University of Namur

MASTER IN COMPUTER SCIENCE

Gender differences in the design of personal websites

application to an academic context

Miche, Magali

Award date:
2005

Awarding institution:
University of Namur

Link to publication

Institute of Informatics
University of Namur
Namur, Belgium

Gender differences in the design of personal websites: Application to an academic context

Magali Miche

Thesis presented in order to obtain a Master's degree in Computer Science
Academic year 2004-2005

Résumé

A ce jour, de nombreuses recherches ont été menées afin de définir les différences existant entre les hommes et les femmes. Les domaines de recherche ont couvert et couvrent encore la psychologie, le comportement social, les capacités cognitives, etc... Le domaine de l'informatique n'a évidemment pas été oublié. Les recherches sur les différences de genre couvrent des sujets tels que l'utilisation d'Internet, les jeux d'ordinateur, la conception d'environnements d'apprentissage pour les enfants, la réalité virtuelle, la communication en ligne, le commerce électronique, la représentation insuffisante du genre féminin dans les professions liées à l'informatique, etc...

Cependant, fait surprenant, très peu d'études ont été menées sur les différences de genre en matière de web design. Les quelques recherches existantes ont été réalisées par des sociologues se focalisant sur des différences spécifiques, à savoir psychologiques (oser mettre sa photo sur son site et autres). Des caractéristiques plus générales, telles que le nombre de pages, de mots, les polices utilisées et bien d'autres, n'ont pas encore fait l'objet d'une attention particulière. Le but de ce mémoire est donc d'apporter une réponse à la question suivante : existe-il un design masculin et un design féminin?

Pour ce faire, les pages web de plusieurs professeurs masculins et féminins issus d'universités australiennes ont été évaluées sur base d'une série de caractéristiques. Cette analyse quantitative se devait d'être accompagnée d'une analyse qualitative. C'est pourquoi un sondage a été réalisé auprès d'étudiants d'un cours de web design. Ceux-ci ont été interrogés sur leurs préférences en matière de conception de sites. Nous comparerons ainsi nos résultats avec ceux issus de recherches précédentes.

Abstract

Many investigations have been conducted so far in order to determine differences existing between men and women. Research topics have been focusing on psychology, social behavior, cognitive skills and so on. Computer science has not been omitted. Research on gender differences targets topics such as Internet use, computer games, design of learning environments, virtual reality, online communication, e-commerce, gender gap in computer science and so on.

However, surprisingly, very few studies have been conducted on gender differences in web design. The few existing investigations were carried out by sociologists who focussed on specific differences, that is to say psychological differences (daring to put one's picture on one's site and so on). More general features such as the number of pages, words, the fonts etc have not been given any particular attention so far. The aim of this thesis is thus to give an answer to the following question: can we find a male way and a female way of designing?

To carry out the analysis, several male and female professors' web homepages have been assessed on the basis of a couple of features. Those professors have been chosen among different Australian universities. Nevertheless, a qualitative analysis had to be conducted as well. To achieve that goal, a survey has been carried out among students of a web design class. They have been questionned about their preferences in web design. We will then be able to compare our results with those from previous research.

I would like to thank my supervisors, that is to say Mrs.Noirhomme (University of Namur), Mr.Simoff (University of Technology of Sydney), Mr.Valenduc (Fondation Travail-Université) and PhD student E.Cuvelier (University of Namur) for their precious help during the training course and the writing of this thesis. I would also like to thank Jenny Edwards (University of Technology of Sydney) and Fay Sudweeks (University of Western Australia) for their useful advice.

I am also very grateful to Christine Marsigny, my brothers Geo and Fred, Mic, Mr.Tomate and Ravi for their very valuable help. Special thanks to my parents for having encouraged me during my thesis. I can't omit to thank my supervisors and my parents again for having given me the opportunity to spend those five months in Australia. It has been such a wonderful experience in one of the most beautiful countries in the world, one of those I will keep in mind for the rest of my life.

Contents

List of Figures

List of Tables

Table of acronyms

ANOVA ANalysis Of VAriance

BN Binary Variables

CTR ConTRibutions

DDS De Digitale Stad

DISTO DISTance to Origin

F Females

FH Feminine-High

IT Information Technology

M Males

MCA Multiple Correspondence Analysis

MH Masculine-High

MDU Multiple User Dungeons or Dimensions

PCA Principal Components Analysis

PGM Predicted Group Membership

QLT QuaLiTy

SMIR Smirnov

SR Squared Ranks

WMW Wilcoxon-Mann-Whithney

Introduction

Many investigations have been conducted so far in order to determine differences existing between men and women. Research topics have been focusing on psychology, social behavior, cognitive skills etc. Computer science has not been omitted. Research on gender differences targets topics such as Internet use, computer games, design of learning environments, virtual reality, online communication, e-commerce, lack of women in IT professions and so on.

However, surprisingly, very few studies have been conducted on gender differences in web design. A question was then raised: is there a male and female way of designing? In order to give an answer to this question, it turned out to be necessary to have a first look at the previous studies that had been conducted regarding possible gender differences in web design. The first chapter of this thesis is therefore devoted to an overview of the very few but yet interesting papers regarding the topic. The first study will give you explanations about the gender gap in computer science. The second highlights gender differences in response to colours. Regarding the other studies, these were mainly carried out by sociologists focussing on special features such as psychological features (daring to put one's picture on one's web site etc). Those investigations form the basis of the thesis, having allowed me to go one step further in this research project. The second chapter will be devoted to the assessment and the analysis of fifteen male web homepages and fifteen female web homepages. These male and female web homepages have been thoroughly selected among different Australian professors' websites.

In the third chapter, you will be introduced to the qualitative analysis that has been carried out among the students of a Virtual Community class. Those students had to learn how to design a website in order to achieve the design of their virtual communities. Therefore, they turned out to be the perfect target for a survey. They have been questionned about their preferences in web design. The questions they had to answer were divided into two categories: the first one questioning them about their design preferences when they navigate on the Internet, the second one which aim is to figure out what kind of designers they are. Finally, the last chapter will consist of the conclusion, drawn according to the results of the previous chapters.

Chapter 1

A literature review

In this chapter, you will read about the very few research that has been conducted regarding the gender issue in computer science. First of all, you will be introduced to a research carried out in order to highlight some explanations regarding the lack of women in IT professions. Then an investigation about the relationship between gender and colours will be presented. The remaining studies focus on web and interface design in general.

1.1 Gender gap in computer science: cultural and psychological factors

This study has been conducted in the context of the Fifth Framework Programme of the European Commission [WBT02].

