, education level and ethnicity

Spatial or geographic data which includes information regarding country, city, population density or ZIP code

Contextual data like site category, referrer url, search engine referrer keywords, site content or content category

Behavioral data includes information about browsing behavior, purchase intent, past purchases, site actions or read-time.

Search intent data like search keywords or search category

Psychographic data including values, interests, lifestyle or attitude.

The collection of desired and valuable profile information takes resources and time. On its very first visit, the online visitor is a rather anonymous customer, this will change whenever the customer or **visitor profile** is extended with data that has been accumulated over time about that visitor along each channel. In marketing this incremental method of information gathering is characterized as **progressive profiling**. It is the progressive nature of the data makes analyzing and reasoning with the data extremely challenging.

1.2 GX Software

This Master Thesis is written in the context of an internship at GX Software. **GX Software** is a global provider of Web Content Management and Online Marketing software. **BlueConic** is GX's customer-driven online engagement product. According to the product's website [1]

> BlueConic was developed to help companies take advantage of todays new cross-channel marketing opportunities. Layered on top of your existing technology, it fosters ongoing communications through profiling and custom content delivery. For everyone – anonymous visitors, leads and customers – BlueConic facilitates relevant dialogues and experiences across channels and in the moment.

The power behind BlueConic is a big data store which contains profiles consisting of visitors' explicit preferences and implicit behavior. These profiles are continuously updated with relevant information about the online visitors in real-time. The BlueConic users have already created more than 100 million profiles all together since June 2012. This research focuses on this data collection and the insight that can be derived from it.

1.3 Information need

The data collection of progressive profiles provides many interesting challenges to address. For the purposes of this thesis we have sought a research subject that addresses one of these interesting challenges, and is also feasible for a Master Thesis within a six-month period. In this thesis we will examine how we can automatically find distinctive features in two groups of visitor profiles. Note that we consider groups of profiles, instead of focussing on profiles individually.

Suppose that there is a group of visitors that has clicked, and another group of visitors that (has seen the banner but) has not clicked on a banner. The idea is that there is a good chance there are some typical characteristics that makes the first group more likely to click the banner. In this case it is not so much the question what the characteristics of the visitors that click are, but more what characteristics do really distinguish the two groups of visitors. This would give marketeers a valuable insight with possible explanations of the reasons why people click or do not click that banner. On the basis of these insights marketeers may, for example, decide to exclusively present the banner to visitors that share a certain characteristic, or they may decide to change the

[1] http://www.blueconic.com/product.htm

content of the banner. Likewise, marketeers may be interested in the distinctive characteristics between

- visitors that have subscribed to a newsletter and visitors that have not
- male website visitors and female website visitors.
- young visitors and old visitors.

These analyzes all provide insights in the typical interests and needs of those groups, and provide information on how to best target these different groups of visitors. This thesis will focus on the discovery and reporting of the differences between groups of progressive profiles. As we will see in later chapters, the dataset contains all data that we have learned about the website visitors up to a certain moment. Some properties of those visitors are fairly easy to measure, while obtaining other characteristics requires much more time and resources. As a consequence, on some areas we have much data, while on others we have very little data. This introduces a lot of uncertainty when our aim is to reason about this data.

1.4 Research goals

The main goal of this research is to find and report the distinctive differences between two groups of progressive profiles. Groups of progressive profiles are simply referred to as segments. However both terms are used interchangeably. We have decomposed our research goal into four subgoals.

1. Given two segments, find how these two segments differ from each other
2. Given a segment, find how that segment differs from the whole
3. Compose a system for above requirement.
4. Build a prototype.

1.5 Thesis outline

- **Chapter 2** gives a more detailed description of what visitor profiles and segments are. It provides an overview of different types of profile attributes based on their distinctive features. Furthermore it explains how continuous attributes can be transformed into discrete attributes using discretization techniques.
- **Chapter 3** covers the significance aspect of group differences. First it provides a short overview of descriptive statistics, which are used to summarize the data. The remainder of this chapter focuses on inferential statistics. These statistics are used to assess whether the differences between two or more groups can be explained through random chance alone or not. Since the model is based on this theory, the rationale behind these statistics is extensively discussed.

- **Chapter 4** addresses the effect of the group differences. Statistical significance of differences does not provide information about whether a difference is meaningful in practise. In fact, very small differences may be considered significant. Now chapter 4 discusses statistics that characterize magnitude of effect, so called effect sizes.

- **Chapter 5** formulates a model for progressive profiling based on the previously discussed theory. It provides the domain with captures the concepts of visitor profiles and segments. Furthermore the model is related to real world situation, and the key differences are mentioned. Finally the model for the assessment of differences between groups of progressive profiles is provided.

- **Chapter 6** applies the model in the form of a prototype, which operates on actual BlueConic data. The first part of this chapter discusses the technical considerations and implementation of this prototype. The second part addresses the design choices that have been made regarding the presentation of the results to the end user. This chapter concludes with a discussion of the findings of the application.

- **Chapter 7** covers the conclusion and recommendations for further research.

Chapter 2

Visitor profiles, segments and attributes

In the introduction we have discussed data that marketeers typically gather about online visitors. This data is structurally recorded in so called profiles or visitor profiles. In this chapter we shortly explain what we mean by a visitor profile, and a segment of profiles, while a formal definition is provided in chapter 5. As we will see, a profile is made up of a number of attributes. In this chapter we outline several distinctive types of attributes using the theory of scale types. We can distinguish a few classes of attributes that correspond to the properties that apply to the values that are used to represent the attribute. This theory enables us to reason about which operations, techniques or even algorithms are applicable to specific types of attributes. Finally we will consider discretization techniques that allow us to transform continuous attributes to discrete attributes.

2.1 Visitor profiles and segments

A **visitor profile** can be seen as a data object which is described by a number of attributes that capture the characteristics of that visitor. This profile has a unique identifier so it can be linked to an entity in reality, and updated at revisits. In the case of BlueConic a profile is associated with a visitor's web browser, via a HTTP cookie. As this is not really relevant for this research we will not go into technical detail here.

Some examples of attributes that are tracked in practice are gender, visited pages, average time on site and football club preference. These attributes are referred to as **profile properties**. Each attribute is assigned some symbolic values, expressed using primitive data types, that are always a representation of some physical values. This symbolic value is just a way to represent the attribute, and thus may have properties that do not correspond to the properties of the attribute. Categorical attributes, such as ID numbers or gender, lack most of the properties of numbers, but they may me represented as numbers. In the next section we will outline different types of attributes, which accurately reflect the properties of the attribute. A profile is called **progressive** when the profile is extended over time with new data, as new information is learned

about the visitor. Basically a **segment** is a group or set of profiles, which is a subset of all profiles. Marketeers typically divide the homogeneous market into several target segments that they approach differently based on their distinctive needs and interests. In this research each identifiable group of visitor profiles, consisting of one or more profiles, is called a segment.

2.2 Theory of scale types

An attribute can be categorized as either numeric or categorical, or respectively quantitative or qualitative. Stanley Smith Stevens developed the **theory of scale types** [18] in which he claimed that all measurement in science was conducted using four different types of scales: nominal, ordinal (both categorical), interval and ratio (both numeric). Respectively, these scales expose the amount of properties that are applicable to the values of the attribute.

Nominal scale attributes provide the least information. The values that this kind of attribute takes are not naturally ordered. Nominal values can only be used to distinguish one object from another. Examples are gender, favorite football club, ID numbers or zip codes. As the name suggest, **ordinal** attributes provide just enough information to order objects. Examples are dichotomous data (young, old) and non-dichotomous data such as grades and scores when for example measuring opinion (good, better, best). At the level of **interval** scale attributes it becomes meaningful to talk about the *difference* between two values. The main characteristic of interval scales is that the zero point is arbitrary, and the attribute can have negative values. Examples include calendar dates and temperature in Celsius. Time attributes can be at interval scale when measured from a certain epoch, for example Unix Epoch. For illustration, we may ascertain that there is a difference of 31 between -1 and 30 degrees celsius, the mean is temperature is 15, but it makes no sense to conclude that the difference is -30 times as high. When both difference and ratios are meaningful we say these attributes are measured on **ratio** scale. Examples include age, counts, length, calculation times and temperature in Kelvin (0K as zero point). All statistical measures that we will discuss in coming chapters can be used for a variable measured at the ratio level.

2.3 Single-valued or multi-valued attributes

In the previous discussion, it was assumed that each attribute was assigned a single value. However it is very common for an attribute, especially for categorical attributes, to have multiple values or a **sequence of values**. We may have an attribute that consists of the items bought by a certain user, the corresponding amount of these purchases (transaction data), the pages the user has visited or its hobbies. Note that these attributes may be considered to be objects itself, where each possible value is an attribute that is assigned either one for occurrence or zero for non-occurrence. Usually occurrences of values are far more rare than non-occurrences, so usually only occurrences are stored.

2.4 Between and within-attribute differences

Recall that objects are described by a number of heterogeneous attributes that capture the characteristics or properties of that object. With these different scale types, it is interesting to examine how differences between attributes relate to each other. *Numeric attributes* are often captured on different ranges of values. When comparing two objects, there might be much variation in the absolute distance between attribute values. Consider we compare two persons on age and income, in absolute terms their difference in income is usually much higher than their difference in age. Moreover the same absolute distance within an attribute can be perceived differently. Intuitively, most people will agree that the difference between two people with age 2 and 12, or age 8 and 18, is greater than the difference between age 50 and 60. This pattern also holds for many other attributes such as temperature and income. The relation between numeric attributes and categorical attributes usually is much vaguer: how is a difference between nominally scaled male/female related to a difference of 20 in age? In order to avoid that the similarity of two objects is dominated by one or more attributes, **standardization** or **normalization** transformations are usually applied. Furthermore, some attributes may be found to be more influential when determining the difference between two objects. Consider that two objects have the same hair color, shoe size or even the same length of toenail. Does that mean that we consider those objects more equal than objects that do not share this attributes, all other things being equal? A common modification is to assign **weights** to attributes to ensure some attributes will contribute more to the overall similarity than others. In this research we consider all attributes independently, and assume all attributes to be equally important.

2.5 Discretization

We have discussed different types of attributes based on their properties. Another way to distinguish attributes is by considering the number of values an attribute can take. A **continuous attributes** can take any value from a continuous domain (real numbers). If an attribute has a finite set of possible values, it is called a **discrete attribute**. Many machine learning algorithms require attributes to be discrete. In this section we will discuss some algorithms to transform continuous attributes to discrete attributes [12].

2.5.1 Unsupervised discretization

In unsupervised discretization an attribute is divided into a number of intervals without making use of class information. One of the simplest methods to discretize a continuous-valued attribute is by dividing the attribute in a specified number of bins.

Definition 2.1. In **equal-width interval binning** the continuous attribute is divided into k equally sized intervals. Assume that we have an attribute x with a sorted set of m values $\{x_1, \cdots, x_m\}$ and a predefined number of intervals k. The interval width w can be determined using

$$w = \frac{range(x)}{k}$$

where

$$range(x) = max(x) - min(x) = x(m) - x(1)$$

Now the cut points are at $x(1) + w, x(1) + 2w, \cdots, x(1) + (k-1)w$. Although this approach is simple and easy to implement, it is very sensitive to outliers that may skew the range.

Definition 2.2. Equal-frequency discretization is a slightly different technique that assigns each bin the same number of values. If m is the number of values and k is the user-specified number of intervals, then each bin has $\frac{m}{k}$ values. The width of each bin may vary but the number of observations in each bin is constant.

2.5.2 Supervised discretization

Supervised discretization techniques use additional "class information" to determine the intervals. In this situation each instance is assigned a certain class label that is used to determine the optimal splits. Some algorithms are based on the chi-square statistic, which we will discuss in section 3.9. We will discuss the entropy-based supervised discretization algorithm proposed by Fayyad and Irani [6].

Entropy-based discretization

Entropy measure

Shannon (1948) [17] defines the entropy of a random variable X with values $x_1, \cdots, x_k$ as

$$H(X) = -\sum_{i=1}^{k} P(x_i)\ log\ P(x_i)$$

where P is the estimated probability of instance x_i. This measure incorporates the average amount of information per event

$$I(x_i) = -log(P(x_i))$$

Minimum Description Length Principle

Fayyad and Irani (1993) use this entropy measure in their discretization method called the Minimum Description Length (MDL) Principle [6]. The **class entropy** defined is as

$$Ent(S) = -\sum_{i=1}^{k} P(C_i|S) \cdot log_2(P(C_i|S))$$

where there are k classes $C_1, \cdots, C_k$, and $P(C_i, S)$ is the proportion of examples in S that have class C_i. The class entropy measures the amount of information needed to specify the classes in S. A smaller entropy means that the class distribution is less even. In other words: consider that we have two classes C_1 and C_2. We have the lowest entropy (of 0) when set S entirely consists of examples of class C_1. The entropy will be the highest when both classes are equiprobable, that is when 50% of the examples are of class C_1 and 50% are of class C_2. To

evaluate a split point, we take the weighted average of the resulting class entropies. For a set S, with attribute A, if S is partitioned into two intervals S_1 and S_2, using cut value T, the information after partitioning is defined as:

$$E(A,T;S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2)$$

The cut point T that minimizes $E(S,T)$ is consider to be the best cut point. This point has the highest gain, the **information gain** for a split point T is the difference between the class entropy of the entire set S and the weighted average of the two resulting classes after splitting:

$$Gain(A,T;S) = Ent(S) - E(A,T;S)$$

The **minimum description length principle** (MDLP) is used as stopping criterion. LIU et al [12] explain this principle as follows

> MDLP is usually formulated as a problem of finding the cost of communication between a sender and a receiver. It is assumed that the sender has the entire set of instances while the receiver has the class labels of the instances. The sender needs to convey the proper class labeling of the instances to the receiver. It says that the partition induced by a cut-point for a set of instances is accepted if and only if the cost or length of the message required to send before partition is more than the cost or length of the message required to send after partition

The partitioning stops if

$$Gain(A,T;S) \leq \frac{log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}$$

where

$$\Delta(A,T;S) = log_2(3^k - 2) - [k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)]$$

and k_1 and k_2 are the number of class labels in S_1 and S_2. N is the number of instances in S.

The algorithm

In this section the steps of the algorithm are discussed in more detail.

Initialisation The algorithm start with the some initialisation.

- We start with an empty set of cut points, $T_A = \emptyset$.
- We sort the set S with N numeric instances of attribute A, in ascending order.
- Then we extract set D which consists of all distinct values in a set S, $\{d_1, \cdots, d_n\}$. Each distinct value represents a number of instances that belongs to one or more of the k possible classes C, $\{C_1, \cdots C_k\}$. For the calculations we can now build a 2-dimensional matrix M, where each element m_{ij} represents the number of instances that have value d_i and belong to class C_j.

$$M = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,k} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & m_{n,k} \end{pmatrix}$$

Iteration: find the best cut point In every interaction of the algorithm we only need the upper and lower bound of D, b_{low} and b_{up}. In first iteration these bounds are $b_{low} = d_1$ and $b_{up} = d_n$.

For each $d_i \in D$, $d_i \geq b_{low} \ \wedge d_i \leq b_{up}$.

$T = \arg\min_{d_i} E(A, d_i; S)$

Stopping criterion The algorithm stops if $Gain(A, T; S) \leq \frac{log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}$. Otherwise the point is added to the set of cut points T_A.

- $T_A = T_A \cup T$
- Subsequently we want to find the best cut points in the resulting subsets S_1 and S_2 after splitting in point T.

  Where $S_1 \subset S$, with attribute values $\leq T$ and $S - S_1 = S_2$ with attribute values $>T$.
- The lower bound for S_1 is equal to the lower bound of S: b_{low}. The upper bound is the new cut point T: $b_{up} = T$. The lower bound for S_2 is the new cut point: $b_{low} = T$. The upper bound is equal to the upper bound of set S: b_{up}.
- With these bounds we can now find the best cut points for S_1 and S_2. This repeats until the stopping criterion is met for all (sub)sets.

After the iteration process the split point should be sorted and the intervals are creating using these split points.

- Sort T_A in ascending order.
- Finally, create intervals using the cut points T_A, d_1 and d_n.

Chapter 3

Significance of group differences

This chapter provides a brief overview of some key concepts of statistics, which play a crucial part in many data mining algorithms. The discussed concepts are fundamental statistics but essential for the remainder of the thesis. As much as possible we try to consider the subtleties of these statistics to ensure we will apply them in the right way and to be aware of the limitations of our approach.

When the aim of study is to only describe the characteristics of the data that has been collected, this field of statistics is called **descriptive statistics**. In contrast, this chapter will mainly focus on **inferential statistics**. In inferential statistics samples are used to make generalizations about the populations from which the samples were drawn. These statistics can be used to consider groups of subjects and to draw conclusions about general differences between these groups. We must keep in mind that these statistics are all defined in a period were researchers had to do with little data, with only a sample of the much larger actual population, and needed tools to draw reliable conclusions.

3.1 Population and samples

The **population** consists of *all* subjects of interest. In actual situations, both in time and resources it is often not possible to gather data about the entire population. Usually only a subset of the population is considered, this subset is called a **sample**. In **inferential statistics** samples are used to make generalizations about the populations from which the samples were drawn. A sample therefore should be *representative*, that is when it accurately reflects the subjects of the entire population. Many statistics therefore require a **random sample**, this means that each subject of the population has an equal chance of being part of that sample.