1.1.1 Theoretical approaches

Developing adequate concepts for capturing whatever cultural and psychological factors may influence the gendering of computing is not an easy task, and the involved disciplines have taken different routes in how they approach this task. Much of the current discourse around the gender gap in computing is grounded in the debate on women in science and technology that dates back to the early 80's. This debate revolved around some of the fundamental theoretical difficulties of addressing gender issues. One of these difficulties is to do with a dualistic notion of the world. Researchers often seem to use the same categorisations as popular accounts of gender, describing “femininity” in terms of values that conform to stereotypes of women: ambiguity, relativity, intuition, connectedness. These stereotypes do not only entrap researchers in normative and culturally conformist images of femininity, they make it difficult to conceptualise variety and otherness in women. It was argued that people with the bodies of women do not necessarily have the minds of women. Much of this older debate has been absorbed by the notion of gender as “performed” [But93]. In her analysis of drag performances (with the film *Paris Is Burning* as her material), Butler claimed that “all gender is drag” and that any “realness” of a performance is “the result of an embodiment of norms, a reiteration of norms, an impersonation of a racial and class norm, a norm that is at once a figure, a figure of a body, which is no particular body, but a morphological ideal that remains the standard which regulates the performance, but which no performance fully approximates". It is this difference between the performance and the ideal that allows women to express different degrees of sameness and otherness and to experiment with varying styles of performing in different contexts and environments.

The notion of gender as performance has been taken up by many feminist scholars and, interestingly, extended to technology. In [Cro99] it is argued that: "We gender a technology by painting it blue and handing it to a boy. We gender the boy in this interaction, providing a frame of reference for appropriate technological and masculine identity associations and expectations. We of course provide a frame of reference for the girl to whom we do not hand it. We perform with technologies. The technology, with its scripts, schemes, and codes, also performs us in that we become subject to its affordances or by happenstance when we start the performance". This argumentation leads to a second difficulty of theorising about technology and gender. Assumptions about the technologies that are examined in studies of gender differences are often quite general and superficial. Not only are computers different from other technologies in ways that may affect the ways women and men interact with them. The range of computing applications dramatically expanded during the last decade and with it the range of computing professions. At the same time the nature of programming as well as practices of systems design changed considerably. Although computers cannot be considered as a uniform study object, few studies of gender differences take account of the great variety of instantiations of the technology. Looking at gender as performance and performed, and at computers as highly specialised and varied technologies, has consequences for the method of "measuring" gender differences. In [Kay92] it has been argued that "to fully understand whether gender differences exist in human-computer interaction, a qualitative, contextual, developmental approach should be employed to examine specific tasks. Kay stresses that without this comprehensive understanding, researchers will continue to identify only pieces of a very complex puzzle" (quoted in [MLM01]).

1.1.2 Analyses of computing culture

The computer as "masculine"

There is a body of literature that argues that the culture of technology is masculine. Part of the disjuncture between women and technology is to do with the idea of masculine tinkering. Men repair, design and build technologies, while women use some of them without taking an active role in their design. In their study of undergraduate computer science majors, Margolis & al. draw a lively picture of gendered attitudes [MFM00]. Women take pleasure in programming and express delight in the sense of mastering or figuring out a program. They tend to contextualise their interest in computers in other areas such as medicine or the arts, valueing the contributions that computers can make to these fields. For men the "love of computing" comes early, and becomes part of their identity and the stories they tell about themselves. They describe a magnetic attraction between themselves and the computer. The computer is the ultimate toy. This affects their women colleagues who tend to feel intimidated by this controlling and mastering attitude. In [Cro99] it is argued that women, although they often don't "have the same access" to the traditional world of tinkering (car repair, woodworking, electronics construction and repair), do in fact tinker. However, it is labelled arts and crafts, cooking, or gardening, which is "feminised". Croissant builds her arguments on the distinction between working-with and working-on a technology [Coc88]: "To fix a car is performed as working on a technology, while to fix dinner is performed as working with a series of artefacts. Both, however, are activities which involve interacting with the known properties and relationships of complex artefacts". The working-on activities are the ones that are most likely perceived as men's domain. Applied to computing this distinction seems problematic, with the boundaries between design and use becoming increasingly blurred. The term "design in use" captures this phenomenon. It denotes the complex set of practices of "interpretation, appropriation, assembly, tailoring and further development of computer support in what is normally regarded as deployment or use" [DEH02]. As responsibility for a specific piece of software may gradually shift "from systems developers and consultants to local technicians, to web designers, service providers and citizens/users", the question becomes: "Who is designing what for whom?". A related line of arguments refers to technologies as objects which social construction is based on a masculine world-view. The notion of science and technology as being based on masculine world-views has been carried into computer science with a view on gender biases in the design of software, computer games and websites.

The concept of ambivalence

German feminist scholars have enriched the debate on "doing gender" and on the masculinity of computers by introducing the concept of ambivalence as characterising women's relationships to technology. Much of this debate has been carried out with regard to computer science and is based on an explicit recognition of the great diversity of computing professions and, hence, the differences of design practice. Women computer scientists, while describing their own relationship to computers as practical-pragmatic, often characterise their male colleagues as practising an identifying, dominant/controlling, and universalising approach. This is what is meant by women's ambivalence: a combination of pleasure and pride in computing with distance from men defining their life around it.

1.1.3 Gender differences from the perspective of psychology

"It has become a cliché to point out that there are gender differences in attitudes towards computers, knowledge about computers, and computer use, and that these differences appear at an early age" [NC97]. Boys/men tend to have more experience with computers. Moreover, their experience seems to be more self-initiated and of the "tinkering" kind: "Computer labs are commonly overwhelmed with the male presence just as male members of the family dominate home machines" [iDGCoc]. Psychological studies focus on the (behavioural) characteristics and orientations of individuals. With respect to computing, they work with concepts like confidence, anxiety, over-use, but also learning style or interactional style. Gender differences with respect to computing orientations have been identified early on, but they are increasingly called into question and much of the earlier work is being refined.

Attribution

"The assumption behind most research seems to be that females' reluctance to use computers stems from the association of computers with math and science and from the general perception of computers as a male domain" [NC97]. Todman and Dick reported that the only gender difference in computing attitudes resides in how much "fun" computers tend to be ([TD93], quoted in [Bro98]). In their research with 127 fifth graders, 58 females and 69 males, Nelson and Cooper have looked into boys' and girls' patterns of attribution for experiences of success and failure with computers, based on Weiner's taxonomy of attributions [NC97]. One of their main findings is that while boys were indeed more master-oriented and self-enhancing, even in the case of failure, girls did not conform to the image of helplessness and self-derogating behaviour. In addition, girls did not show any special tendency to attribute failure to lack of ability. However, boys felt in general more confident than girls about their performance and more relaxed than those when using the computer program. This echoes empirical evidence about women's lower self-efficacy, which is partly to do with the fact that girls often have less prior experience with computers than boys. However, in [iDGCoc] it is stated: "Unfortunately, for every study supporting the hypothesis that increased experience promotes computing confidence, another study refutes it".

Self-efficacy

Self-confidence or self-efficacy as defined by Bandura is an important theme in psychological research: "Broadly speaking computer self-efficacy can be seen as a measure of an individual's judgement of their own abilities with computers, an assessment of self-confidence" [DHL00]. Several studies show that women tend to enter computer programs with fascination and excitement and leave them without a sense of competency and extreme low levels of self-esteem.

Examining positive orientations

Charlton refers to gender-oriented cognitive schema such as the common assumption that women place more value upon interpersonal communications and relationships than men who are more instrumental (goal-oriented) in their behaviour [Cha99]. He argues that: "The extent to which an individual utilizes such a gender-oriented cognitive schema during their cognitive development determines the extent to which they become sex-typed". It is important to note this is different from the notion of "doing gender" which stresses the situatedness and context-dependence of gender performances. One of the findings of his study is that "sex differences in positivity of orientations toward computers are diminishing as a result of the large increase in the number of applications for which computers are now used, this leading to severance of the previously strong psychological link between computers and stereotypically masculine activities such as mathematics and game playing". Expressiveness, as part of the concept of femininity, was found to be a positive attribute with respect to the development of computing orientations. In [MLM01], Mitra & al take up the idea that women and men may differ in their computing orientations in terms of the type of usage is raised. They constructed their longitudinal study of liberal arts college students around the concept of innovation adoption. The students were asked to respond to a questionnaire that assessed their attitude toward an institution-wide computerisation of the teaching/learning process. Innovation adoption was operationalised in terms of relative advantage, compatibility, triability, and observability. Their finding is that gender plays a role in the ways the students reacted to the implementation of computers, with men reporting that "technological enrichment has a relative advantage in enhancing the learning process, although this difference in perception of relative advantage seems to disappear as the innovation gets entrenched within the university environment. At the same time, women continue to feel that the technological enrichment was not particularly necessary within the environment of a liberal arts institution and thus reported lower perception of compatibility of the computerization effort with the overall culture of the university". The authors conclude that women tend to be more cautious in their interpretation of technological enrichment than men. It is often argued that women have a more pragmatic approach to computers, need a clear purpose for working with the machine and often contextualise their interest in computers in other arenas [MFM00].

Attachment begins at an early age

The home evidently has some influence of gendered attitudes toward computing. In [iDGCoc] a series of studies of young children and computing are quoted. The "family" computer is typically set up in the son's bedroom, which makes it difficult for daughters to get access. In kindergarten, schoolgirls are taught to be docile and polite, while boys may be undisciplined to a certain extent. Boys therefore tend to monopolise their teachers' attention. Also "while boys tend to "jump in" and explore computers without explicit permission during class, girls wait to be told what to do and often ask permission to explore the computer "on their own" [iDGCoc]. Many children are first introduced to computers by a parent. For male students it has been the father who provided them with the first contact with computers before they started exploring them on their own. Fisher & al found: "Females' stories are filled with descriptions of watching their dad work on the computer, or having their older brother show them how he programs the machine [FMM97]. From there, their interest is sparked, and some do become active in computing activities in high school, but their participation is much more qualified than the males". Interestingly, mothers' competence in handling computers and other technical devices at home seems to have a positive influence on their daughters: "The ability to interact with their mother interested in expanding her own technical knowledge will provide young girls with the state of mind and skills needed to succeed in other areas of computer science" ([HRTT99], quoted in [iDGCoc]). Hapnes and Rasmussen reported that "mothers using computers at work frequently urged girls to learn how to operate a computer and get a home PC" [HR00]. Hapnes and Rasmussen also see the purchase of a home computer as related to the parents' identities. For example they found fathers in traditional craftsmen's occupations having no interest in computers at all, while those working in the care sector encouraged their children to use computers. Another interesting finding is that young women who know a male friend who uses a computer are less computer anxious. Hapnes, who studied much younger girls, observed that many of those who did not have a computer at home "had access through school, friends, neighbours, relatives and boyfriends [HR00]. For a great number of the girls, school had been the arena where they had become interested in the Internet, and this was by far the most popular use among the girls"[HR00].

1.1.4 Gendered practices of computer design and use

A central argument in the debate on gender and computing is the gender bias that has been built into the machine, that is to say the programs, applications, and tools. This argument can be found in studies of programming styles, computer games, and the use of the Internet. There is also some literature examining gendered aspects of web design.

Programming styles: gendered approaches

The notion that women may develop specific programming styles goes back to [TP90], an influential article in which they identify two ideal-type programming styles: the style of the bricoleur who works in an associative mode, negotiating and rearranging his/her way through a program; and structured programmers who think in an analytic and rule-oriented way. Approaches to computer programming, they argue, can be arranged on a spectrum characterised by degree of closeness to objects and styles of organising work (in terms of concrete versus abstract thinking). In a later publication, Sherry Turkle identified two dominant computer cultures [Tur95]. The culture of calculation, which she also terms "hard", a top-down structured engineering approach to systems design. She considers simulation as more consistent with a "soft" style with which women seem to feel more comfortable. There are few researchers who take up the issue of gender and programming, and much of the debate on programming styles is carried out in the context of teaching. Carter and Jenkins observed that female students tend to seek help and this indicates a different learning style [CJ99].

In [CJ02], computer science educators were asked to determine whether the authors of particular fragments of code were male or female. Less than one fifth could so, and there was no consensus about the criteria for identifying gendered ways of programming. In [CMK98], mixed teams of girls and boys (10-12 years old) were investigated during a three-month computer project, during which the students simultaneously learned new information and designed a relevant product (a multimedia encyclopaedia) reflecting their knowledge. The focus was not so much on programming as such, but on the status of girls in these mixed-gender teams. From the students' perspective high-status activities were programming and graphic art, Internet research, leading a software demo and consulting (helping others). In the beginning of the project, "low-status" activities such as reporting on group progress and resolving interpersonal problems in the group were assumed by the girls. Boys started contributing on this level only when group meetings were introduced. While most boys worked on individual stations, calling one another for help, girls preferred to work collaboratively, giving programming advice by glancing over to another's screen. The most discouraging finding of [CMK98] was that the girls at the end of the project "had not expanded their planning repertoire to include more bottom-up strategies" and that boys "developed a more flexible view than the girls of what it takes to plan and manage a project".

Computer games

It is widely argued that computer games until recently were exclusively targeted at boys on the one hand, that playing computer games is of limited attraction to girls/women on the other hand. In [iDGCoc], the major findings are summarized as follows: boys' games "encourage competition, shooting, violent graphics, and loud noises, most of which do not appeal to girls. Interestingly, it's the repetition of the music and game activities that girls do not like. Girls tend to prefer games encouraging collaboration with other players and involving storylines and character development" ([GK99], quoted in [iDGCoc]). Few women are incorporated into video games (although there is now an increasing number of games with women heroes on the market).

An interesting finding in this context is that children's ability to solve a problem strongly depends on the wording and imagery of the problem. A UK study found that when "kings, pirates, and mechanical forms of transportation such as planes and ships were used, the boys outscored the girls. However, when the problem was re-written to include "honeybears", a pony, and balloons, the girls outscored the boys" ([Newtm], quoted in [iDGCoc]). In [iDGCoc], gender biases in how stores present and sell software are put forward: "It is painfully obvious that the software for fun titles target young boys while the educational packages target the female population. This only reinforces the typically female view that computers are tools and not devices used for fun". Some studies identified girls' play patterns. According to these, girls prefer collaboration, enjoy nonclosure and exploration, prefer to use puzzle-solving skills and complex social interaction, like characters that behave like people they know, feel attracted by virtual reality applications (e.g. bungee jumping, shopping, conference calls, travel, talk-show hosting), are fond of magical transformations, like rich texture and good audio quality [GM00]. Barbie Fashion Designer is referred to as one of the first computer games that became popular with girls and gave them an opportunity to see that computers can be entertaining. "Many girl games on the market tend to create games that focus on shopping or putting on make-up, rather than the interactive storytelling or educational games that girls crave. In addition, creating an interactive storytelling game is potentially more technically difficult than a standard shoot'em up game" [iDGCoc]. Designing computer games for girls so that they feel attracted by them is a point of debate: "The reality is that many girls want exactly what society has taught them to desire. The issue is whether to reinforce gender stereotypes in the design of games for girls or not".