Example 3.1.1. Imagine a researcher is interested in which hair colors are most common in the Netherlands. The researcher has a large Dutch family, so he has made a list with the original hair color of 80% of its family. He sees that 50% of them is red-haired, 40% brown-haired and 10% has black hair. Using this data the researcher concludes there are far more red-haired people in the

Netherlands than black haired people.

There are a few flaws in this research. First, the researcher makes a generalisation about the people in the Netherlands, using a sample of its family members. This type of sample is called a **convenience sample**. It is true that the researcher has a Dutch family, but his family members might not be representative for all people in the Netherlands. As we know each family has its own different background. It may very well be that there are accidentally many red-haired people in his family. Using this sample the researcher may only conclude about the population from which this sample was drawn, that is its own family. Now let us assume that the user takes his own family as the population he want to make statements about. The second thing is that the sample is not randomly taken from the population. It may be the case that the researcher studied a certain side of the family with many red-haired people, and the 20% that he did not study are all black-haired. However, when a sample is sufficiently large, that chance is high that it accurately represents the population, although it is not a random sample.

3.1.1 Paired versus independent samples

Usually we work with more than one sample of observations. We can distinguish samples on whether or not their observations are *independent*. In this thesis we always we will reason from **independent samples**.

Independent samples Suppose that we have two samples with values, or scores, for a certain dependent variable. The samples are said to be independent if the probability of a specific value occurring in one sample is not influenced by the values that occur in the other sample.

Paired samples In paired samples it is possible to uniquely match or pair the observations in the first sample with an observation of the second sample. This will typically occur in pre-test/post-test studies where a variable is measured before and after an intervention (repeated-measures design). Another option is to create samples were each subject in a sample is purposefully paired, based on some variable of interest, to a subject in the other sample (matched-subjects design). An example is to match subjects on age, so that each subject in the first sample is matched with a subject in the second sample having the same age.

3.2 Descriptive statistics

Descriptive statistics are used to quantitatively summarize and describe the data and characteristics of a population or sample of subjects. A measure of the population is called a **parameter**. A descriptive measure associated with a sample is called a **statistic**. By convention, Greek symbols are used for population parameters and the Roman letters for sample statistics.

3.2.1 Measures of Central Tendency

A **measure of central tendency** is a single value that describes a set of data by identifying the central position within the data. The most commonly used

measure of central tendency is the mean, the average value. The sample **mean**, or arithmetic mean, is defined as

$$\text{mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

where n is the number of observations in the sample and x_i is the ith observation. The disadvantage of the mean is that is is susceptible to the influence of outliers. In some situations the median will provide a better description of the data. The **median** is the middle observation when data have been arranged in order from the lowest to the highest value. For a sorted set of observations $\{x_1, x_2, \cdots, x_n\}$

$$\text{median} = \begin{cases} \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ is even} \\ x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \end{cases} \tag{3.2}$$

The least used measure of central tendency is the mode, it is used when one is interesting in the most commonly observed value in the data. The **mode** indicates the most frequent value within the data.

3.2.2 Measures of Variability

Measures of central tendency are mainly useful when summarizing data. However these measures are limited in the information they provide, because they do not provide information about the variation within the data. A **measure of variability** describes the spread or dispersion of a set of data. These measures indicate if values are widely spread out or relatively concentrated around a single point such as the mean.

The total (or maximum) spread or dispersion in a distribution is often expressed in terms of the range. The **range** is the difference between the maximum value and the minimum value of a distribution. For a sorted set of observations $\{x_1, x_2, \cdots, x_n\}$, where $x_1 = x_{min}$ and $x_n = x_{max}$:

$$\text{range} = x_{max} - x_{min} \tag{3.3}$$

Because this range measure is based on the two most extreme values in the data, it is very sensitive to outliers, more robust measures are often used such as the **interquartile range**.

The average amount of spread within the distribution is often expressed with the variance and standard deviation. The difference or distance between a single point and the population or sample mean is called its **deviation**. The **variance** computes the average squared deviation from the mean. The population variance, denoted by σ^2, is given by the formula

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{3.4}$$

where N is the size of the population and μ is the population mean defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

A slightly different formula is used for the sample **variance** s^2

$$\text{variance} = s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1} \tag{3.5}$$

where n is the size of the sample and $\bar{x}$ the sample mean as defined in equation (3.1). The sample size, the number of observations, is corrected by one: $n-1$. This is because the true variance is underestimated by using the sample mean instead of the population mean.

Note that to make the deviation scores positive, the deviation between each observation and the sample mean is squared, also known as the **squared deviation**: $(x_i - \bar{x})^2$. The sum of the squared deviations for all observations is known as the **sum of squared deviations** or **sum of squares (SS)**.

$$\text{sums of squares} = SS = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.6}$$

The sum of squares is divided by the number of observations to get the average squared deviation. By squaring the deviation scores, we have changed the original scale of measurement. There is another measure that compensates for this by simply taking the square root of the variance: the standard deviation. The **standard deviation** is defined as the average deviation from the mean.

$$\text{standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.7}$$

3.3 Hypothesis testing

Whenever we take a random sample of the population, we will have to a greater or lesser extent different observations. When we see a particular variation between a specific sample and the population, or two sample distributions, we want to test if this can be explained through random chance alone or not. In other words, we want to decide whether the result of a statistic is **statistically significant**.

The usual way to handle this is to state two mutual exclusive and exhaustive hypotheses, of which either will be true. Unless we find enough evidence, the null hypothesis, that suggests an absence of effect *in the population*, is assumed to be true.

The null hypothesis (H_0) is the assumption that the difference among observations are due simply to *random sampling error* or *chance*.

The alternative hypothesis (H_A or H_1) states that the difference can not be explained by random chance alone.

If the null hypothesis is true, the **p-value** is the probability to get a difference this large or larger. In other words it is the probability that this evidence would arise given the null hypothesis is true [2] [5]

$$P(E|H_0)$$

In a two sample t-test (section 3.7.2) we ask the question: what is the chance to get this mean difference (E), given that the two samples were obtained from the same population (H_0)? Note that we *do not* consider the probability that the obtained groups were sampled from same population, which would be $P(H_0|E)$. A p-value of 0.05 means that when we randomly draw two samples of this size from the same population we expect to see a mean difference this large or larger 1 out of 20 times. Consequently, a p-value near 1 means that we always expect to see a difference this large or larger. The **significance level** α is the threshold at which we decide to reject the null-hypothesis, which is usually set at $\alpha = 0.05$. Even with a strong significance level there is still a chance that our rejection of the null hypothesis is wrong, we refer to this as the **type I error**.

Note that this hypothesis does not conclude anything about whether the observed difference is small or large. For example, with a sample of a million people, both a difference of 10 and a difference of 1000 might be enough to result in very small p-values, and thus rejection of the null hypothesis. With a very small sample however, even a difference that we would consider as very large, can be considered as present simply due to chance. Because of the design of the statistics, we can be very confident in small differences, and very unconfident about large difference, just due the size of our sample. Some researchers suggest that the significance testing is too heavily influenced by the sample size [11]. Moreover, in data mining, we often can work with sufficiently large amounts of data. In these situations there is no need to verify the probability of a difference occurring due to chance. Differences between very large samples will be considered statistically significant anyhow because of the design of the inferential statistics.

The bottom line is that the statistical significance is not informative about the **substantive significance** or practical importance of the observations. The substantive significance is concerned with the meaning of the observation: is the difference large enough to be meaningful in practice? Therefore we need some other statistic such as for example the **effect size**. However this does not mean that statistical significance is unimportant, it is wise to consider both. We will discuss the concept of practical or substantive significance in chapter 4. This chapter continues with the discussion of inferential statistics.

3.4 Independent and dependent variables

Recall from chapter 2 that an **attribute** captures a characteristic of an object. The variable is the operationalized way in which the attribute is represented. In an experimental design variables can be divided into in two major types, independent and dependent variables. An **independent variable** is the variable that is presumed to have an effect on the **dependent variable**. The independent variables usually are considered predictor variables because they predict the dependent variables. The independent variable may be

- a different condition to which a subject is exposed. In this case the researcher has control over the variable. A classical example is in medical studies, where one group is given a certain (real) treatment, while the control group is given a placebo treatment. In advertising for example, one could randomly expose subjects to different advertisements, to determine the effect on the click rate.

- characteristics that the subject brings in a research situation. A researcher could for example be interested in the differences between men and women, subjects with different income levels or subjects within certain age ranges.

It depends on the research context which of the variables are the independent and which are the dependent variables. In some situations the variable might be considered as an independent variable, while in others that same variable is considered to be the dependent variable.

3.5 Design of inferential statistics

The statistics that are used in inferential statistics to determine statistical significance, which will be discussed in the coming sections, share a common design. In inferential statistics we are concerned with the question whether a phenomenon we see in the sample represents an actual similar phenomenon in the larger population. A sample is not expected to perfectly represent the population: we always expect to see a deviation between the sample and the population. We refer to this variance, that we reasonably would expect when we randomly select a sample of the population, as the standard error (section 3.6). In inferential statistics we usually examine whether the sample statistic is large of small compared to the expected variance

$$\frac{\text{sample statistic}}{\text{standard error}}$$

Using an appropriate distribution we finally can determine what is the chance to get a ratio as extreme as this.

3.6 Standard error

A sample is supposed to represent the larger population. However, it is inherent in sampling that we make errors. There are generally two causes for differences between the sample and the population

random sampling error This is the difference between the sample and the population caused by chance. Each time we take a sample from the population, our observations deviate to a greater or lesser extent on what we see in the population.

sampling bias These are errors that are systematic due to inadequate design of the sampling process. The sample is collected in such a way that some members of the intended population are less (or more) likely to be included than others.

The inferential statistics do not correct for sampling bias. The assumption is that the sample is taken randomly from the population, so there is no sampling bias. When our sample is sufficiently large, there is a high probability that even a biased sample approximates the population. Therefore, all else being equal, statements based on larger samples may be considered as more reliable than those on smaller sample sizes.

The statistics do correct for **random sampling error**. This is why we talk about the *standard error*. The standard error is the standard deviation of the sampling distribution of a statistic. The **sampling distribution** can be seen as the distribution of the statistic for all possible samples of a given size from the population. In other words, when we repeat our sampling procedure many times, the average deviation we find from our statistic will be the standard error. However in most practical cases we do not have information about the whole population, and thus the true value of the standard error. In these situations we make an estimation of the standard error. The **sample standard error** is obtained by dividing the sample standard deviation (equation 3.7) by the square root of the number of observations n in the sample. We will now focus on the standard error of the mean. The basis formula for the standard error of the mean (SEM) is

$$\text{standard error of the mean} = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Note that the sample standard deviation is our best guess of the population standard deviation. When in the population the values are highly concentrated around the mean, the sample means of different samples drawn from that population will not differ much. Vice-versa, when the values within the sample are far apart, it is likely that the values within the population are far apart. When there is much average deviation of the mean, this is when there is much variation between the values within the sample, we expect a larger standard error. Oppositely, when in our sample all observations are very centered around the mean, we would expect to see this pattern when we select another sample of the same size. Furthermore, the larger the sample, the greater the likelihood that our sample accurately represents the population. Therefore larger sample sizes produce smaller standard errors. As a consequence, all else being equal, larger sample sizes produce higher statistics, and are more likely to be judged statistically significant. In section 3.3 we already highlighted that significance testing is heavily influenced by the size of the sample.

Finally, we have now reached the point that we can discuss inferential statistics for statistical significance tests of sample differences. In the following sections we will discuss three different statistics.

Student's T-test The t-test addresses whether the means of two groups are statistically different from each other.

Analysis of variance (ANOVA) ANOVA tests significance of differences between the means of two or more groups.

Chi-squared test The chi-squared test is a statistical hypothesis test to examine the difference between two or more distributions.

3.7 Student's T-test

A **t-test** is a statistical test to determine whether there is a significant difference between the means of two groups. Technically, a t-test is any statistical test that uses the Student's t-distribution. The **t-distribution** is a family of probability distributions that is used when the sample size is small ($n < 120$) and the

population standard deviation is unknown. Otherwise the normal distribution is appropriate. Dependent on the size of the sample specific t-distributions are available. We can distinguish three types of t-tests, based on the number and type of samples

- One-sample t-test
- Independent two-sample t-test
- Paired two-sample t-test

We have explained the difference between independent and dependent samples in section 3.1.1. The paired samples t-test is not discussed because it is not relevant for this study.

3.7.1 One-sample t-test

In a one-sample t-test the sample mean is compared to a known population mean or a meaningful fixed value. The *t-statistic* is observed by dividing the difference between the population and the sample means (mean difference) by the sample standard error

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where μ is the population mean, and $\bar{x}$ the sample mean (3.1).

3.7.2 Independent samples t-test

The **independent samples t-test** is used to compare the means of two unrelated, non-overlapping samples on a given variable. In this test we have a single categorical independent variable with exactly two levels or categories (e.g. male/female). The aim of the t-test is to find if the means on a quantitative dependent variable differ significantly between the two levels of the independent variable. We want to know if the difference of the means of the two samples is large compared to the standard error. So that it is likely that the difference reflects a true population difference, and is not caused by random chance alone. The equation that provides the t-value is

$$t = \frac{\text{observed difference between sample means}}{\text{standard error of the difference between the means}}$$

The numerator is the difference of the two means: $\bar{x_1} - \bar{x_2}$. The denominator, the standard error of the difference between the means, depends on the two samples and their characteristics. Now we can use the appropriate t-distribution to find out what is the probability to get a sample statistic as extreme as the calculated t-statistic.

3.8 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) provides a statistical test to compare the means of *two or more* groups. It generalizes the t-test to more than two groups. Although there are several versions of ANOVA, only **one-way ANOVA** is discussed in this section.

3.8.1 One-way ANOVA (*F*-test)

One-way analysis of variance is a statistical technique to compare the means of *two or more groups* for a quantitative dependent variable to see whether there are statistically significant differences among them. The term *one-way*, or one-factor, indicates that there is a single categorical independent variable (also called the treatment), with two or more of levels. If each subject is only exposed to one treatment, this is called **between-subjects** one-way ANOVA. In this case the subjects in each condition group are mutually exclusive. We talk about **within-subjects** ANOVA when each subject is exposed to several levels of treatment. In case the independent variable has only two levels, that is when we compare two means, we will come to the same conclusions whether we use an independent samples t-test or one-way ANOVA. For independent variables of two levels: $F = t^2$.

The *null hypothesis* is that the population means are all equal

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

where k is the number of conditions. The alternative hypothesis states that at least one mean is different

$$H_A : \exists i, j : \mu_i \neq \mu_j$$

We know that the mean of each sample is the best guess we can get of the population mean. Each subject within a sample has a certain deviation, error, from this sample mean. For each sample we could also determine an average squared deviation from the mean, which we called the *variance* (SS/df). In one-way ANOVA this concept of variance is used, the major difference is that we consider multiple samples together instead of individual samples. When we have multiple samples, there is a difference or variance within all samples. Thereby, also between the different samples means there is a certain deviation. One-way ANOVA is based on the idea that when comparing samples the variance can be divided into two components

Between-groups variance This is the average variance between the means of the groups

Within-groups variance This is the average variance within the groups, the variance around the group mean

ANOVA basically determines the ratio between the average amount of variation between *each* of the samples and average amount of variation within *each* of the samples, which is expressed as:

$$F = \frac{\text{Between-groups variance}}{\text{Within-groups variance}}$$

This is operationalized by dividing the mean square between groups by the mean square within groups:

$$F = \frac{\text{mean square between}}{\text{mean square error (mean square within)}} = \frac{MS_{between}}{MS_{error}}$$

The general formula for the mean squares MS is, just like the sample variance, $\frac{SS}{df}$. We will now discuss the two mean squares in detail.

Mean square between groups

The **mean square between groups** is the average amount of variation between the groups. The interesting part is the **sum of squares between groups**. We have seen the notion of sum of squares when discussing the sample variance. This time we consider the deviation of the mean of each sample with the grand mean. The **grand mean** $\bar{X}$ is the mean of all samples combined. Let X be the collection of samples, and let X_i be the ith sample of K samples. Let n_i be the number of subjects in the ith sample, so $X_i = \{x_1, \cdots, x_{n_i}\}$, and N be the total number of subjects. X_{ij} is the jth subject of the ith sample.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n_i} X_{ij}$$

The sum of squares between groups is given by

$$\text{sum of squares between groups} = SS_{between} = \sum_{i=1}^{K} n_i (\bar{X}_i - \bar{X})^2 \tag{3.8}$$

where $\bar{X}_i$ is the mean of the ith sample. This difference between the sample mean and the grand mean applies for each element in the sample. Each sample has a number of subjects and can be of different size, we take this into account by n_i, which is the size of each sample. We finally divide this sum of squares by the degrees of freedoms to get a kind of an average

$$MS_{between} = \frac{SS_{between}}{df} = \frac{SS_{between}}{K - 1} \tag{3.9}$$

Mean square within groups

The **mean square within groups** (MS_{within}) or the **mean square error** (MS_{error}) is the average amount of variation within each of the samples. For each individual sample i we can compute the sum squared deviations by

$$SS_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Here $\bar{X}_i$ is the mean of the ith sample. Note that by dividing SS_i by df_i we would just get the variance as in equation 3.5. In the **sum of squares within** (SS_{within}) we combine the deviations all of the K samples in a single estimate.

$$SS_{within} = \sum_{i=1}^{K} SS_i = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \tag{3.10}$$

$$df_{within} = \sum_{i=1}^{K} df_i = \sum_{i=1}^{K} (n_i - 1) = N - K$$

$$MS_{error} = MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{SS_{within}}{N - K}$$

F statistic

Note that the difference between an individual subject and the grand mean, is equal to the sum of the difference between the subject and its sample mean *and* the difference between the sample mean and the grand mean.

$$(X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X}) = X_{ij} - \bar{X}$$

We can determine the sum of squared deviations for all subjects with the following formula, that we call SS_{total}

$$SS_{total} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

The interesting thing in ANOVA is that this squared difference between each subject, of all samples together, with the grand mean is equal to the sum of two things:

- For all samples, the squared difference between each subject and the group mean
- For all samples, squared difference between the group mean and the grand mean times the size of the sample.

These two parts are exactly what we get by SS_{within} and $SS_{between}$, our two SS components.

$$SS_{total} = SS_{within} + SS_{between}$$

We saw that the mean squares are the estimation of the average of these sum of squared deviations (estimates of variance). The **F-ratio statistic** is the ratio of these two estimates

$$F = \frac{MS_{between}}{MS_{within}}$$

What this statistic basically tries to answer is whether variability that we see between the different samples, is large or small compared to the variability that we see within the samples. Using this statistic and the family of F-distributions we can determine the probability of finding this ratio, and test the *null hypothesis* whether the groups means are the same. The statistic tends to be larger if the alternative hypothesis is true, than if null hypothesis is true. If the null hypothesis is true, if the population mean of the samples is roughly the same, we expect the $MS_{between}$ to be small.

3.9 Chi-square

Previously we discussed statistics to determine statistical significance for numeric variables. This section will focus on discrete or categorical variables.

3.9.1 Contigency tables

A **contingency table** is a type of table (in matrix format) used to display and analyze the relationship between two or more *categorical* variables. The cells of the table display the frequency distribution of variables, which can be either frequency counts or relative frequencies. It is most used for analyzing two variables presented in a 2-dimensional contingency table. Consider for example this 2 x 3 contingency table, which divides a group of people by gender and whether those are right-handed, left-handed or ambidextrous. Based on this

	Gender		
Handedness	Male	Female	Total by Handedness
Right-handed	50	60	110
Left-handed	15	10	25
Ambidextrous	10	5	15
Total by Gender	75	75	150

Table 3.1: 2 x 3 contingency table

table we can easily compute some probabilities. We can see for example that we have equal chance of seeing males or females.

$$P(Gender = Male) = P(Gender = Female) = \frac{75}{150} = 0.5$$

Ten percent of this group is male and left-handed.

$$P(Gender = Male \cap Handedness = Left - handed) = \frac{15}{150} = 0.1$$

Given that someone is female, the chance that she is right-handed is 80 percent.

$$P(Handedness = Right - handed \mid Gender = Female) = \frac{60}{75} = 0.8$$

In this group 10% of the people is ambidextrous.

$$P(Handedness = Ambidextrous) = \frac{15}{150} = 0.1$$

In the table we can see that the proportion of right-handed female is greater than the proportion of right-handed male. It would be interested to see whether this difference is significant or might occurred due to chance. We can use the chi-squared test to see whether there is a dependency between Gender and Handedness.

3.9.2 Pearson's chi-squared test

Technically, a chi-squared test is any statistical test that uses the chi-squared distribution. Usually, when referred to the chi-squared test, the Pearson's chi-squared test is meant. The **Pearson's chi-squared (χ^2) test** is a statistical

hypothesis test to examine the difference between two or more distributions. The general statistic is defined as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Here O and E are respectively the observed and expected, theoretical, frequency for each cell in the contingency table. Pearson's chi-square test has two types

- Goodness of fit test
- Test of independence

3.9.3 Goodness of fit test

The **goodness of fit** test is used to check whether an observed distribution matches an expected or theoretical distribution. In other words: it tests how *well* the model *fits* the actual data. It is also called **one-way chi-square** because the data is classified using one variable. The *null hypothesis* states that there is no significant difference between the expected and observed frequencies. Expected frequencies can be determined using prior knowledge, or one can assume that each category has an equal frequency.

Example 3.9.1. In our contingency table example, we had a total sample size of 150 people. We might expect that when we draw a random sample the number of men and women are equal in frequency. So we would expect 75 men and 75 women. Our observed frequencies were indeed 75 men and 75 women. But what if that would be 85 and 65?

$$\chi^2 = \frac{(85-75)^2}{75} + \frac{(65-75)^2}{75} = 2.67$$

The probability of observing a value as extreme as this (with $df = 1$) is

$$p = 0.1$$

With an α level of 0.05 we conclude that there is no significant difference between the expected and observed frequencies: the data fits the model.

3.9.4 Test of independence

This test examines whether two nominal variables are independent of each other, it is also called the **two-way chi-square** test. When we take two nominal variables A and B, the null and alternative hypothesis are

H_0**:** Variable A and variable B are independent.

H_A**:** Variable A and variable B are not independent.

The degrees of freedom for this test is equal to

$$df = (r-1)(c-1)$$

where r is the number of levels for variable A and c the number of levels for variable B, which represents the rows and columns in the contingency table.

The expected frequency for each cell in the contingency table can be computed using

$$E_{rc} = \frac{n_r \cdot n_c}{n}$$

For each cell we take multiplication of the total of its row and the total of its column, divided by the total sample size. Then the chi-squared statistic can be computing using the formula

$$\chi^2 = \sum \frac{(O_{rc} - E_{rc})^2}{E_{rc}} \tag{3.11}$$

Using the chi-square distribution for the degrees of freedom df, we can determine the probability of observing a sample statistic as extreme as the test statistic.

$$P(\chi^2 > CV)$$

Example 3.9.2. We take our example in section 3.9.1. The two variables in this example are *gender* and *handedness*. In the given sample, we see that the probability of being right-handed is slightly larger when the gender is female.

$$P(Handedness = Right - handed | Gender = Male) = \frac{50}{75} = \frac{2}{3}$$

$$P(Handedness = Right - handed | Gender = Female) = \frac{60}{75} = 0.8$$

Now we want to conclude whether there is a dependency between these two variables, *gender* and *handedness*. We can use the independence test for this. First we compute the expected probabilities for each combination.

$$E_{male,right-handed} = 75 \cdot \frac{110}{150} = 55$$

$$E_{male,left-handed} = 75 \cdot \frac{25}{150} = 12.5$$

$$E_{male,ambidextrous} = 75 \cdot \frac{15}{150} = 7.5$$

$$E_{female,right-handed} = 75 \cdot \frac{110}{150} = 55$$

$$E_{female,left-handed} = 75 \cdot \frac{25}{150} = 12.5$$

$$E_{female,ambidextrous} = 75 \cdot \frac{15}{150} = 7.5$$

Note that the sum of all expected values is equal to the sample size. As in this example we have an equal amount of males and females (both 75), when there is no dependency we expect that the number of right-handed people in the sample is divided evenly between them. We now investigate whether the differences between the observed values and the previously computed expected values, are sufficiently large to reject our null hypothesis that there is no dependency between gender and handedness.

H_0**:** *gender* and *handedness* are independent.

H_A**:** *gender* and *handedness* are not independent.

$$\chi^2 = \frac{(50-55)^2}{55} + \frac{(15-12.5)^2}{12.5} + \frac{(10-7.5)^2}{7.5}$$
$$+\frac{(60-55)^2}{55} + \frac{(10-12.5)^2}{12.5} + \frac{(5-7.5)^2}{7.5} = 3.576$$

We have two degrees of freedom: $df = (2-1)(3-1) = 2$, so

$$\chi^2 = 3.576, df = 2$$

The p-value, the probability of observing a chi-square value as extreme as this is

$$p = 0.17$$

With a α level of 0.05 we stick to the null hypothesis, and conclude that there is no dependency between the two variables.

Chapter 4

Effect of group differences

In the previous chapter we discussed statistics that can be used to determine statistical significance related to group differences. Statistical significance is concerned with whether a research result is due to chance or sampling error. We saw that statistically significant differences can be found with very small differences, if the sample size is large enough. Thereby, the presented inferential statistics provide no information about whether a difference is meaningful in practise [23]. Sawyer and Peter [16] state that

> marketing researchers should become more aware of the limited value of classical statistical significance tests

Furthermore

> empirical results should be described and analyzed such that the size and substantive significance of obtained effects are emphasized and not merely the p-values associated with the resulting test statistics.

Moreover, when there is sufficient data available, there is no need to reason about sampling error and chance. Differences between very large samples will be consider statistically significance anyhow because of the design of the inferential statistics. As Berry and Linoff [1] state

> One difference between data miners and statisticians is that data miners are often working with sufficiently large amounts of data that make it unnecessary to worry about the mechanics of calculating the probability of something being due to chance.

There certainly is a need for more than null hypothesis significance testing (NHST) alone. **Practical** or **subjective significance** is concerned with whether the result is useful in the real world [9]. In this chapter we will discuss statistics that characterize magnitude of effect, so called **effect sizes**. These statistics are in the class of descriptive statistics. Unstandardized measures refer to raw, absolute differences in the dependent variable, such as the mean difference

$$\bar{x}_1 - \bar{x}_2$$

There are different types of effect sizes suited for different research situations. Two categories of effect size measures we will discuss are

Standardized effect sizes Standardized measures express the difference in standardized units of difference.

Variance-accounted-for statistics. These statistics reflect the amount of "explained" variance within an experiment that is attributable to an independent variable.

The measures provide quantitative outcomes that enable to assess the magnitude of differences on a scale tied to the real world. However only the user can put the results into context; the measures assist the user to put it into real world context. Note that "practical significance" is a subjective concept: an effect size of 0.30 might interpreted as small for some, while it is large for others. Usually there are some general guidelines in literature what can be considered as a small or large effect.

To summarize, when we consider the effect size as an indication for the *magnitude* of the effect, we can consider statistical significance as an indication for the *certainty* about the existence of an effect.

4.1 Effect size of numeric variables

Standardized effect sizes of numeric variables are usually expressed in **standard-deviation units**. The effect size for the mean difference between two populations is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

4.1.1 Cohen's d

Cohen's d [4] is a measure of effect size that can be used with t-tests. It is defined as the difference between the group means, divided by the standard deviation. The idea is that the standard deviation of either group could be used when the variances of the two groups are homogeneous.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$$

The interpretation of this measure is that the difference of the mean values is d standard deviations. Whether the value of d should be considered as large or small in practise is dependent on the application. In general, $d < 0.20$ is interpreted as a trivial effect size, $d \geq 0.20$ to < 0.50 is a small effect size, $0.5 \leq d \leq 0.8$ is a medium or moderate effect size and $d > 0.80$ is a large effect size. Usually the **pooled standard deviation** is used for two independent samples with unequal variances

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

The formula of the resulting effect size statistic is

$$\frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

We refer to this formula when we talk about Cohen's d effect size.

4.1.2 Eta-squared

Eta-squared (η^2) is a measure of effect size for use in ANOVA. This is used when we compare multiple groups on a numeric variable. It is a variance-accounted-for statistic. Eta-squared reflects the percentage of variability in the dependent variable that can be explained by the independent variables in the sample data. In the context of one-way ANOVA the formula is

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

Another commonly used modification of this statistic is **partial eta-squared** (η_p^2).

$$\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{error}}$$

However for one-way ANOVA, which we have discussed in this thesis, eta-squared and partial eta-squared are equal. Recall that we have seen $SS_{between}$ and SS_{total} in section 3.8. We have also seen that

$$SS_{total} = SS_{within} + SS_{between}$$

where $SS_{within} = SS_{error}$. So $SS_{total} = SS_{between} + SS_{error}$. Because we do not use one-way ANOVA and Eta-squared in the current version of our model, we do not provide an effect size interpretation for this statistic.

4.2 Effect size of categorical variables

4.2.1 Cramér's V

Cramér's V is a measure of association for two nominal categorical variables. The formula is defined as

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

where
χ^2 is the chi square formula (equation 3.11),
n is the total number of cases,
k is smallest number of categories of the two variables, that is the smallest number of the total number of rows or columns in the contingency table.

If one of the categorical variables is dichotomous, Cramér's is equal to the phi statistic

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Cramér's V ranges from zero to one. The closer V is to 0, the smaller the association between the two categorical variables.

$$0 \leq V \leq 1$$

For the interpretation of Cramér's V we can consider Cohen's effect size interpretation rules-of-thumb. Using several sources [3] [14] we have derived a five-level scale Cramér's V interpretation that can be found in table 4.1. Here $df = 1$ means that the smallest dimension is 2.

Interpretation	$df = 1$	$df = 2$	$df = 3$
Very strong	≥ 0.7	≥ 0.49	≥ 0.42
Strong	< 0.7	< 0.49	< 0.42
Moderate	< 0.5	< 0.35	< 0.29
Weak	< 0.3	< 0.21	< 0.17
Negligible	< 0.1	< 0.07	< 0.06

Table 4.1: Interpretation table for the Cramér's V statistic

Chapter 5

A model of progressive profiling

In this chapter we define a formal model that enables us to reason about the research domain and to provide a solution for the research goals. We start by providing the domain which captures the concepts of visitor profiles and segments. Subsequently we will discuss how this model relates to the real world situation that we want to reason about. Finally we outline our approach for the assessment of differences between groups of progressive profiles. This approach is strongly based on the theory that is discussed in the previous chapters. This discussion includes how the statistical results should be presented to marketeers, which usually do not have a scientific background and lack deep statistical knowledge. Our aim is to translate the statistics into graphical or textual reports that are highly understandable and actionable for marketeers.

5.1 The domain

We have a collection U of **visitor profiles**. The collection A of profile attributes or **profile properties** is the set of all possible profile properties. We have a collection V of values that consists of the values that are assigned to a profile property.

Definition 5.1. A single **visitor profile** $u \in U$ is a partial mapping of profile properties to values, so

$$U \subseteq A \mapsto V^+$$

Example 5.1.1. Consider a profile $u \in U$, profile properties $a_1, a_2, a_3, a_4, a_5 \in A$ and values $v_1, v_2, v_3, v_4, v_5, v_6 \in V$

$$u = \{a_1 : [v_1, v_2], a_2 : [v_3], a_3 : [v_4], a_4 : [v_5], a_5 : [v_6]\}$$

This may be the abstract representation of the concrete profile

$$\begin{aligned} u = \{&\text{hobbies} : [\text{tennis}, \text{football}], \\ &\text{age} : [33], \\ &\text{subscription} : [1], \\ &\text{gender} : [\text{male}], \\ &\text{lastvisit} : [1359365228355]\} \end{aligned}$$

Definition 5.2. A **segment** S is the subset of profiles that satisfy a certain condition on its attributes. We simply define

$$S \subseteq U$$

Definition 5.3. A **target segment** T is just that segment that has been chosen for certain marketing purposes. It consists of a set of profiles that satisfies the target condition

$$T = S \subseteq U$$

We define the set of **excluded profiles** E, the set of profiles that does not satisfy the target condition, as

$$E = U - T$$

5.2 Relationship between model and reality

A model is by definition, and so is our model, an abstract representation of reality. It represents the reality by taking out the essentials from the real world. We try to reason about the visitors in reality using this model, therefore it is worthwhile to identify where the differences are between model and reality. In this section we discuss how our model relates to the real world, considering the following aspects

- the relationship between the reality and the model
- the completeness of knowledge in the model
- the validity of knowledge in the model

Relationship between the reality and the model

A visitor is an actual person that for example likes some books, has an age, visited a few pages and bought some products in the web shop. The visitor profile model represents this visitor by capturing the contextually **relevant** and **measurable** properties. This usually includes information about actual visitor's interests, characteristics and interactions with the system. In the ideal situation all relevant information is either implicitly or explicitly provided to the system. However, visitor profiles are built up progressively and our model only incorporates **observed facts**, the things we have learned about each visitor up to a certain point in time.

Completeness of knowledge in the model

We do not have complete knowledge about reality. We have discussed that profiles are built up progressively, this means knowledge about the visitor is gathered incrementally instead of all at once. Furthermore some facts may be measurable, and our model is capable of representing those facts, but are practically infeasible to acquire. In large-scale practical applications it is impracticable to get to learn the ages of all our visitors, or even harder all books that they like. In contrast, it is relatively easy to capture all visited pages or all bought products for all visitors.

Validity of knowledge in the model

Interests and characteristics change over time. So facts that we observed in the past may no longer hold in the current situation. To correct this the model might incorporate the **invalidation** of observed facts. This is a rather tricky problem that we do not discuss here. We assume that everything we observed about a visitor in the past is still true or valid in the future.