Gendered designs

Research examining gendered design of software or applications is scarce, with the exception of computer games. In [PL00], kindergarten children's preferences to varying designs of multimedia interfaces are examined, looking at gender. Passing and Levin distinguish between the display, conversation, navigation, and control aspect of an interface. In addition, they use two concepts from the world of the cinema: mise-en-scene and montage. The children were exposed to interactive computer stories which they enjoyed reading: "Each child can read the story alone and can act in his own way making the reading experience personal, easy, enjoyable and more interesting". They found that as a first priority, boys preferred learning interfaces that deal with navigation and control and girls those dealing with display. Furthermore, "girls emphasised writing, colours, drawings, help and a calm-moderate game; boys on the other hand emphasised control over the computer, sharp moves and movement on the screen". One of the outcomes of this type of research may be how to design learning interfaces that are not biased in favour of boys/men.

In [RvOO99], another design issue has been taken up in the context of an analysis regarding a virtual city gate for the City of Amsterdam (DDS). Interestingly, the founder of this website is a woman with experience in media, the arts, and politics. Rommes & al take a "social shaping view" on the development of DDS and they use the concept of gender scripts

[RvOO99]. This concept is based on work done in [Akr95] on how technology developers imagine and represent future users: "technologies contain scripts: they attribute and delegate specific competences, actions, and responsibilities to their envisioned users. When these scripts reveal a gendered pattern, we call them gender scripts" [RvOO99]. It is known that designers consider their own preferences and skills to be representative of users. DDS was designed by a network into which journalists, graphic designers, and people from non-profit organisations were recruited: "This personal network did attract women to the project, most of whom are still very active in computer networks. Nevertheless, looking at the gendered division of labour within DDS, women were mostly found in creative, assisting and policy-making positions, whereas male hackers dominated the programming tasks". Having both, women and men in the team, did not result in extra attention given to female users. The choice of software together with the idea to offer all the functionalities of the Internet resulted in an interface which is complicated to use, favouring more experienced computer users. The main idea, as expressed by the main (male) programmer was one of discovering through trial and error: "You have to keep things exciting; discovering is important. This has to do with the way in which I discovered the Internet and all its possibilities, you discover more and more, and that is fascinating. So you will have to let people discover things; that is fun".

Rommes & al conclude by referring to [Har86] in which it is proposed to discuss gender on the structural, the symbolic, and the identity level: "At the identity level, the designers were personally interested in politics, fascinated with all the new technical possibilities of computer networks, and endowed with a masculine learning-style. This masculine identity was reflected in their representation of users, and thus in the technology they designed. This is remarkable given that, at a conscious level, the designers of DDS were very idealistic and took great pains to make their design user-friendly for everybody". A study of a professional development website for teachers featuring "virtual classroom visits" reveals a similar pattern [HMS02]. The website was developed in a participatory way. Teachers may submit video clips and there is an asynchronous forum for discussing these. Herring & al refer to relevant former studies showing that "men tend to dominate mixed-sex discussion on academic and professional topics: they post more and longer messages, and are more likely to be assertive, self-promoting and critical of others, whereas women are more likely to post shorter messages, be polite and attenuated, and express support for others". They focus their analysis on one aspect of the site, the video clips, trying to understand if and how this feature encourages or discourages participation in the site. Their main finding is there are significantly fewer female videos (75% of videos being submitted by men and only 25% by women), that female videos are shorter in length, and that female participation in the discussions is lower. Users chose video clips on the basis of gender similarity, with female users responding preferentially to female video clips even if they had to scroll down the page to do so. A discourse analysis of discussions revealed that females used more hedges and expressive language and that they tended to be more critical and offer more advice. Discourse styles varied by member category, but on the whole a more female discourse style dominated. In their interpretation of their findings, Herring & al refer to previous research showing that "females are less likely than males to seek out or welcome public attention" [HMS02]. They also argue that females might feel less confident in claiming to be good teachers. The video clips are expected to be models of "good teaching" and research suggests that younger teachers may tend to emulate them. One of the authors' conclusions is that "the under representation of female teachers in the videos, although not an intentional feature of the design, nonetheless has clear and important consequences for the ability of the site to meet its stated goals of fostering online community". They think that there are technical solutions to this problem of having female videos included in proportion to the number of female teachers and of having them more visible and easier accessible. However, Herring & al are also reasoning that non-intentional gender-biases like the ones they identified may not be easy to tackle. For example, "25-30% female representation in mixed-sex public contexts tend to be perceived by both women and men as gender equality".

An entirely different type of study was carried out in [Sim01]. Simon looked at the impact of culture and gender on websites. He used a sample of 160 female and male students, with and without computing background, from four areas (Asia, Europe, Latin and South America, and North America), exposing them to real-world functioning sites. Measuring the effect of gender was based on a selectivity model, according to which "males often do not engage in comprehensive processing of all available information but instead are selective. Males tend to employ various heuristic devices that serve as surrogates for more detailed processing. In contrast, females tend to use a "comprehensive strategy" and attempt to assimilate all available cues". Culture was measured by using Hofstede's dimensions: power distance, uncertainty avoidance, masculinity-femininity, individualism-collectivism, and long-term time orientation. Simon reported clear differences in women's and men's perception of the websites: "Forty-seven percent of female respondents (full sample) indicated that they would have preferred all sites to provide additional information as compared to a much smaller percentage (17%) of male respondents. Additionally, females indicate that they overwhelmingly (84%) prefer sites that are less cluttered, with minimal use of graphics and sites, which avoid multiple levels of sub-pages to drill through. Females (52%) suggested that sites making use of pull-down menus are easier to navigate than those with levels that require them to click through to achieve their objective. Males, on the other hand, indicate that sites making extensive use of graphics and animated objects are clearly their preference (77%)". The limitations of approaches to measuring cultural effects in survey research make an interpretation of differences in computing orientations difficult.

In [Gef00] and in [GS97], it is argued that gender-related social expectations have roots in national cultures. Gefen and Straub found that both cross-cultural difference and gender influence how people perceive social presence, usefulness and ease of use (quoted in [Ahu02]). Greenhill & al argued that cultured notions of femininity, together with philosophies favoured by some Asian cultures that stress collectivism, uncertainty avoidance and long-term orientation, may help formulate new meanings of computing [GvHNP97]. Simon points out some design issues resulting from his study, namely the importance of creating culturally and consumer specific sites, e.g. by tailoring the site based on past behaviour and inference from like-minded people [Sim01]. The main value of such studies for understanding the gender gap seems to be their uncovering of hidden biases in the ways computer applications or websites have been designed. Designers' representations of users are indeed often quite one-sided and unexamined, with the (young, white) male often being projected as the model user.