5.3 Analysis of segments

Practically speaking, in this research we are interested in the differences between two groups of website visitors, which are actual human beings. This would be rather straightforward when we would have access to all information about these visitors. However, as we previously discussed, our representation of reality only incorporates the things we have learned about each visitor up to a certain point in time. This means our knowledge about the visitor is incomplete and this brings a large amount of uncertainty about the visitors. Despite this uncertainty our goal is to make reliable statements about group differences. Therefore our method is consciously influenced by the progressive nature of the profile dataset.

In the next sections we will discuss how to analyze the differences between two groups of progressive profiles, called segments. First we will distinguish some segment types based on their dependency. We have defined a segment as a set of profiles, $S \subseteq U$. Note that we use the terms groups or sets of progressive profiles and segments interchangeably. We will see that when we take two segments these may be somehow related to each other. Our approach is always based on two independent segments.

5.3.1 Segment types

Let A and B be two segments: $A, B \subseteq U$. These are two sets that contain visitor profiles. If two sets A and B have no elements in common, we say that A and B are disjoint or independent. This is when the intersection of the two sets is empty

$$A \cap B = \emptyset$$

We use this terminology from set theory to say that two segments A and B are **independent segments** when they have no profiles in common.

Independent segments

We can split all visitors U for which the gender is known into two segment, a segment A of men and a segment B of women. In that case we have $A \cap B = \emptyset$, so these two segments are independent.

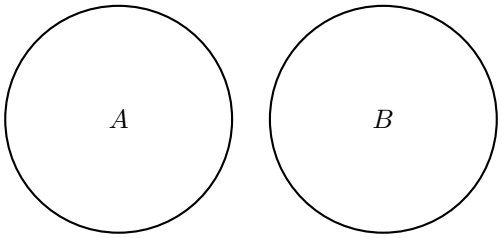


Figure 5.1: Two independent segments

The target segment T and the excluded profiles E are independent segments by definition.

Dependent segments

In case we have two more arbitrary segments, these segments will usually intersect. Two segments are called **dependent segments** when they have one or more profiles in common. For example, let A be the segment of all visitors that visited the "contact" page, and B the segment of all visitors that visited the "news" page. The intersection of both segments $A \cap B$ is equal to the visitors that visited both pages.

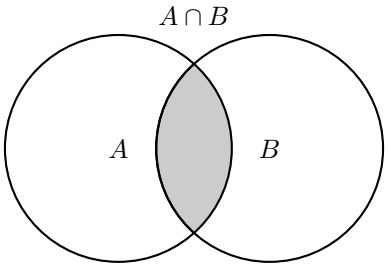


Figure 5.2: Two dependent segments

We can define the relative complements A^* and B^*, which are the nonintersecting parts of both segments

$$A^* = A - (A \cap B) = A - B$$

$$B^* = B - (A \cap B) = B - A$$

A special case is when a segment is a subset of another segment, we refer to this as **segment containment**. When A is a subset of B, or A is contained inside B, that means all profiles in A are also in B.

$$A \subseteq B$$

$$A \cap B = A$$

This is graphically visualized as

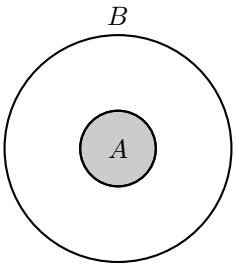


Figure 5.3: Two dependent segments, where A is contained inside B

We see that the nonintersecting part of A and B, is equal to $B^* = B - A$. Our model assumes segments to be independent. In case two segments intersect, we will always try to compare two nonintersecting parts, which are independent. See the figure 5.4.

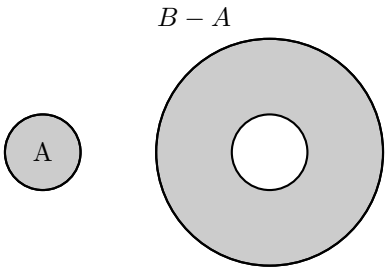


Figure 5.4: The two nonintersecting parts of figure 5.3

5.3.2 Segment differences

In the remainder of this chapter we will connect the theory we have discussed in the previous chapters with our model. The progressively accumulated profiles are an elegant way to incrementally gather profile data over a longer time period. The direct consequence of this approach is that the profiles in our dataset will vary greatly in the amount of information they contain. Another way to state this is that only a small fraction of profiles have values assigned to a certain profile attribute. Now from a segment point of view this means we usually only have information about a small subset of the profiles that belong to that segment. Moreover, in many cases we do not exactly know which profiles belong to a certain segment. As a consequence our model contains a lot of uncertainty about whether the profiles reflect the true characteristics of that segment in reality. We can look at it as being the segment the population we want to make statements about. We operationalize these "statements about the population" by making statements on attribute level. Now because of progressive profiling we will only have data about a small part, a sample, of the population on each individual attribute. Our aim is to use this data to generalize about the population. We approach this by considering the aspects of **(un)certainty** about the difference and **magnitude** of the observed differences. Fortunately the statistics we have discussed offer the tools to address these aspects. Because the substantial difference between these attribute types, we will discuss numeric (quantitative) and categorical (qualitative) attributes independently of one another. This discussion does not include multi-valued attributes. Multi-valued

attributes should be transformed into multiple single-valued attributes to be analyzed.

Assumptions

We start by formulating the conditions under which the statements apply. Although the discussed theory (e.g. ANOVA) enables us to compare multiple groups, we only consider two groups in our discussion. Now we assume that

samples are independent The samples being compared have no subjects in common.

numeric attributes are normally distributed If a numeric attribute is considered it is assumed that the population follows a normal distributions. This will hold for many numeric attributes, but obviously not for all numeric attributes.

sample subjects are randomly taken from the population The statistics all have the condition that a random sample is taken from the population. We violate this assumption because we say that our sample consists of all data we have progressively gathered about the profiles that belong to a segment.

In the remainder when we talk about a sample we mean all data that we have about the concerning segment on a certain attribute, here each value represents a profile.

5.3.3 Differences on numeric attributes

Uncertainty about mean value

Single-valued numeric attributes either contain one numeric value or are empty. We begin the discussion with how we can express the uncertainty about the mean value of a single sample. We know the sample mean is always an estimate of the population mean and varies from sample to sample. Informally stated our sample consists of all information we have progressively gathered up to a certain point in time about a certain characteristic. Formally, when S is a segment and $a \in A$ is a numeric attribute, then $S(a)$ consists of all values assigned to this attribute for all profiles $u \in S$.

Definition 5.4. We define that $\tilde{\mu}$ is the estimated mean of this segment, and $\tilde{\sigma}$ is its standard deviation (equation (3.1) and (3.7))

$$\tilde{\mu} = \frac{1}{|S(a)|} \sum_{x \in S(a)} x \tag{5.1}$$

$$\tilde{\sigma} = \sqrt{\frac{1}{|S(a)| - 1} \sum_{x \in S(a)} (x - \tilde{\mu})^2} \tag{5.2}$$

This mean is just a point estimate, our best guess, of the mean that we would get if we would have complete knowledge (e.q. no sampling error).

We have discussed the concept of standard error in section 3.6, recall that it provides us with a quantitative value of the error that we reasonably would expect when we take a sample of this size from the population:

$$\frac{\tilde{\sigma}}{\sqrt{|S(a)|}}$$

We can use the concept of **confidence intervals** to make an educated prediction about the upper and lower bounds of the population parameter. Usually researchers use confidence interval of either 95% or 99% (95% or 99% CI). With a CI of 95% we expect in 95% of the cases, 19 out of every 20 times, (we take a sample of the same size from the population) that our confidence interval covers the true population parameter value. This does not mean that the interval has a 95% chance of containing the true parameter value. The confidence interval either contains μ or does not contain μ. However, there is 95% chance of creating an interval that does contain μ.

Definition 5.5. We determine a 95 CI estimation with upper and lower bound using

$$\tilde{\mu}_{lower} = \tilde{\mu} - t_{95}\frac{\tilde{\sigma}}{\sqrt{|S(a)|}} \tag{5.3}$$

$$\tilde{\mu}_{upper} = \tilde{\mu} + t_{95}\frac{\tilde{\sigma}}{\sqrt{|S(a)|}} \tag{5.4}$$

where $t_{95}\frac{\tilde{\sigma}}{\sqrt{|S(a)|}}$ is also called the margin of error. The value t_{95} can be calculated using the family of t distributions, we will not discuss that in detail here. Note that $\frac{\tilde{\sigma}}{\sqrt{|S(a)|}}$ is the standard error.

The width of the confidence interval reflects the **precision** of the estimate. A narrower interval indicates a more precise point estimate; wider intervals reflect greater uncertainty about the estimate. Usually increasing the sample size, which increases reliability, will narrow the interval with, and so increases precision. Confidence intervals do not correct or control for inadequate sampling design, which we characterized as sampling bias. If our sample is biased the actual error may be greater than the CI indicates.

Certainty and magnitude of difference

It is relatively easy to see whether or not there is a difference between two point estimates, such as two sample means. However, we have seen that the estimation of these points involves uncertainty, and so the difference between these points. More interesting is whether the observed difference in the sample actually reflects a true difference in the population. What we ultimately want is to make a certainty statement about the observed difference.

The second thing is whether an observed difference should be considered large or small in practise. Obviously only domain experts can judge about whether a difference is large or small. Moreover, some differences are considered small by some while considered large by others. We will discuss some general approaches to determine the magnitude of effect.

We know that even when we randomly draw samples from the population, so everyone in the sample has an equal chance of being assigned to either of the

treatment groups, our observations will vary from sample to sample. Recall that we hypothesized that the two treatment groups are the same, so that difference among observations are simply due to chance. Now the p-value provides us with a likelihood that the observation (the difference between the means) is due to chance alone. The consideration is whether we have enough evidence to reject our null hypothesis and conclude that the difference is not due to chance alone. Statisticians often use a rather pessimistic approach and use a significance level of 0.05. This means the null hypothesis is rejected when a difference as extreme or more extreme could have happened by chance alone less than 5% of the time. As Greenfield et al [7] state

> The cut-point or significance level of 0.05 is arbitrary, and may ignore important, clinically meaningful findings.

Note that the rejection of the null hypothesis does not imply that we conclude that the two groups are the same. That we do not proof a difference does not mean a proof of no difference. A lower p-value (than the α-level) means we think there is enough evidence, so we are enough confident to conclude a difference between the groups.

Definition 5.6. We define the concept of **confidence in difference** between the segments on a numeric attribute as

$$\text{conf}_{\text{difference}} = 1 - p \tag{5.5}$$

Here the value of p is calculated using the independent samples t-test (section 3.7.2) and Student-t family of distributions. A $\text{conf}_{\text{difference}}$ of 0.99 means that there is still a 1% probability (1 out of 100) to see a difference greater than this that is caused by random chance alone.

With large sample sizes we can be very confident about even very small differences. Now such small differences might be considered meaningless in practise. That is why we also assess the effect, or practical value, of the difference.

Definition 5.7. We use Cohen's d effect size (section 4.1.1) to assess the **magnitude** or **effect of the difference**. Variables $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ are the mean and standard deviation of the first and second segment on a numeric property. The first segment has a value on n_1 properties, and the second segment on n_2 properties. Now the formula for the effect size is given by

$$\frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\sqrt{\frac{\tilde{\sigma}_1^2(n_1-1)+\tilde{\sigma}_2^2(n_2-1)}{n_1+n_2-2}}} \tag{5.6}$$

5.3.4 Differences on categorical attributes

There is no way to summarize categorical attributes by providing a central position within the data. A good alternative is to report the mode, which is the value that appears most often in a set of data. In this model we solely focus on the difference between two segments on a categorical property. Just as with numeric attributes we define the concepts of confidence in difference and effect of the difference here for categorical attributes.

Definition 5.8. For the **confidence in difference** we rely on the chi-square test of independence, discussed in section 3.9.4. With this test we try to assess whether there is a dependency between being part of either the first or second segment and the value on the categorical attribute. If there is a strong dependency between the segments and the attribute, this means there is a large difference between the two segments on this categorical attribute. Using the chi-square distribution the p-value can be computed with the test statistic. The p-value is the probability of observing a sample statistic as extreme as the test statistic. Now the formula for the confidence in difference is equal to that of numeric attributes.

$$\text{conf}_{\text{difference}} = 1 - p \tag{5.7}$$

Definition 5.9. We have discussed Cramér's V effect size, which measures the strength of association for nominal categorical variables, in 4.2.1. This measure is used to assess the **effect of the difference** for categorical attributes. In our model we limit ourselves to the consideration of two segments. This means the smallest number of categories is always two. Consequently, using $k = 2$, Cramér's V formula in our model is equal to phi, ϕ

$$V_{k=2} = \phi = \sqrt{\frac{\chi^2}{n}} \tag{5.8}$$

The definition of the χ^2 formula is provided in equation 3.11. Here n is equal to the number of profiles in the two segments that we can classify in one of the categories of the categorical attribute.

5.4 Interpretation of model statistics

When analyzing data in the right way and with the right techniques, possibly interesting findings can be uncovered. As we have seen in the previous sections, the result of a statistical analysis is usually just a set of numbers on some statistics. Comprehensive understanding of these numbers requires knowledge on how these statistics are computed. Therefore just presenting these raw numbers in an actual application to the user requires a high level of prerequisite knowledge. This is rather undesirable because it increases the barriers to use the system. It is worthwhile to investigate the possibilities the present the result in a more user-friendly way. In this section a guideline for translating the statistical results to a semantic meaning is provided.

5.4.1 Interpretation and reporting

In order to clarify statistical findings these may be represented using tables and charts. Textual expressions or statements may help to summarize and interpret these results. Miller [13] provides guidelines on how best to present information on both statistical and substantive significance of regression results. As he clarifies

> An important part of writing a thorough description of multivariable regression results involves striking the right balance between presenting inferential statistical results and interpreting the substantive meaning of those results in the context of the particular research question.

According to Miller the description of a research finding should include

concepts one should write in terms of specific real-world concepts (male, female, age, income) instead of generic references to the dependent and independent variable.

units of measurement incorporating units of measurements (years, kilograms, euros) makes it easier to assess real world implications of the findings.

direction knowing that there is an association between the independent and dependent variable usually is not enough. To make the result more actionable it is necessary to include the direction of association. This provides information on questions such as: do males have higher income than females, or vice versa? Is there a positive or negative relation between age and income?

magnitude Besides the direction of association it is interesting to assess the magnitude of association: is the difference large or small in practise. Very small differences may be considered statistically significant with large sample sizes. Small differences however are usually not very interesting in practise. Instead of observing that males have higher incomes than females, we want to know whether the incomes of males are extremely much higher or just slightly higher than females.

statistical significance After reporting the subjective aspects it is still important to report the statistical significance statistics which give us the confidence in the existence of the association or difference.

Using Miller's approach in the context of this thesis we want to make clear statements like (these examples are just for illustrative purposes)

> "People who clicked on the banner are a much older (8.4 years at average) than people that did not click ($p < .05$)."
>
> "People who clicked on the banner spend little more time on the site (12 seconds at average) than people that did not click ($p < .05$)."
>
> "People in the target segment visit a few more pages (4 pages at average) than people that are not in the target segment ($p < .0001$)."
>
> "Males are more likely to click (20% of males) on the banner than females (4% of females) ($p < .05$)."
>
> "People who spend 80-130 seconds on the page are much more likely to click the banner than people who spend 20-60 seconds ($p < 0.05$)."

Although these textual statement provide a clear interpretation of the findings, these should always be supported by tables and charts which provide better insight into the data. Based on some discussions during the internship at GX Software, it appears that significance levels are hard to understand for marketeers. An alternative approach for interpreting significance levels is discussed next.

5.4.2 Levels of statistical significance

Recall that the conventional significance level of 5% ($\alpha = 0.05$) is rather arbitrary. Other popular levels of significance in scientific literate are 10% (0.1), 1% (0.01), 0.5% (0.005) and 0.1% (0.001). A common approach is to use the "three-star system" to indicate the different levels of significance [10]. Here $*$ indicates $p \leq .05$, $**$ means $p \leq .01$ and $***$ indicates $p \leq .001$. Sometimes four asterisks are used that indicate significance at .0001 level. The smaller the p-value, the greater the significance and thus our "confidence" in the existence of a difference or association. More precisely formulated: the closer the p-value is to zero, the more confidence we have that the difference is not caused by chance alone. We propose the following system, which is a little more extensive and also provides a practical interpretation This scheme enables to directly translate the

Symbol	P-value	Meaning	Confidence in difference
ns	> 0.1	not significant	not confident
nqs	≤ 0.1	not quite significant	not really confident
*	≤ 0.05	quite significant	little confident
**	≤ 0.01	significant	confident
***	≤ 0.001	clearly significant	very confident
****	≤ 0.0001	very clearly significant	highly confident
*****	≤ 0.00001	extremely clearly significant	extremely confident

Table 5.1: Interpretation table for statistical significance

inferential statistics to their practical meaning. However, one must always keep in mind that the results are based on some assumptions that may have not been met. Even with very significant result, one should be suspicious about how the data has been obtained. To be completely clear: we do not intend to provide perfectly validated scientific results, our goal is to uncover relevant insights for marketeers.

5.4.3 Levels of effect

In chapter 4 we have already provided interpretations for different levels of effect sizes. These levels are used to report whether we think an effect is negligible, weak, moderate, strong or very strong. In order to help the user to interpret the results we also construct natural language statements as proposed by Miller [13]. The grammar of this language is fairly simple. For numeric properties the structure of the statements is: "people in `segmentA` have `effect` values on `property` than people in `segmentB`". Here `segmentA` is the name of the first segment, `segmentB` is the name of the second segment and `property` is the name of the property under analysis. Depending on the value of the statistic, `effect` is translated to one of the following values

- "very slightly higher"
- "slightly higher"
- "higher"

- "much higher"
- "very much higher"

An actual example of a statement we have generated using this grammar is "people in Age between 50 and 60 have slightly higher values on Average order value than people in Age between 20 and 30".

For categorical properties the statements are formulated differently because we do not talk about differences in values in this case: "`property` is `effect` with being subject of either `segmentA` or `segmentB`". The `effect` is translated to

- "very weakly associated"
- "weakly associated"
- "associated"
- "strongly associated"
- "very strongly associated"

An actual example of a statement that we have generated for a categorical property is "Searched for: www is weakly associated with being subject of either Age between 20 and 30 or Age between 50 and 60". In contrast to the Miller's advice we do not incorporate units of measurements. The reason behind this is that this information is not available to us in the prototype.

Chapter 6

Distiller: applying the model

In the previous chapter we have described a theoretical model for the assessment of differences between independent groups of progressive profiles, which we refer to as independent segments. Now we will consider the practical value and application of the discussed theory. We will discuss how these statistical findings can be converted into relevant insights and knowledge in a business context. One of the goals of the research was to make the rather theoretical findings of this study tangible. We will consider how the model can be incorporated in an actual application, and we build a prototype given a set of requirements that uses actual data. This application is named **Distiller**, as it finds the properties that distinguish two segments from each other. The output of this prototype is carefully analyzed and discussed, and acts as a guide for recommendations to GX Software.

6.1 Practical application

In this chapter we will concretize the practical value of the discussed theory. In order to do this we will construct a prototype that works for both simulated and actual data. The main goals that we try to achieve with Distiller are to

- demonstrate that the literature and theory that is discussed in this thesis, and the resulting model, can be incorporated in an actual application
- demonstrate that the model works on simulated data
- demonstrate that the model works on actual data
- demonstrate that the analysis leads to both meaningful and interpretable results
- investigate on which aspects the model falls short in practice, which allows us to reason about areas for improvement

6.2 Technical requirements

In an actual situation a segment can consist of millions of profiles. GX relies on open source Apache projects, including Apache Solr and Apache Cassandra, to handle this large amount of data. An important selling point for GX Software is that the product operates in real-time: the user should not have to wait for an analysis to be computed, so everything should be available within milliseconds. However, we have only limited time and resources to build the prototype. Our prototype should run on a single machine with limited memory and computation power. Furthermore we do not have access to the profile database directly, but instead we have to make use of CSV exports. Taking all this together, we are satisfied when our prototype can read, analyze and export 150 properties in two segments with up to a maximum of 5000 profiles per segment, in reasonable time (less than 5 minutes execution time).

The analysis part of the prototype is written in JAVA. For the statistics and algorithms discussed in the previous chapters we use existing libraries, when available, instead of writing these ourselves. Two main libraries we use are

Commons Math is a library of lightweight, self-contained mathematics and statistics components addressing the most common problems not available in the Java programming language or Commons Lang. We heavily use this library for its implementation of many mathematical or statistical functions, distributions and algorithms that are discussed in thesis.

WEKA (Waikato Environment for Knowledge Analysis) [8] is an open source library for machine learning. We use this library solely for its implementation of Fayyad & Irani's MDL supervised discretization algorithm, which we discussed in section 2.5.2.

The results of the analysis are written to a JSON file. Furthermore we provide a web-based user interface, that presents the results of the analysis to the user. This part is written in HTML, CSS, Javascript, and uses jQuery and the Google Visualization API. The **Google Visualization API** enables use to draw tidy charts that clarify the statistical findings.

6.3 The operation of the prototype

Now we will discuss the functionality of Distiller. Let us first define *what* our application should do

> Distiller should find those properties in which two sets of progressive profiles are most dissimilar. These findings should be presented to the user in a clear and understandable manner

The remainder of this section discusses *how* we approach this in a real application, and illustrate this with the prototype. In this discussion we consider our prototype as two individual parts: an analysis part and a presentation part, which we respectively refer to as the analyzer and the presenter. The interesting computations take place in the analyzer. The discussion about the analyzer focuses on how the data is processed and the statistics are computed. Which is essentially a translation of our model to source code. The presenter takes care

of presenting the results of the analyzer to the user. Here the emphasis is on the important considerations regarding the user interface.

6.3.1 Analyzer

The input of the **analyzer** is two sets of profiles, with for each profile its values on a set of properties. We have formally defined a profile in section 5.1. These sets of profiles can be either simulated or constructed on the basis of real data.

Listing 6.1: Simulation of values

```
public int[] generateNumericValues(int n, int lowerBound, int
    upperBound, int bins, int[] dist);
public String[] generateCategoricalValues(int n, List<String> list
    , int[] dist);

// generate 100 numeric values between 0 and 80
generateNumericValues(100, 0, 80, 6, new int[] { 15, 25, 25, 25,
    5, 5 });
// generate 100 categorical values (10% yes, 90% no)
List<String> YESNOMAYBE = Arrays.asList(new String[] {
      "yes", "no", "maybe" });
generateCategoricalValues(100, YESNOMAYBE, new int[] { 10, 90, 0})
    ;
```

Simulated profiles For the simulation of profiles we have defined two generation functions whose definitions are listed in 6.1. These functions can generate a specified number of values on numeric and categorical properties.

The `generateNumericValues` function generates n numeric values between the lower and upper bound. Here the range between the lower and upper bound is equally divided in a specified number of bins. Now the distribution array tells what proportion of values to generate for each these bins. So if we have four bins, with lower bound 0 and upper bound 80, we have four bins of length 20. Our distribution may prescribe to randomly generate 30% of the values in the first bin, 40% in de second bin, 20% in the third bin and only 10% in the last bin.

For the `generateCategoricalValues` function the bins are given in advance by a list of the possible categories. Now the distribution array says what proportion of values to generate for each of the categories.

Each of the values that we generate represents a single profile. So with a n of 100 we have generated values for 100 profiles on a certain property. By calling the function twice we can generate values for two segments on a certain property. We can now simulate differences between two segments by varying the distribution of values over bins to a greater or lesser extent.

Actual profiles BlueConic provides the ability to export sets of profiles to a comma-separated file (CSV). Our prototype can read up to a maximum of about 7000 rows of profiles and 180 columns of profile properties. For numeric properties empty values are not considered. For categorical properties, in some cases the empty value is not considered, in others the empty value is seen as a separate `EMPTY` category. We have manually indicated the desired behavior for each property.

As we have seen in the discussion of the model, the analysis is strongly dependent on the type of the profile property. Therefore the system must have knowledge of the type of each profile property: either numeric or categorical. This can be either indicated in advance or, with some uncertainty, determined based on the data. For example, with our automatic detection technique, integer ID numbers will be interpreted as numeric while in fact these should be considered categorical. We do not have access to property type information, so the prototype derives these types for us.

Details on how the data is parsed are not considered in this discussion. The final result of this process is two segment objects, both consisting of a set of profile objects. These profile objects have values assigned to a set of profile properties. In the next sections we comprehensively describe how the profile properties are analyzed.

Numeric property analyzer

For clarification, we will provide some parts of the source code. Many details however are hidden in the packages we have used, and some code fragments are simplified for illustrative purposes.

We have declared two variables `myFirst` and `mySecond`, that refer to the two segment objects. Math3's `DescriptiveStatistics` class is used to compute the summary statistics for a segment on a numeric profile property. Only the profiles within that segment that have a value on that profile property are regarded here. This statistics class includes the sample mean, standard deviation and sample size. We have extended this class to support the 95% confidence interval for the mean. This method is shown in listing 6.2. Listing 6.3 shows how we can request the descriptive statistics on a certain numeric property `prop`.

Listing 6.2: Confidence interval

```
/**
 * Compute the width of the confidence interval for the current
     numeric attribute
 * @return  confidence interval width
 */
public double getConfidenceIntervalWidth()
{
  // t distribution with n-1 degrees of freedom
  TDistribution tDist = new TDistribution(getN() - 1);
  // compute the probability quantile function of this
      distribution
  double a = tDist.inverseCumulativeProbability(1.0 - 0.05 / 2);
  return a * getStandardDeviation() / Math.sqrt(getN());
}
```

Listing 6.3: Segment descriptive statistics

```
// get the summery statistics on both segments
ExtendedDescriptiveStatistics statsSegmentA =
  myFirst.getStats(prop);
ExtendedDescriptiveStatistics statsSegmentB =
  mySecond.getStats(prop);
```

These descriptive statistics are used to individually summarize the two segments on a numeric property. Furthermore these are the foundation of many other

statistics. Recall we have already extensively discussed these statistics in chapter 3.

The remainder will focus on the difference between the two segments on a numeric property. The most simple way to describe the difference between two numeric properties is the mean difference, which implementation is shown in listing 6.4. However this statistic does not tell us anything about the certainty and effect of the difference. As discussed in the model, for this we use the two-sample t-test (listing 6.5) and Cohen's d effect size. The Cohen's d effect size (listing 6.8) measure incorporates the pooled variance or pooled standard deviation (see section 4.1.1), which are listed in 6.6 and 6.7

Listing 6.4: Mean Difference

```
/**
 * difference between means of two segments
 * @return mean difference
 */
double getMeanDifference() {
  DescriptiveStatistics statsSegmentA = myFirst.getStats(prop);
  DescriptiveStatistics statsSegmentB = mySecond.getStats(prop);
  double  m1 = statsSegmentA.getMean(),
          m2 = statsSegmentB.getMean();
  return Math.abs(m1 - m2);
}
```

Listing 6.5: Independent sample t-test

```
/**
 * Get the p-value for a independent sample t-test on the
     specified property.
 * @param prop
 * @return p-value
 */
double tTest() {
  TTest test = new TTest();

  DescriptiveStatistics statsSegmentA = myFirst.getStats(prop);
  DescriptiveStatistics statsSegmentB = mySecond.getStats(prop);

  return test.tTest(statsSegmentA.getValues(), statsSegmentB.
      getValues());
}
```

Listing 6.6: Pooled variance

```
/**
 * Computes the pooled variance.
 * @param s1  variance first segment
 * @param s2  variance second segment
 * @param n1  sample size of first segment
 * @param n2  sample size of second segment
 * @return    the pooled variance of segment s1 and segment s2
 */
static double getPooledVariance(final double v1, final double v2,
    final double n1, final double n2)
{
  return ((n1 - 1) * v1 + (n2 - 1) * v2) / (n1 + n2 - 2);
}
```

Listing 6.7: Pooled standard deviation

```
/**
 * Computes the pooled standard deviation.
 * @param s1  standard deviation first segment
 * @param s2  standard deviation second segment
 * @param n1  sample size of first segment
 * @param n2  sample size of second segment
 * @return    the pooled standard deviatin of segment s1 and
     segment s2
 */
static double pooledStandardDeviation(final double s1, final
    double s2, final double n1, final double n2)
{
  double v1 = Math.pow(s1, 2), v2 = Math.pow(s2, 2);
  return FastMath.sqrt(getPooledVariance(v1, v2, n1, n2));
}
```

Listing 6.8: Cohen's d

```
/**
 * Computer cohen's d effect size
 * @return cohen's d effect size
 */
public double cohensD() {
  DescriptiveStatistics statsSegmentA = myFirst.getStats(prop);
  DescriptiveStatistics statsSegmentB = mySecond.getStats(prop);

  double m1 = statsSegmentA.getMean(),
    m2 = statsSegmentB.getMean();

  double poolStandardDeviation = pooledStandardDeviation(
    statsSegmentA.getStandardDeviation(),
    statsSegmentB.getStandardDeviation(),
    statsSegmentA.getN(),
    statsSegmentB.getN()
  );
  return (m1 - m2) / poolStandardDeviation;
}
```

Furthermore the numeric property analyzer has several complex methods for the discretization of a set of continuous numeric values into discrete intervals. Note that we have two sets of values on a numeric property: the values on that property for the first and the second segment. So for each value of the numeric property we know whether the value belongs to the first or the second segment (this corresponds to the class label). The discretization algorithm tries to optimally split the continuous range into a few bins wherein each bin provides maximum information about the segment. To illustrate: when most values of the first segment are between 10 and 30, and most values of the second segment are between 40 and 60, the discretization algorithm will indicate these two ranges as bins. The prototype uses the Fayyad & Irani's MDL method which we have discussed in section 2.5.2 and is implemented in the WEKA library.

Categorical property analyzer

This section discusses the technical implementation of the categorical property analyzer. Most of the methods we describe here are part of the `CategoricalPropertyAnalyzer` class and frequently use the `counts()` method

that returns a two-dimensional `long[][]` array. This array contains for each category of the categorical property, how many profiles within the two segments belong to that category. For example, let us assume we have a categorical property with three possible categories: `yes`, `no` and `maybe`. Assume that 50 people in the first segment belong to `yes`, and 50 in the first segments belong to `no`. For the second segments, all 100 profiles belong to the `no` category. Now the counts array looks like this: `{{50,0},{50,100},{0,0}}`. We can use this two-dimensional array for a chi-square test of independence, and return the associated p-value. This method is shown in listing 6.9. The implementation of `ChiSquareTest` is given by the Commons Math library.

Listing 6.9: Chi-square test of independence

```
/**
 * Returns the observed significance level, or p-value, associated
     with a
 * chi-square test of independence
 * @return p-value
 */
public double chiSquare() {
  ChiSquareTest testStatistic = new ChiSquareTest();
  return testStatistic.chiSquareTest(counts());
}
```

For the effect of the difference we have implemented Cramér's V effect size measure. As in our situation `nCols` is always equal to two (the two segments), the number of degrees of freedom (`df = nRows < nCols ? nRows - 1 : nCols - 1`) is always equal to one. So in fact we use the phi measure of effect size.

Listing 6.10: Cramer's V

```
/**
 * Computes Cramer's V
 * @return Cramer's V of phi (for df=1)
 */
  public double CramersV() {
    long[][] counts = counts();
    int nRows, nCols, df;
    double n = 0;

    nRows = counts.length;
    if (nRows > 0) {
      nCols = counts[0].length;

      for (int row = 0; row < nRows; row++) {
        for (int col = 0; col < nCols; col++) {
          n += counts[row][col];
        }
      }
      df = nRows < nCols ? nRows - 1 : nCols - 1;

      if (n > 0 && df > 0) {
        ChiSquareTest testStatistic = new ChiSquareTest();
        return Math.sqrt(testStatistic.chiSquare(counts) / (n * df
            ));
      }
    }
    return 0; // error
  }
```

Furthermore the analyzer class has several utility functions for processing the categories, generating tables for the presenter and translating the results to their semantic meaning.

6.3.2 Presenter

The analyzer outputs two JSON files. The first file contains all data that is needed to report the results of the analysis to the user. The other file maps the system variable names to real names. For instance, `RANGE_PROPERTY_VISTIS` is mapped to `Monthly visits`. Both files are input to the presenter. This presenter is a small web-based application that can be opened within the web browser. This application is involved with processing the JSON files, building the HTML DOM and handling the interaction of the user. We do not provide details on its technical implementation here. Rather we will focus on the design choices regarding the user interface.

Our prototype can analyze up to a maximum of 150 properties. For a user to consider the results for those 150 properties is already a lot of work. We therefore want to make a ranking of most interesting to least interesting properties, so we can present it to the user accordingly. Now we have to consider what properties do we find most interesting. The properties are all scored on two points, the confidence in the existence of a difference, certainty of the difference, and the effect of the measured difference. It is possible to see a very large effect, but a very low confidence in the existence of the difference. This is usually the case when we have little data that is widely distributed among the two groups. In some cases our confidence is large, but the effect is small. This is common when we have a lot of data for both groups, but there is only a small or even negligible difference between the two groups.

It seems reasonable to take a mixture of the confidence score and the effect score as ranking criterion. The confidence score is always somewhere between 0 and 1. It must be close to one (≥ 0.95) to make us actually believe in the difference. Recall that the confidence score is equal to $1-p$. We have previously provided an interpretation table of this p-value in section 5.4.2. The effect size for categorical properties also ranges from 0 to 1. However for numeric properties the effect size can become greater than one. Categorical properties are considered large and very large at respectively 0.5 and 0.7, as this is 0.8 and 1.3 for numeric properties. Unfortunately, we have not been able to find a golden formula that combines these two scores, confidence and effect, in a single ranking score. We find that this would be an interesting topic for further research, which could also focus on whether marketeers would manually rank the results in the same way.

We have decided to display a list of all properties, and sort this list on the significance score. Consequently, the differences that we are most confident of appear on top of the list. One of the design goals is that the application should be easy to use and understand: no deep statistical knowledge should be required. Instead of presenting raw numbers, we focus on meaning and interpretation of the numbers. The confidence and effect of the difference on each property are therefore presented in natural language. To further clarify this we assign a color to each score on the 5 level scales, from green (best) to red (worst). Non confident and negligible results are displayed in gray. We also have a button in the interface to filter these results from the list. Figure 6.1 shows a screenshot

of the resulting list in the actual application, with actual data. We refer to this screen as the **list view**.