Internet use

The gender gap in Internet use is an issue that is attracting increasing attention. US surveys indicate a closing of the ratio of males and females using the Internet from 20:1 to 2:1 in a four-year period. Little systematic research has explored the reasons for these gender differences. Schumacher studied a group of incoming undergraduate college students, first in 1989/90 and then in 1997, trying to understand the relationship between computer experiences and Internet use [SMM01]. In this period home use of computers increased from 50 to 87%. Although gender differences had diminished, males in 1997 continued to report more experience in programming and playing computer games, and more of them had their own computer. They also had spent more time online than females. Morahan gives an overview of some of the more substantial findings concerning gender and the Internet [MM98b]. She reported that females find it more difficult to find information online than do men. Men seem to use more Internet applications or channels and they use it for finding economic information, for doing research, and for participating in newsgroups. They are also more likely to use Internet games generally referred to as MUD being intensely absorbing and requiring staying online for an extensive period of time. Many of them are action-ridden games oriented towards gaining mastery and status.

Much has been written about the different communication styles of women and men and "most of the published work on issues surrounding gender and networks emphasizes areas of tension or exclusion" [KF94]. In [Her96], men are portrayed as practicing an adversarial style (which in its extreme may lead to flaming), while women valued harmonious interpersonal interaction. Also, men tend to monopolise online conversations. One area where men do not dominate is the use of email. Indeed, women are often portrayed as using the Internet more for communication with friends and family. Kaplan & Farrell have conducted an ethnographic study of a small community of young women using the Internet [KF94]. In their analysis they concentrate on two focal cases: a young woman who intends to study science and one that shows no interest in programming. The women have some interests that they share in their board meetings, among them science fiction and fantasy texts and music. Kaplan & Farrell describe these young women as being focused on communicating over the Net and on describing their experiences and emotions. These young women's dialogues are directed towards maintaining connections and forging relationships. The boards constitute an important social space. They use it for supplementing their almost daily contact with each other. Kaplan & Farrell contend it is important to learn more about what makes the Internet attractive to young women. Although these women do not feel compelled to learn programming, "the Webs they weave, after all, consist not just of the warp and woof of their electronic messages but of the totality of their lived experiences, combining virtual and material worlds". They also develop a sense of mastery and participation".

1.2 The meaning of color for gender

In [Kho], a review of a few studies regarding gender differences about colours is proposed. What we feel and interact with is in color, including both natural and built environments. About 80% of the information which we assimilate through the sense, is visual. However, color does more than just giving us objective information about our world; it affects how we feel. The presence of color becomes more important in interior environments, since most people spend more time inside than outside. Is there a gender difference in response to color? Although findings are ambiguous, many investigations have indicated that there are differences between gender regarding preferences for colors. Early investigations done by Guilford on the harmony of color combinations highlighted that a person is likely to see balance in colors that are closely related or the opposite [Gui34]. Guilford also found some evidence that more pleasing results were obtained from either very small or very large differences in hue rather than medium differences, this tendency being more frequent for women than men. A review of color studies done by Eysenck in early 1940's notes the following results to the relationship between gender and color [Eys41]. Dorcus (1926) found yellow had a higher affective value for the men than women and St.George (1938) maintained that blue for men stands out far more than for women. An even earlier study by Jastrow (1897) found men preferred blue to red and women red to blue.

Guilford and Smith found men were generally more tolerant toward achromatic[1] than women [GS59]. Thus, Guilford and Smith proposed that women might be more color-conscious and their color tastes more flexible and diverse. Likewise, McInnis and Shearer found that blue-green was preferred among women, and that women also preferred tints to shades [MS64]. They also found 56% of men and 76% of women preferred cool colors, and 51% men and 45% women chose bright colors. In a similar study, Plater found men had a tendency to prefer stronger chromas than women [Pla67]. Kuller conducted a study on the effects of color in two opposite environments [Kul76]. Six men and six women were asked to stay in two rooms, one room was colorful and complex while the other was gray and sterile. Electroencephalograms and pulse rates were recorded throughout the period, as well as the individuals' subjective emotional feelings. The results showed heart rates were faster in the gray room than in the colorful room. Moreover, men tended to have stress reactions more often than women. Men also became more bored than did the women in the gray room. Kuller also stated that men could not achieve the same degree of mental relaxation as women.

Thomas, Curtis, and Bolton interviewed seventy-two Nepalese and asked them to list the names of all the colors they could think of [TCB78]. There was a significant difference between men and women. Although the women consistently listed more color names than men did, the cultural context of this study must be noted since Nepalese women traditionally wear more colorful clothing than men do. A similar study by Greene examined the color identification and vocabulary skills of college students [Gre95]. They were asked to identify the colors of twenty-one color chips. The results showed that women recognized significantly more sophisticated colors than did the men. Findings also indicated that gender different responses in color identification may be attributed to a difference in the socialization of men and women. Another study examined the appropriateness of colors used on the walls of a simulated domestic interior furnished in one of the following styles: Georgian, Art Nouveau and Modern. Whitfield reported that internal consistency among women is higher than for men [Whi84]. When the study was broadened to include marital status, wives achieved significantly more internal consistency in each of the three styles than did the husbands. More recently, Radeloff has found that women were more likely to have a favorite color [Rad90].

[1] Having no color or hue; without identifiable hue. Most blacks, whites, grays, and browns are achromatic colors

In expressing the preferences for light versus dark colors, there was no significant differences between men and women. However, in expressing the preference for bright and soft colors, there was a difference, with women preferring soft colors and men preferring bright ones.

1.3 Using gender schema theory to examine gender equity in computing: a preliminary study

Agosto conducted a study with eleven teenage girls [Ago04]. These ones had to fill a form at first in order to divide them into two groups: the first one being the **feminine-high group**, thus girls having a feminine profile, and the second one being the **masculine-high group**, thus girls having a masculine profile. After having completed this first step, they were requested to perform the assessment of a couple of sites in a computer lab. The results show that the preferences differ between both groups. Here are the results for these:

DESIGN EVALUATION CRITERIA
(FEMININE-HIGH GROUP)

1. GRAPHIC QUANTITY
   (a) concentration
   (b) diversity
2. GRAPHIC QUALITY
   (a) color preferences
   (b) detail
   (c) art style preferences
3. VISUAL ENGAGEMENT
   (a) movement
   (b) composition/design manipulation
4. MULTIMEDIA QUANTITY
   (a) audio/video concentration
   (b) audio/video diversity
5. MULTIMEDIA QUALITY
   (a) audio/video clarity
   (b) audio/video workability

CONTENT EVALUATION CRITERIA
(MASCULINE-HIGH GROUP)

1. INFORMATION ENGAGEMENT

   (a) subject interest
   (b) interactivity

2. INFORMATION QUANTITY
   (a) depth
   (b) supplementary information
   (c) art style preferences

3. INFORMATION QUALITY
   (a) contextual explanation
   (b) topicality
   (c) clarity
   (d) functionality

4. INFORMATION ACCESSIBILITY
   (a) organization
   (b) ease of use

The pattern codes presented in uppercase letters represent the major evaluation criteria on which the corresponding groups based their site quality judgements. The pattern codes in lowercase letters represent subcategories of each major evaluation criterion. The categories appear in descending order of evaluation significance. For example, for the FH group model, the major criterion pattern code GRAPHIC QUANTITY and its subcategory pattern codes concentration and diversity indicate that the FH participants judged web site quality first and foremost according to their assessments of graphic content quantity. The major gender difference was design versus content. For the FH group, design characteristics (especially the overall visual appearance of a site) were foremost in determining whether or not the participants deemed a site to be of high quality.