The second part of the presenter, the **detail view**, provides detailed information on each individual property. This view can be accessed by selecting one of the properties from the list. It is a concise report that displays in an insightful manner where the differences are between the two segments for a certain property. There are two versions of the detail view, a version for numeric properties and a version for categorical properties. An example report for a numeric property is found in three parts in figure 6.2, figure 6.3 and figure 6.4. In figure 6.5 and 6.6 the report for a categorical property can be found.

The report always starts with a statement in natural language that summarizes the finding, the structure of these statements is discussed in section 5.4.3. Instead of using colors now bars are used to indicate the strength on a five-point scale. There is a "show details option" to reveal the raw numerical results from the statistics. For numeric properties the report shows some summary statistics such as the mean and confidence intervals as we have discussed in the model. The power of textual reports is limited. We visualize the results using histograms to make clear at a glance how the data is distributed within the segments (see figure 6.3, 6.6). Furthermore the report summarizes categorical results in tables by row and by column. Let us discuss what the tables in figure 6.6 show. The interpretation of the first table is that 10.9% of the people with an age between 50 and 60 (835 profiles) searched for the keyword "www". Only 1.76% of the people between 20 and 30 searched for "www". The second table tells us that from all people (910 profiles) that searched for "www", 91.76% is between 50 and 60 while only 8.24% is between 20 and 30 years old. What we may conclude from this report is that people between 50 and 60 years old are more likely to search for the term "'www" than people between 20 and 30 years old.

Figure 6.4 shows the results of the MDL discretization technique. This table is only visible when the algorithm is able to distinguish a few characteristic intervals within this numeric property. We see that the algorithm has distinguished three groups. Proportionally many Feyenoord supporters (almost 50%) fall within the first category (visits between 0-105.5). Instead proportionally many Ajax supports are in the last category (visits between 298.5 - 4096). This emphasizes the finding that Ajax supporters have (slightly) higher values on Visits than Feyenoord supporters.

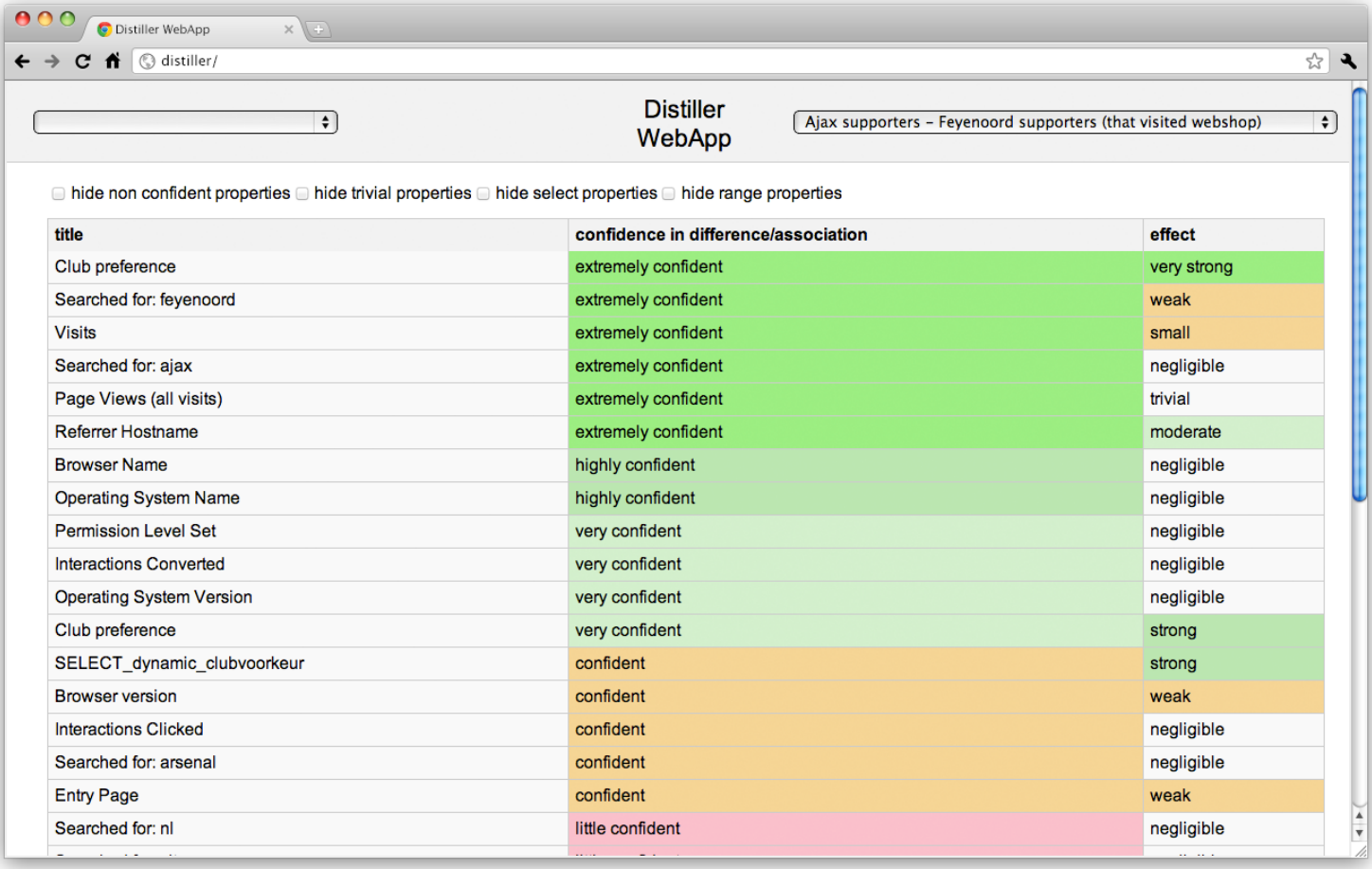
6.4 Findings of the prototype

We have tested our prototype on both simulated and actual BlueConic data. For the simulated data the prototype behaves as we had expected. Therefore the focus is on the findings using actual data. We have tested Distiller on four different datasets. These datasets are summarized in table 6.1. The columns "size first" and "size second" refer to the number of profiles within each segment. A few profiles were corrupted, so the actual numbers may be very slightly lower. In this final section we will summarize our general findings.

- We have assumed that the numeric attributes are normally distributed. It turned out when inspecting the datasets that many numeric attributes are not normally distributed. See for example figure 6.3, where most values are close to zero.

First segment	Second segment	Size first	Size second
Ajax supporters	Feyenoord supporters	789	3617
Men to 30 years old	Women to 30 years old	4561	2705
People between 20 and 30 years old	People between 50 and 60 years old	4259	7669
People that bought a product in shop between 6:00 - 10:00	People that bough a product in shop between 20:00-24:00	2656	4307

Table 6.1: Actual datasets exported from BlueConic



title	confidence in difference/association	effect
Club preference	extremely confident	very strong
Searched for: feyenoord	extremely confident	weak
Visits	extremely confident	small
Searched for: ajax	extremely confident	negligible
Page Views (all visits)	extremely confident	trivial
Referrer Hostname	extremely confident	moderate
Browser Name	highly confident	negligible
Operating System Name	highly confident	negligible
Permission Level Set	very confident	negligible
Interactions Converted	very confident	negligible
Operating System Version	very confident	negligible
Club preference	very confident	strong
SELECT_dynamic_clubvoorkeur	confident	strong
Browser version	confident	weak
Interactions Clicked	confident	negligible
Searched for: arsenal	confident	negligible
Entry Page	confident	weak
Searched for: nl	little confident	negligible

Figure 6.1: A screenshot of the Distiller WebApp list view

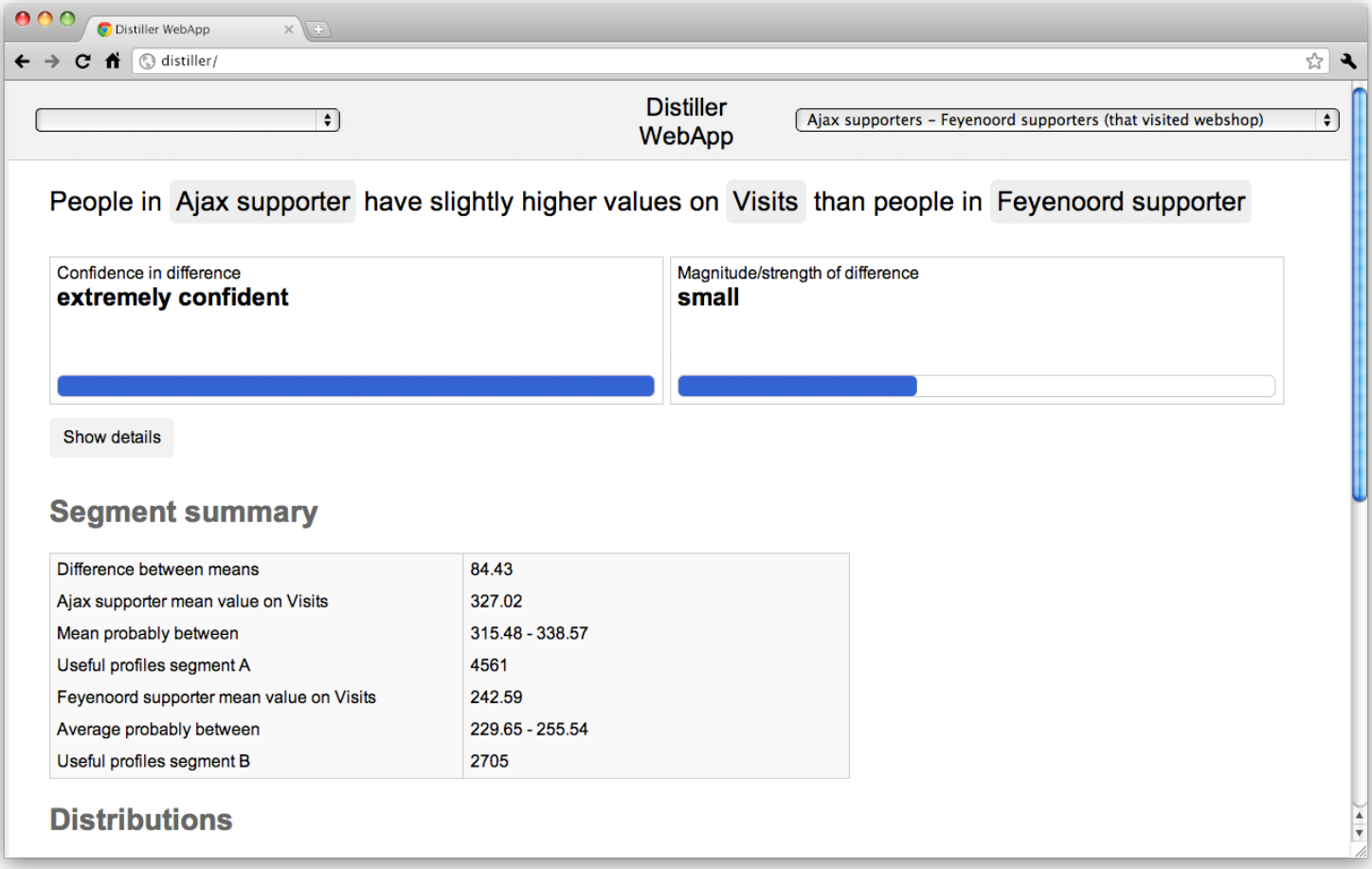


Figure 6.2: A screenshot of the Distiller WebApp detail view for numeric properties (part 1 of 3)

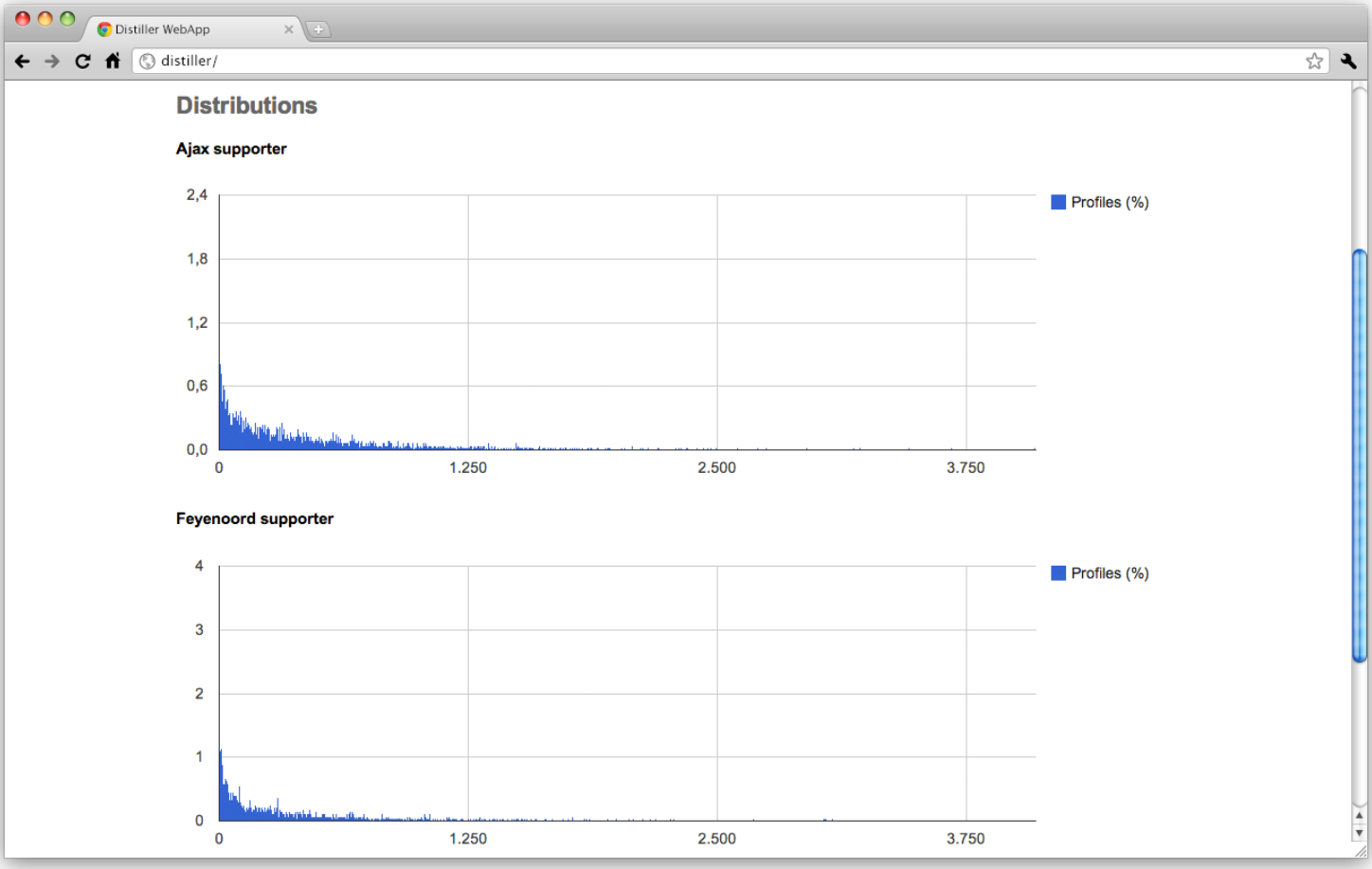


Figure 6.3: A screenshot of the Distiller WebApp detail view for numeric properties (part 2 of 3)

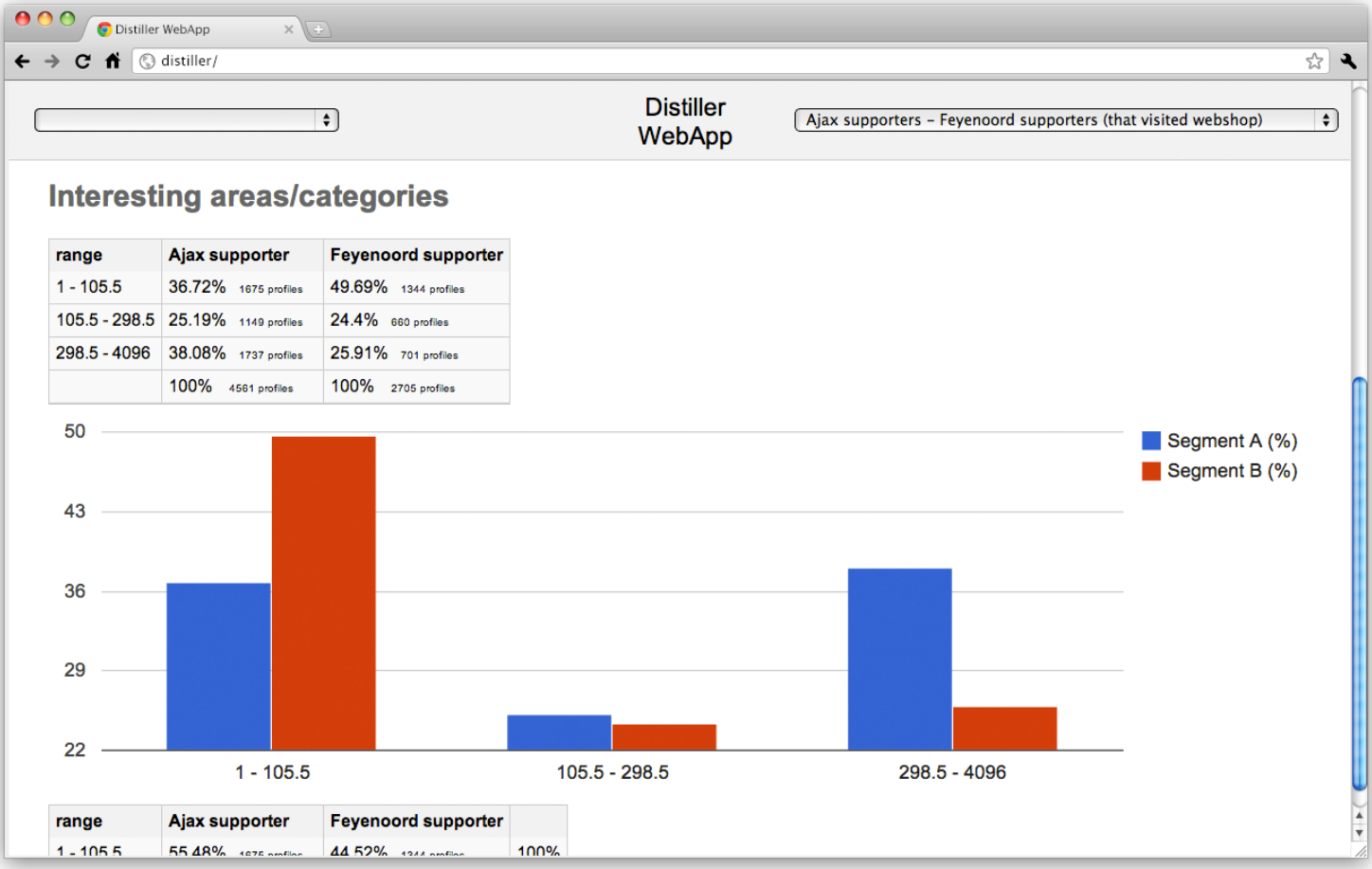


Figure 6.4: A screenshot of the Distiller WebApp detail view for numeric properties (part 3 of 3)

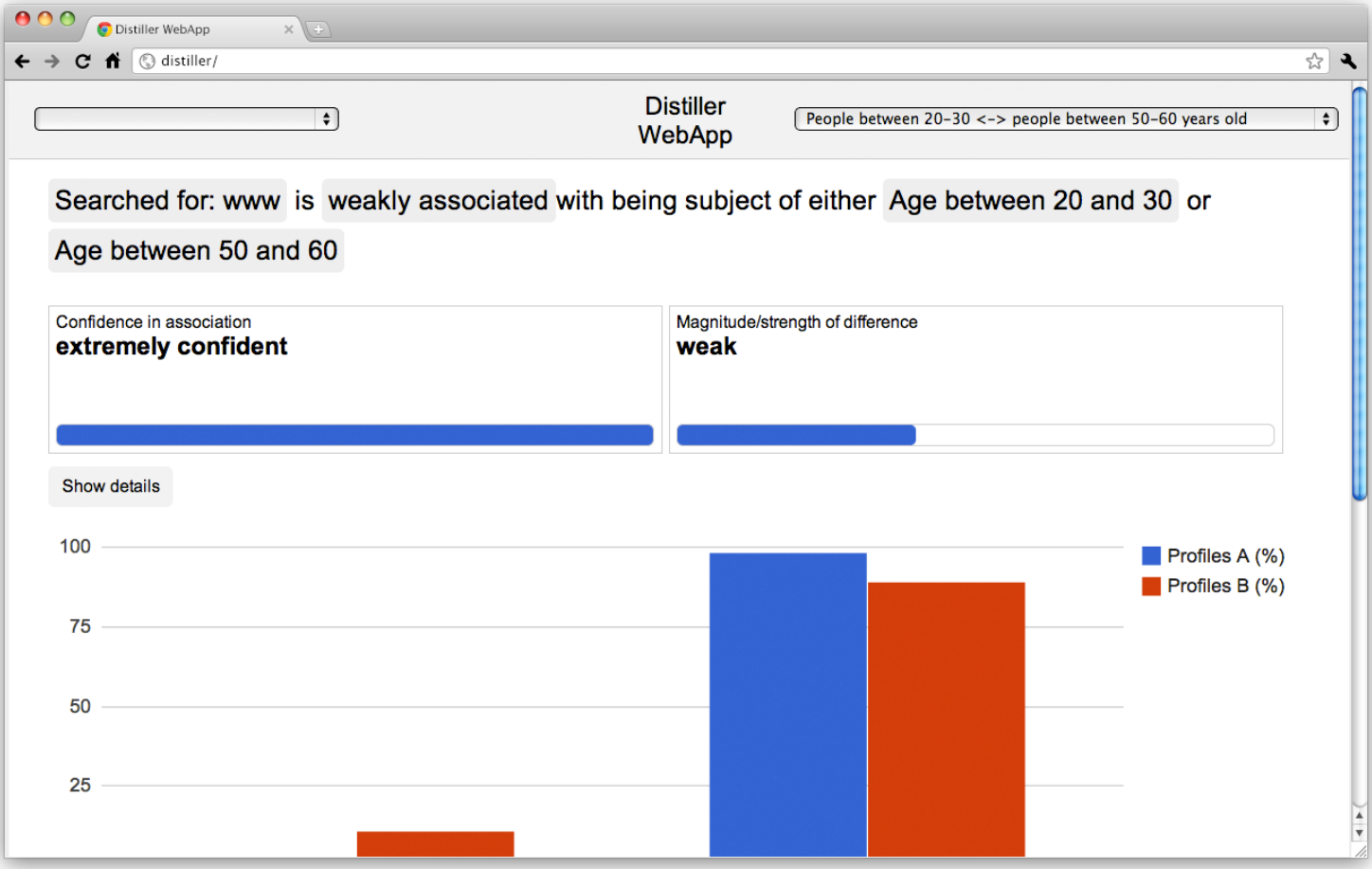


Figure 6.5: A screenshot of the Distiller WebApp detail view for categorical properties (part 1 of 2)

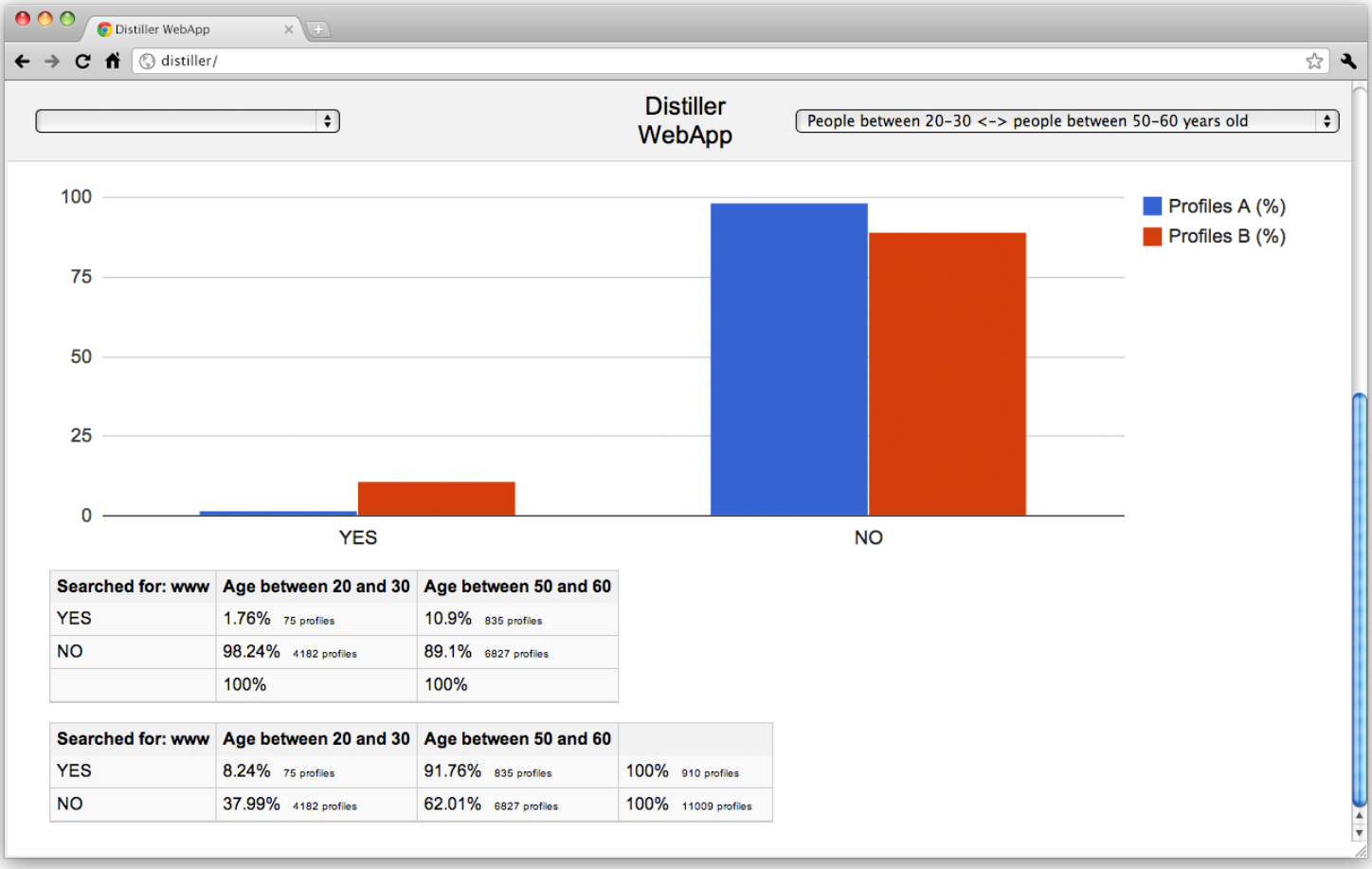


Figure 6.6: A screenshot of the Distiller WebApp detail view for categorical properties (part 2 of 2)

Our explanation is that many attributes in the dataset are **count attributes**. These attributes represent something that can be counted in whole numbers. Usually this is visitor behavior such as the total number of website visits, the total number of monthly visits, the total number of page views, the number of products bought this month, the number of ecards sent, etcetera. This data usually follows a **Poisson distribution**. Sun et al. discuss how they treat count data from Facebook in [19]. It is advisable to look into other statistics to assess these count attributes.

- For many instances the value on a profile attribute is empty. For some properties this means we do not know the value. The visitor just did not tell us the value so far. However in many other cases this emptiness has a certain meaning.

  To illustrate this: emptiness on the numeric property "product bought" means 0 products bought. For "visits this month" emptiness means the user did not visit the page this month, so the value should be zero. Emptiness on the categorical property "subscribed" means the profile belongs to the "no" category.

  We have manually indicated this for each property to improve the results of the analysis. However we advise to specify the meaning of emptiness at the definition of each property.

- The dataset has several properties that are multi-valued. These properties can contain more than a single value per profile. Examples are "search keywords entered", "pages visited" and "products viewed". We have not been able to find methods for the assessment of confidence in difference and effect of difference for multi-valued attributes.

  However an alternative that we have tested is to translate multi-valued attributes into multiple single-valued categorical attributes. We call these properties **virtual properties**.

  This is tested for the "search keywords" property. We have created 50 virtual properties for the top 50 most popular keywords. Now for each profile we have indicated for the fifty virtual properties whether (yes) or not (no) that keyword is in the profile's list of search keywords. The results are promising. With this method we have for example uncovered that: "Searched for: voetbal is (weakly) associated with being subject of either men to 30 years old or women to 30 years" and "Searched for: www is (weakly) associated with being subject of either Age between 20 and 30 or Age between 50 and 60" (see figure 6.5 and 6.6). It may not be surprising that males do search more for "voetbal" (football). But we cannot explain exactly why older people search for "www" more often than younger people.

- In some cases the effect size interpretation table is inconsistent with our interpretation. Figure 6.5 and 6.6 show that "Searched for: www is weakly associated with being subject of either Age between 20 and 30 or Age between 50 and 60" ($\alpha < 0.00001$). The associated Cramér's V or phi effect size is 0.16. The table shows that 10.9% of the more than 7000 profiles between 50 and 60 years old searched for "www". That is one out of ten profiles. In contrast, less than two in every hundred profiles

between 20 and 30 years old searched for "www". This is a difference of nine in a hundred profiles. Our intuition says this is a strong difference or association, however using the effect size interpretations discussed in this thesis it is considered as a weak effect. More research is needed to say something meaningful about the origin of this interpretation difference.

- Some categorical attributes have hundreds of categories. In addition, often there are just a few profiles that belong to each category. Analysis of this data with the chi square test of independence and Cramér's phi often leads to highly significant results with moderate to strong effects.

  An example of a categorical attribute with lots of categories is "Entry page". The entry page is the page where visitors arrive at the website. In an actual case, we have seen 1455 different entry pages, so 1455 categories, for the segments "Men to 30 years old" and "Women to 30 years old". For most entry pages, there was only one profile with that entry page in one of the two segments. The conclusion of the prototype is that "Entry Page is strongly associated with being subject of either Men to 30 years old or Women to 30 years old".

  We do not trust outcome of the statistics in this case, as there is very few data available for each category. Furthermore with so many categories the prototype is not able to present the results in an insightful way. This can be seen in figure 6.7. However the figure only shows 4 rows of the 1455 rows in total.

  A general rule of thumb, which is sometimes referred to as Cochran's rule, when using chi-square is that [26] [22]

  > No more than 20% of the expected counts are less than 5
  > and all individual expected counts are 1 or greater

  The data in our example certainly does not satisfy this rule. To increase cell frequencies we advise to combine categories, if possible. For example, with entry pages we could merge categories on domain name. An alternative is to only consider the top 5 or top 10 most frequent categories in analysis.

Although the prototype revealed several areas for improvement, overall, the results are promising.

First, with the application we have been able to identify on which properties the two segments are distinctive. In the four test cases the vast majority of the properties was considered statistically insignificant. This means that only for a small proportion of profile properties we were (moderate to highly) confident that there actually is a difference between the two segments. For example, for the segments with Ajax and Feyenoord supporters, only 20 of the 160 profile properties were considered significantly different. As the number of profiles grows, we see that more profile properties are considered significantly different between the two segments. With the segments "People between 20 and 30 years old" and "People between 50 and 60 year", 79 of the 184 properties were considered significantly significant. Nevertheless, this leads to an enormous reduction in the number of profile properties a marketeer needs to consider. Furthermore the application distinguishes valuable (the ones that have the greatest

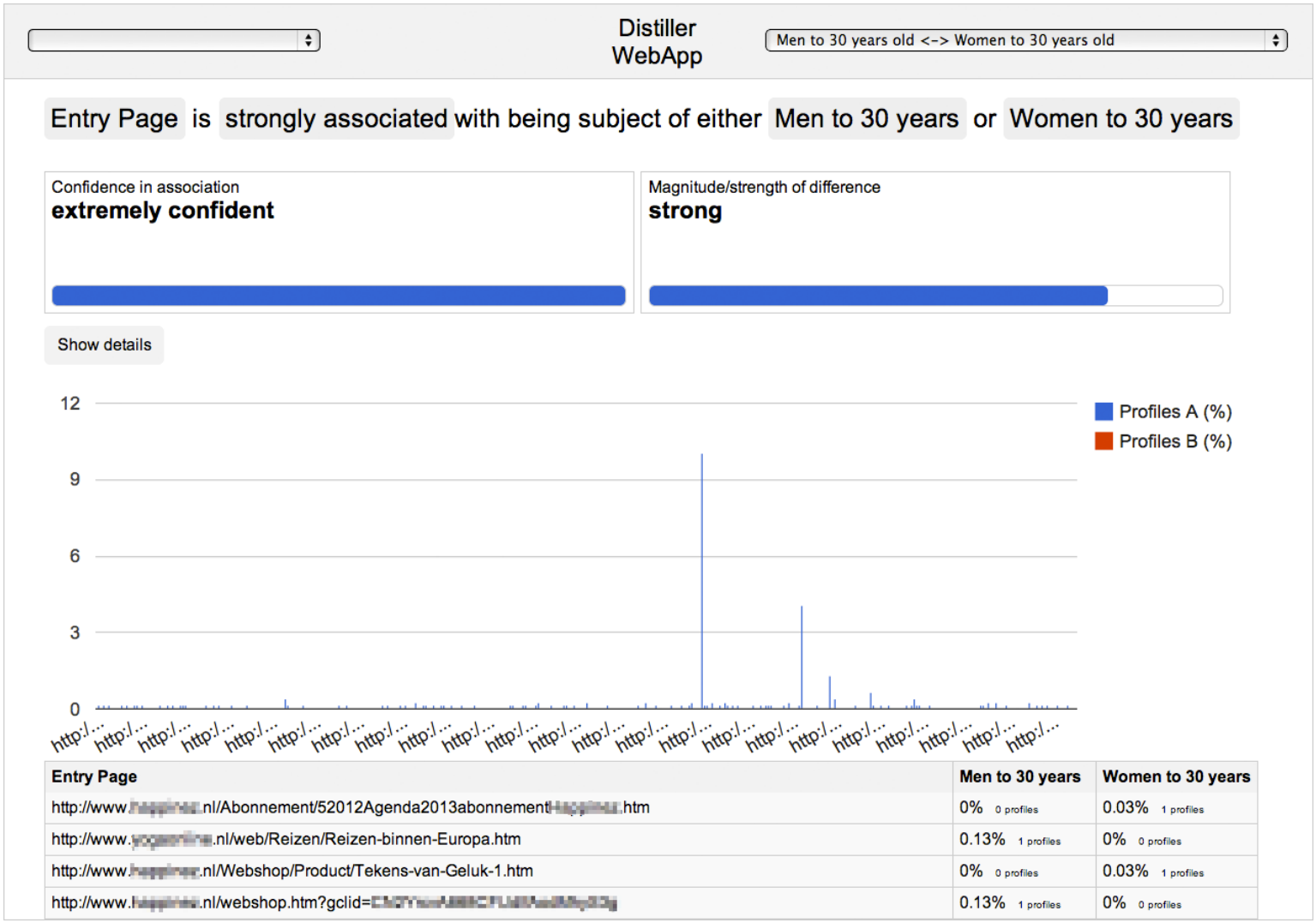


Figure 6.7: A screenshot of the Distiller WebApp for a categorical attribute with many categories. The table for which four rows are shown is 1455 rows long.

effect) and less valuable results from one another. In the football club example, twelve of the twenty significant properties were classified as trivial or negligible in terms of effect. Consequently, there were just eight properties that are statistically significant and have a weak to strong effect size. The three strongest of those eight distinctive properties measure the same visitor characteristic in a different manner: the football club preference. It is good to see that this actually comes up in the analysis. Furthermore we have found that Ajax supporters have a slightly higher number of visits than Feyenoord supporters. Feyenoord supporters search (slightly) more for the keyword "Feyenoord" than Ajax supporters (4.99% to 1.23%).