For the FH group, GRAPHIC QUANTITY was the most important characteristic considered when evaluating a site. Increased graphic **concentration** (a greater number of individual graphic components per individual web page) was greatly preferred, and FH group participants frequently critiqued sites' graphics for lacking **diversity** or for being too similar in visual appearance to other graphics contained within the same pages or sites. GRAPHIC QUALITY was also of great importance in assessing the sites. Correspondence with participants' **color preferences**, increased levels of **detail** in individual graphics, and correspondence with participants' **art style preferences** were used as graphic quality assessment criteria. VISUAL ENGAGEMENT also proved to be significant for the FH group. Any site design that supported graphic element movement (such as animated gifts placed on a page or text scrolling across a page) was praised, as were designs that enabled composition/design manipulation (enabling users to move objects around the screen space or enabling them to change design elements to suit their own preferences, such as selecting from a number of page wallpaper designs). Increased MULTIMEDIA QUALITY in terms of increased **audio/video concentration** and increased **audio/video diversity**, was also highly preferred. The FH group members agreed that information presented through multiple media formats increased their engagement with the informational content by reducing the monotony of presentation. Similarly, **audio/video clarity** and increased **audio/video workability** were being preferred. That is, the participants critiqued photographs for being unpleasantly fuzzy or grainy and audio and video clips for taking too long to download, for requiring helper applications to run, or not functioning at all.

For the MH group, INFORMATION ENGAGEMENT was the most significant evaluation criterion. Increased subject interest (correspondence with topics of personal interest) and increased site interactivity (the ability to input information into a site) were the major factors in determining the degree of information engagement. INFORMATION QUANTITY, measured in terms of the depth of information presented and the extent to which supplemental information was provided (either in the form of separate sections of a site or in the form of links to other sites), played a key role in determining site quality assessments. Although not as important to the participants, INFORMATION QUALITY did play a role in site assessments, on the basis of the provision of contextual explanation, the degree of **topicality** of the task at hand, the perceived **clarity** of the information, and the perceived **functionality** of a site as a whole, independent of the topicality of the task at hand. Lastly, INFORMATION ACCESSIBILITY played a role in site evaluation, on the basis of the utility of a site's organization and its overall ease of use. Information accessibility was largely a function of the participants' desire to use their time on the web efficiently. They wanted to be able to grasp immediately how it was organized and how best to use it. The FH group, on the other hand, was more open to serendipitous surfing within a site.

1.4 Gender and web homepages

In 1995, Arnold and Miller found out that most personal homepages echoed well-known print forms of self-presentation like the pen-pal letter, the CV, the high school year book, that few pages were authored by women, and few women had pictures of themselves on their pages [AM99a]. By 1998, the proportion of women who had personal homepages had increased enormously (though they were still outnumbered by men's pages by about two to one) so that they were able to start an analysis of gender differences in homepages. They concentrated on the textual content of the pages, and found that men's pages were shorter, that there was more variety in length and self-reference in women's pages, and that women made more reference to the reader and seemed to be showing more awareness to those.

In this study, Arnold and Miller looked at homepages produced by people in institutional or commercial settings. Given that it is often suggested it is particularly in such settings that women find it difficult to have their status, authority and credibility recognised, it seemed worthwhile to see how the "official" personal web pages of women and men might differ in these aspects. They selected some academics' websites and found different styles of presentation. Some are confidently self-effacing (particularly amongst high status men) and others (mainly women) are friendly yet clearly feel obliged to display "credentials" (full CV, list of degrees and honours, etc). Often amongst the women (though they suspect in some cases with heavy irony) there was a "feminine" style of self-image, but they have not found any women's pages that use jokey pictures of themselves, as some men do. Nor were women's presentations as overtly confident as on some of the men's sites.

On the web people can "belong" to a group of people, e.g. in a department or subject grouping, which is dominated by a house style. Yet even here, gender differences intrude in the cyberspace equivalent of "fluffy" feminine (such as the use of a substitute picture e.g. "flowers") compared to technical "images" (e.g. a computer) used by men. The place that is the homepage may have both a "front door" (ie the person may have been found "at home" because of real-life status) or via the "back door" (from a trail of links through a subject research). For the women found via the homepage "front door", they present themselves as open, "friendly" and smiling (with a suitable picture), but also include a full CV or list of honours, degrees, titles or membership of esteemed professional bodies. The men, on the other hand, are able to be confident about the way they present themselves and their work (which they can assume is the reason for the "visit" to their page) and the discovery of credentials will be possible, but is not the most important "presenting" feature on their page.

1.5 A Nomad Faculty: English Studies' Online Representations of work, product and workplace

In [Hes], it was found out that while several male academics also include family within their online representations, men's pages tend to focus more on presenting a self-image to the viewer. Women's pages, in contrast, often feature more pictures of family members than themselves, and in many cases completely exclude their own image. In the survey related to the case study, some women even said they had chosen to represent themselves through graphics instead of showing themselves on the web.

1.6 The Presentation of self in WWW homepages

For the purpose of this study, Miller and Mather looked at 35 women's and 35 men's personal homepages [MM98a]. For the analysis, the length of each page was measured, in half pages of A4 printout (roughly equivalent to screens). Most pages were one, two or three screens (means of 3.1 for women, 2.5 for men). Regarding the style of the pages, two main types of pages were found: "low-content" pages mainly made up of links to other pages/other sites, and "high-content" pages, with more information on it. Pages were classified as low-content, high-content or unclassifiable. There was no gender difference at all here, with 22 low-content, 11 high-content, and 2 unclassifiable pages in each group. A traditionally-identified gender difference has been between "expressive" and "instrumental" orientation [Bem81]. Miller and Mather examined this by looking at what was mentioned and linked to on the page. A more expressive style would focus on feelings, people, and relationships, while the instrumental style might show itself in reference to abilities and achievements, material goods, and organisations and products rather than people. Various measures might relate to this dimension. They counted links to other people, compared with links to non-personal sites. Women did put up more links to other people (mean of 1.8 compared with 1.2 for men), but they also had more links to non-personal sites (12.0 vs 9.4). Women also show more awareness of, and engagement with, the visitor to the site. Women's pages had a mean of 4.5 references to the reader (using words like "you", "yours", or expressions of awareness to the reader), whereas men's pages had 2.6. Guestbooks were more common on women's pages as were counters (21 to 13).