Finally, what we did not evaluate is whether there are more interesting differences within our data that do not come up in our application. As we have discussed, at the moment, the prototype is not capable of a good analysis of multi-valued attributes and attributes with many categories. Therefore we are pretty sure there are interesting differences in the data that do not come up in the current version of the application. It may be interesting to examine whether the prototype finds all distinctive properties that a human expert can find within the data. What remains to say is that the results always strongly depend on the quality of the data. If there are few significant differences within the data, then little significant properties will come up in the application.

Chapter 7

Conclusion

We have started this research with a rather practical issue. The goal of this study was to automatically find distinctive features in two independent segments of progressive profiles. We have thoroughly investigated the research problem, formulated a model for progressive profiling, and ultimately we have reflected the results on the practical problem in chapter 6.

First, in chapter 2 we have carefully examined the characteristics of the available data. The high degree of uncertainty about profile values has led us to a statistical approach. We consequently have build a solid foundation of statistics in chapter 3 and chapter 4. In chapter 5 we have incorporated these statistics in a formal model for progressive profiling. This model addresses both the significance and the magnitude of effect of segment differences. We have decided to indicate the significance as the "confidence in the difference", as we find this term better reflects what we actually try to assess. The interpretation of the statistics is an important part of the model. We have provided general guidelines for the interpretation of the outcomes of the statistic. Finally we have applied the model. The result is a working prototype that operates on actual datasets. This application has exhibited a few weaknesses of the model. We have discussed these weaknesses and outlined possible solutions or improvements to the model.

Overall, we think the results are promising. Distiller can be used without technical knowledge about statistics. The results of the analysis are presented to the marketeer in a clear and understandable manner. It saves marketeers a lot of time, as it makes clear at a glance what the distinctive characteristics are between two groups of profiles. These are valuable, and sometimes unexpected, insights that may cause to approach certain groups of customers in a different manner. After some further research we think the model can be successfully applied within BlueConic. We will discuss our recommendations for GX Software in the next section. Finally in section 7.2 we conclude with some subjects for future research.

7.1 Recommendations

The findings of this thesis lead to the following recommendations for GX Software

- Due to time and resource constraints the prototype has not been programmed in a very efficient way (section 6.2). We however expect that when the application is programmed in a more careful way, and is built on top of powerful search systems such as Apache Solr or ElasticSearch, it may generate the analysis within milliseconds (in real time).

- For each profile, information is stored on several profile properties. Currently, little is specified about these properties. This increases flexibility but allows for strange data on some properties. For example: we have seen negative, or extremely large values on the property "age". On dichotomous attributes we have seen more than two categories, such as 'yes', 'no', 'false', 'null'. Automatic data analysis strongly depends on the quality of the data.

  We think it is advisable to specify some characteristics (range or possible categories) about the properties to keep the dataset clean. Furthermore, as discussed in 6.4, we advice to specify the meaning of emptiness of a property in the definition of profile properties.

- The prototype has demonstrated some weaknesses in our model. We have highlighted these findings in 6.4. Before applying the model in BlueConic we therefore strongly recommend to look at the subjects for future research, which we will address in section 7.2.

- Finally, the prototype should be incorporated in the existing BlueConic software product. One should take care finding the appropriate form and position within the application. In its current form the most likely decision seems to create a separate analysis section within the application.

7.2 Future research

We will conclude this thesis with some subjects for future research. These subjects largely reflect the findings in section 6.4.

- In the model we assume that data on numeric attributes follows a normal distribution. We have found that many numeric attributes that measure visitor behavior follow a Poisson distribution instead of a normal distribution. Subject of future research should be to adapt the model with Poisson models for the analysis of count attributes.

- The model does not consider multi-valued attributes. Alternatively, we have suggested a method to translate multi-valued attributes into multiple single valued attributes in section 6.4. Future research could focus on how to accurately assess confidence and effect of differences on two segments for multi-valued attributes.

- The model does not produce reliable outcomes on categorical attributes with many categories, where most of these categories have a small number of instances. We have suggested some solutions to reduce the number of categories in section 6.4. Future research on these attributes may potentially lead to new insights.

- Based on a variety of literature we have proposed guidelines for the interpretation of the used statistics. Especially on the aspect of effect it is important that these guidelines are consistent with the interpretation of marketeers. The guidelines may be, for example, too strong within a marketing context. We have already discussed an interpretation issue in section 6.4 that encourages further research. This may be investigated in a practical study that considers how marketeers value the statistical results, and relates these assessments to the guidelines discussed in this thesis.

- As discussed in section 6.3.2 the results are ranked according to the confidence score. However, with this ranking, results may appear on top of the ranked list that we are very confident of, but are meaningless in practice (have a negligible effect). Future research may address finding a single relevance score that combines both confidence and effect.

- Throughout this thesis we have extensively discussed that visitor profiles are built up progressively. On some profiles much data has been gathered over time, while on others we come to know very little. In practice we say that some profiles contain much information, and others contain very little information. This makes sense, because profiles that are filled on many attributes provide much information about that website visitor. However, not all attributes are equally informative. One might argue that rarely filled attributes are more informative than frequently filled attributes. What "information" exactly means is a bit vague. In addition, it is difficult to express how informative a certain profile is. In order to solve these issues we have made a step towards defining **profile informativeness**. Although we think it this is a very interesting topic and gives rise to further research, it was partly outside the scope of this thesis. Therefore we have decided to discuss our measure for profile informativeness, which is inspired by the tf-idf statistic, in Appendix A.

Bibliography

[1] M.J. Berry and G.S. Linoff. *Data Mining Techniques:*. Wiley, 2004. 28

[2] Ronald P Carver. The case against statistical significance testing. *Harvard Educational Review*, 48:378–399, 1978. 16

[3] J. Cohen. Quantitative methods in psychology: A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. 30

[4] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge Academic, 2 edition, January 1988. 29

[5] L.J. Cronbach and R.E. Snow. *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington New York, 1977. 16

[6] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1029. Morgan Kaufmann, 1993. 10

[7] Mary Lou VH Greenfield, John E Kuhn, and Edward M Wojtys. A statistics primer confidence intervals. *The American Journal of Sports Medicine*, 26(1):145–149, 1998. 39

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. 45

[9] Roger E. Kirk. Practical Significance: A Concept Whose Time Has Come. *Educational and Psychological Measurement*, 56(5):746–759, October 1996. 28

[10] Erin Leahey. Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology. *Social Forces*, 84:1–24, 2005. 42

[11] Timothy R. Levine, Ren Weber, Craig Hullett, Hee Sun Park, and Lisa L. Massi Lindsey. A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research*, 34:171–187, 2008. 17

[12] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. Discretization: An enabling technique. *Data Min. Knowl. Discov.*, 6(4):393–423, October 2002. 9, 11

[13] Jane E. Miller and Ph. D. Interpreting the substantive significance of multivariable regression coefficients. 40, 42

[14] D. Muijs. *Doing Quantitative Research in Education: With SPSS*. SAGE Publications, 2004. 30

[15] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:2004, 2004. 70

[16] A.G. Sawyer and J.P. Peter. *The Insignificance of Statistical Significance Testing in Marketing Theory Testing Research*. Working paper series (Ohio State University. College of Administrative Science). College of Administrative Science, Ohio State University, 1982. 28

[17] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948. 10

[18] S S Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946. 8

[19] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas Lento. Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the Third International Conference on Weblogs and Social Media*, San Jose, CA, May 2009. AAAI Press, AAAI Press. 60

[20] Barbara G. Tabachnick and Linda S. Fidell. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, 5 edition, March 2006. 66

[21] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. 66

[22] Merle W. Tate and Leon A. Hyer. Inaccuracy of the X 2 Test of Goodness of Fit When Expected Frequencies Are Small. *Journal of the American Statistical Association*, 68(344):836–841, December 1973. 61

[23] Bruce Thompson. The "significance" crisis in psychology and education. *The Journal of Socio-Economics*, 33(5):607–613, November 2004. 28

[24] T.C. Urdan. *Statistics in Plain English, Third Edition*. Taylor & Francis, 2010. 66

[25] Klaus P. Wiedmann, Holger Buxel, and Gianfranco Walsh. Customer profiling in e-commerce: Methodological aspects and challenges. *Journal of Database Marketing*, 9(2):170–184, 2002. 3

[26] D.S. Yates, D.S. Moore, and G.P. McCabe. *The Practice of Statistics: TI-83 Graphing Calculator Enhanced*. W.H. Freeman, 1999. 61

Appendix A

Profile informativeness

A.1 Information and surprise

We have described a visitor profile by the values that it takes on a finite set of attributes. Each time we assign a value, content, to a profile property for a profile, we add a some information to that profile. We state that each profile holds a certain amount of **information**. Our aim is to provide a quantitative measure that describes the total amount of information that a profile contains. More formally, if $u \in U$, we want to define a function $I_u(u)$ that assigns a quantitative value for the amount of information included in that visitor profile.

We will first clarify what we exactly mean by information. In essence, a profile is made up of multiple pieces of data. Which can be representations of events, actions, preferences or characteristics. Each piece represents some information. The information within a profile will increase each time a piece of data is added. The key part here is that we do not find all pieces of data within a profile equally *important* and *informative*. Based on prior knowledge we always have some idea of what observations we expect for each profile attribute. The more unexpected an observation, the more surprised we are. Each observations has its own level of surprise that directly relates to how informative we find that observation.

Example A.1.1. We will provide an example. Someone visits our website for the first time and we observe the following events: the visitor enters the website on the homepage, visits the news section, searches for the keyword 'Obama' and provides us its age of 73. Either implicitly or explicitly all of those events contain some information. That one enters our website on the homepage does not really surprise us. For most websites most inbound traffic will be generated by people entering the URL in their browser, although also much traffic is from search engines. Slightly more interesting might be that the fact that this person visited the news section, obviously only a part of the visitors will visit this page. The search keyword provides even more information, it provides some clue about what the visitor is searching for on this website. When the chance that someone searches that keyword is low, this probably really tells something about this visitor and its intension. Finally also the age provides us valuable information. Here, when this age is very uncommon it provides more information than when the age is close to the average visitor age.

What we see in the example is the idea is that each attribute has its own probability distribution, that tells us the probability for a specific value in this distribution. When we observe values that are infrequent we find this more informative than when we observe common values. We will now discuss the related **information theory** concepts of **surprise** and **entropy**.

Definition A.1. Let X be a random variable, taking values in X with a probability density function P, $0 \leq P(X = x) \leq 1$. The **surprise** or **information** I of observing $X = x$ is defined as

$$I(X = x) = log_2 \ \frac{1}{P(X = x)} = -log_2 \ P(X = x) \tag{A.1}$$

The suprise measure ranges from zero for $P(X = x) = 1$ to infinity for $P(X = x) = 0$

$$0 \leq I(X = x) \leq \infty$$

Note that the lower the chance of X taking value x, the more surprised we are to see that value, the higher our outcome for I. Vice versa: if the chance to see x is very high, our surprise will be very low, so will be I.

Example A.1.2. The concept of surprise is a crucial part in well known information retrieval algorithms. Let us consider the very popular $tf - idf$ (*term frequency-inverse document frequency*) statistic [15]. It provides a quantitative measure for the evidence that a document is relevant for a given term.

Obviously the more times we see the given term in that document, the more relevant we would consider the document. The *term frequency* refers to number of times a term occurs in a document. Sometimes a *normalized term frequency* is used to compensate for bias towards longer document. Because there is no linear relation between the relevance and term frequency, logarithmic normalization is usually applied. The *inverse document frequency* idf_t quantifies how surprised we are to see term t in a document. It is a measure of the informativeness of this term about the document identity. The document frequency of t, the number of documents in D that contain t is defined as

$$df(t|D) = |d \in D : t \in d|$$

Now the likelihood or chance for t to occur in any document is defined as

$$P(t|D) = \frac{df(t|D)}{|D|}$$

When we insert this chance in the previously defined equation (equation A.1) of surprise we get the *idf* of t

$$idf(t|D) = log \frac{1}{\frac{df(t|D)}{|D|}} = log \frac{|D|}{df(t|D)}$$

Because usually the word "The" occurs in many documents, and the word "Formula" in only a few, we are obviously more surprised to see the word "Formula" in a document. Consequently, seeing the word "Formula" says much more about the identity of the document, than seeing the word "The". The inverse document frequency provides us with a numeric measure for that. The complete equation is

$$tf - idf(t, d|D) = (1 + log\ tf(t, d)) \cdot log \frac{|D|}{df(t|D)} \tag{A.2}$$

Definition A.2. We have discussed a related measure when discussing the MDLP discretization algorithm. The average surprise on discovering the outcome of a random experiment is given by the Shannon entropy. The **Shannon entropy** defines the amount of information that random variable X contains as

$$H(X) = -\sum_{x \in X} P(X = x)\ log_2\ P(X = x) \tag{A.3}$$

If all observations are equally likely, the entropy is maximized.

A.2 Formalizing profile informativeness

We saw that IDF is a measure of the informativeness of a term about the document identity. The aim is to define a similar measure that provides a quantitative amount directly related to the information that a value of a profile property discloses about the **profile identity**. Again this will be based on the underlying idea that unexpected values, values that surprise us, contain more information than common values. We start by defining a measure that provides us with the total informativeness of a profile.

Definition A.3. We will call that measure the **profile's informativeness** I_u. For each profile $u \in U$ and a set of attributes A

$$I_u(u \in U|A) = \sum_{a \in A} w_a\ I_a(u(a)|\ U, a)$$

The informativeness of a profile is just the sum of the informativeness of its attributes. We may find some attributes more important in affecting informativeness than others, so a different weigth can be assigned to each attribute w_a.

Finally, we have to define the key part of the formula, **attributes's informativeness** I_a. This is a measure of information seeing the sequences of values $u(a)$ for attribute a. The order of values is disregarded. We will provide a possible solution here for categorical attributes. The rest is up to future research.

Definition A.4. Given value $x \in a$ for a profile property $a \in A$, the chance to see any profile with profile property with that value is

$$P(x|U, a) = \frac{|\{z \in U : x \in z(a)\}|}{|U|}$$

The surprise (in bits) seeing the value of x for this property is

$$I(x|U, a) = log_2 \frac{|U|}{|\{z \in U : x \in z(a)\}|} \tag{A.4}$$

If we assume that the occurrence of values within an attribute are statistically independent, we can combine occurrence of multiple values by addition. To

prove this, let x and y be two values

$$\begin{aligned} I(x \text{ and } y) & \\ &= -log\ P(x \text{ and } y) \\ &= -log\ P(x)P(y) \\ &= -(logP(x) + logP(y)) \\ &= I(x) + I(y) \end{aligned}$$

Therefore the **attributes's informativeness** I_a for user u on attribute a can be defined as

$$I_a(u(a)|U,a) = \sum_{x \in u(a)} I(x|U,a) = \sum_{x \in u(a)} log_2 \frac{|U|}{|\{z \in U : x \in z(a)\}|} \tag{A.5}$$