The authors identified four categories for self-image on the page:

- **straight**: an image which is meant to be a straightforward likeness
- **joke**: a distorted or caricatured or unrepresentative image, e.g. cartoon, baby photo, author just after falling off bike into mudhole, author caricatures as frog, etc
- **symbolic**: an image which represents a human being, but not the actual person who posted the page. This is often a piece of clip art, like a cherub or a generic silhouette
- **none**: no images of humans

The authors of this study counted blurred or pixellated photos which *might* be of the author, but were so unclear that they didn't really represent an individual. They were also a bit surprised to find there were several (15 out of 35 for both groups) pages with no images at all. Men's pages had more "real" images (10 compared with 6) as they expected. The big difference was in the other two categories. Joke images only featured on men's pages (on 4), and symbolic images only on women's (on 10 pages, the most common form of image on women's pages).

1.7 Gender differences in visualization

This article is a case study about gender factors associated with visualization tasks on the computer [Khu04]. In [HS04], it is stated that most of the time websites are designed without considering these differences [HS04]. In fact the design is usually biased towards males. Male and female perceive, interprete and understand interfaces differently from each other. When it comes to real life, research has shown that men outperform women in particular spatial tasks. On the other hand research has shown that women outperform men in verbal tasks. One of the theories of gender differences in spatial abilities is the Hunter & Gatherer theory in [SE92] that states that female are better at keeping track of things and objects in their environment and men are better at travelling in unfamiliar environments, which supports the findings of the Passing and Levin study where boys keep navigating the environment until they find their way around while girls ask for help as soon as they get stuck when they are both exposed to an unfamiliar computer environment [PL00].

Chapter 2

The quantitative analysis

In this chapter, you will be able to read about the quantitative analysis that has been carried out, consisting of the assessment of 15 male and 15 female sites. You will be told about the selection of the sites, the list of features that has been established, the hypotheses that have been formulated and of course, the results of the assessment.

2.1 Site selection, features and hypotheses

The first step in the assessment of the sites was of course the selection of these. Academics' web homepages have been therefore selected among Australian IT departments. To try to reduce the influence of the cultural factor, the sample didn't include any Asian professor. According to professional web designers, the Asian culture has different requirements in web design. The main difference when designing for Asians is the need of having very colourful sites. In appendix A, you will find the list of the different sites that have been selected. The list of features consists of thirty characteristics that can be classified into four categories: *design*, *content*, *functionality* and *media use*. A coding scheme has also been written in order to explain the meaning of each item included in the list and the way to assess it. For the first category, that is to say the *design category*, you will find the following features:

- **number of pages**

  In [MM98a], the length of each page was measured, resulting in a higher score for the males than for the females. That is why it was interesting to assess the length of the web pages in A4 printout in order to know if the male academics tend to have more pages on their sites than the female academics. To assess the number of pages, each link included in the website, apart from links to other people's websites and non-personal websites, has been opened with Microsoft Word used as a browser. The *Word Count Tool* has been used to count the number of pages for each link.

- **number of words and characters**

  In [Khu04], it is stated women are better at verbal tasks than males. That is why we will analyse the number of words in order to know if women are more expressive in terms of text use. We will also focus on the number of characters to figure out if women tend to write longer words. The procedure to assess the number of words and characters is the same as the one for the number of pages.

- **number of characters with spaces**

  Here the purpose is to use the number of characters with and without spaces in order to compute a percentage of white spaces. In his study, Simon reports 87% of the females being part of his sample prefer websites that are less cluttered [Sim01]. That is why it seemed of interest to take into account this feature as well as the number of paragraphs in order to confirm or invalidate this finding. The procedure to assess the number of characters with spaces and paragraphs is the same as the one for the number of pages.

- **number of fonts**

  In their study, Julie Fisher and Annemieke Craig studied gender responses to key elements in effective web design, mentioning text design as being a major feature [FC00]. It thus turned to be interesting to have a look at the number of fonts in order to know if females and males had a different use of these.

- **type of fonts**

  The number of fonts couldn't be examined without taking the type of fonts into account. The purpose was of course to determine if women tend to use more "fluffy" feminine fonts (curved etc) on their websites. Regarding men, the tendency would thus be the use of classic fonts. The fonts were therefore classified into two categories: classic and "girlish". The latter consists of all curved fonts.

- **number of colours used for text and hypertext**

  Many gender studies regarding colours state women and men are different on this point of view. Women can list more sophisticated colours as stated in [Kho], prefer websites that are colourful and thus lively. That is the reason why this feature as well as the number of background colours had to be examined in this study. To assess the number of colours for (hyper)text, each link included in the website, apart from links to other people's websites and non-personal websites, has been searched thoroughly. Regarding the number of background colours, each page of the site has been viewed.

- **type of colours used for text and hypertext**

  As said in [Kho], women prefer red and men blue. That is why it seemed important to figure out if we could validate this finding by examining the use of reddish and blueish colours in the frame of our sample. As stated in Fisher and Craig's study, black fonts are considered as "boring" by the males and "a good choice" by the females. Thus this feature was taken into account as well. Regarding white and grey fonts, it was interesting to include them in the list since they seemed to be used more often by the males further to the site selection.

- **number of words for the main page**

  During the selection of the sites, it appeared men tended to put more text on their "front" page than women. That is why this feature has been included in the list. To assess it, the first page of each website being displayed has been opened with Microsoft Word to use the *Word Count Tool*.

- **technological level of the site**

  Here again, this feature was included during the site selection period in order to figure out if there was a difference between genders regarding the sophistication level of the sites. Thus the whole website has been thoroughly searched in order to find any technological feature like a search engine, flash use, sophisticated design etc.

- **type of background colours**

  Two categories were added compared with the type of colours for text: soft and dark colours. In [Kho], it was stated men and women did not differ regarding the preference for soft or dark colours. Can we say the same in our context?

- **type of background**

  Many gender studies highlight the fact men dare more to show jokey pictures or items in general. During the site selection period, it appeared original backgrounds were only found among the males. Thus this feature had to be included in the list and was charaterized into two categories: classic and original.

For the content category, you will find these features:

- **presence of self-description**

  In [AM99a], [Hes] and [MM98a], the self-presentation was mainly discussed. That is why it turned out to be of interest to include the presence of a self-description in order to know if we can find a gender difference on this point. To assess this feature, the whole website has been thoroughly searched in order to find any self-description consisting of a couple of sentences describing the academic.

- **number of words for the self-description**

  Since the presence of a self-description was examined, it appeared to be interesting to have a look at the number of words for this self-description as well in order to highlight a possible gendered difference. To achieve this, each self-description has been pasted into a Word document to use the *Word Count tool.*

- **type of self-description**

  Many gender studies argue men tend to use self-mockery or jokey elements (pictures etc) more often than women. Thus a question was raised in this context: could we find this kind of difference when an academic describes him-/herself on the net? To carry out the assessment, self-descriptions were classified into two categories: professional or private. A professional description only consists of references to the academic's investigations whereas a private self-description consists of personal elements or elements of self-mockery.

- **proportion of personal pages**

  During the site selection period, it seemed men included more pages with personal content than women. So, it seemed interesting to study this feature thoroughly in order to know if this tendency could be confirmed on a statistical point of view. To achieve this, the whole website has been thoroughly searched to count the number of pages containing personal elements. The number of personal pages has been divided by the number of pages for the whole site to obtain a percentage.

- **focus on credentials**

  In [AM99a], it is clearly stated women presented their credentials (full CV, list of degrees and honours etc) as a main feature of their site. That is why the whole website has been thoroughly searched to determine if the academic was focusing on his/her credentials.

- **presence of graphic accents**

  Many studies regarding online communication state women use more emoticons than men. Thus the whole website has been thoroughly searched in order to find any graphic accent.

The functionality category consists of:

- **number of links**

  In [MM98a], no difference was found regarding "low-content" (a low number of links) and "high-content" (a great number of links) sites. This feature was included in the list in order to know if we can state the same in this context. In order to assess it, a Java program has been written and can be found in appendix B. The *Page Info tool* of a Mozilla Internet browser has also been used to see all the links contained in a specific page. The CSS link type has been excluded.

- **number of links to other people's pages**

  In [MM98a], a gendered difference was found with females using a more expressive style regarding links to other websites and males using a more instrumental style. This means women tended to put up more links to other people's pages and men links to non-personal websites. Thus it turned out to be interesting to examine such a feature in our context by visiting each link in order to determine whether it referred to another person's web homepage or to a non-personal site.

Finally, in the media use category, you have:

- **number of self-photos**

  In [Hes], it is stated men's pages tend to focus more on presenting a self-image to the viewer compared with women. That is why it seemed interesting to include this feature in order to compare with the finding of that study. Therefore, the whole website has been thoroughly searched in order to count the number of photos portraying the academic.

- **type of self-photos**

  In [AM99a], it is argued women try to present a suitable picture as much as possible. That is why the first type of self-photos to be taken into account is the "official picture", in order to know if we can find a gender difference in our context as well. Secondly, in [MM98a], it was found men tend to show more joke pictures of themselves. The second type is thus the "non-official picture". Further to the site selection, six other categories were examined as well: family (academic with family), friends (academic with friends), colleagues (academic with colleagues), pets (academic with pets), computer-related (academic alone performing an activity related to his/her job).

- **quality of self-photos**

  The authors of [MM98a] examined the quality of self-photos, finding a gender difference (women having more blurred self-photos). That is the reason why this feature was included in the list, the quality of the self-photos depending on the pixellization level of the pictures.

- **number of photos**

  Since the number of self-photos was included in the list, it turned out to be necessary to study the number of photos as well, since it was said above women had a greater tendency to put up more photos excluding their own image. Thus the whole website has been thoroughly searched in order to determine the number of photos that were not showing the academic.

- **type of photos**

  During the site selection, six categories of photos could be identified: academic's family, academic's friends, academic's colleagues, academic's pets, computer-related (items related to academic's job). They could thus be examined in a statistical way in order to highlight any possible gender difference.

- **quality of photos**

  Since the quality of self-photos was studied, it could be interesting to study the quality of photos as well (depending on the pixellization level of the pictures).

- **number of graphics**

  In [Ago04], the feminine group's first criterion for rating a website was the graphic quantity. It turned out to be interesting to count the number of graphics in our context in order to know if women tended to use more graphics than men. To assess this feature, a Java program you can find in appendix B has been written. The *Page Info tool* of a Mozilla Internet browser has also been used to distinguish graphics from self-photos and photos.

- **type of graphics**

  During the site selection, five types of graphics could be highlighted: basic (no frills graphics), trendy, artistic, comics/cartoons, computer-related. We could thus question a possible gendered difference regarding the type of graphics.

Before the analysis of the results, hypotheses had to be formulated. These hypotheses will have to be confirmed or invalidated by the analysis. For the design category, they consist of:

1. **Men are more expressive than women**

   - Men's websites contain more pages than women's
   - Men's websites contain more words per page than women's
   - Men's websites contain longer words than women's

2. **Text in women's websites is more spaced out than in men's**

   - Women use more white spaces than men
   - Women use more paragraphs per page than men

3. **Women tend to use more different fonts than men**

4. **Men and women both use classic fonts**

5. **Men and women don't use girlish fonts**

6. **Women tend to use more colours for text and hypertext**

7. **Men and women differ in the type of colours used for text and hypertext**

   - Women tend to use more reddish colours
   - Men tend to use more blueish colours
   - Both use black
   - Men tend to use white more often than women
   - Men tend to use grey more often than women

8. **Women tend to have fewer words for the main page**

9. **Men tend to have more technological websites**

10. **Women's websites have a more colourful background**

11. **Women and men differ regarding the type of colours used for their backgrounds**

    - Women use softer colours
    - Men use darker colours
    - Women use more reddish colours
    - Men use more blueish colours
    - Men use black more often
    - Men and women do not differ regarding the use of white
    - Men use grey more often

12. **Men and women differ regarding the type of background**

    - There are more women's websites with a classic background than men's sites
    - There are more men's websites with an original background than women's sites

For the content category, they are formulated as below:

1. **There are more men describing themselves on their websites than women**

2. **Men tend to describe themselves in a longer way than women**

3. **Men and women differ in the way they describe themselves**

- Men tend to describe themselves in a private way
- Women tend to describe themselves in a professional way

4. **The number of male academics having personal content in their website is greater than the number of female academics**

5. **By and large men tend to have websites with a higher personal content**[1]

6. **Men and women do not differ regarding the focus on credentials**

7. **There are more women using graphic accents than men**

Regarding the functionality category, the hypotheses are the following ones:

1. **Men and women do not differ regarding the number of links included in their websites**

2. **Men and women differ in the type of links they have on their websites**
   - There are more men who insert links to non-personal pages than women
   - By and large, men's websites include more links to non-personal pages[2]
   - There are more women who insert links to personal pages
   - By and large, women's websites include more links to personal pages[3]

For the media use category, you have:

1. **There are more men showing self-photos than women**

2. **Men show more self-photos than women**[4]

3. **Women and men differ regarding the type of self-photos**
   - Women and men both show the official picture
   - Men show more non-official pictures of themselves than women
   - Men show more family pictures with themselves than women
   - Men show more pictures of themselves with friends than women
   - Men and women do not differ regarding pictures of themselves with colleagues
   - Men and women do not differ regarding pictures of themselves with their pets
   - Men show more pictures of themselves in their leisure time than women
   - Men show more computer-related pictures with themselves

4. **Men and women both show good-quality self-photos**

[1] The hypothesis formulated here is independent of the previous one. Actually, there could be more women having personal content than men, but overall, if we sum the quantity of personal content for the males, we will obtain a greater percentage than for the females

[2] same remark as above

[3] same remark as above

[4] same remark as above

5. **There are more men showing photos apart from self-photos than women**

6. **Men show more photos apart from self-photos than women**[5]

7. **Women and men differ regarding the type of photos**

    - Men show more pictures of their families than women
    - Men and women do not differ regarding pictures showing their friends
    - Women show more pictures of their colleagues than men
    - Men show more pictures of their pets than women
    - Men show more pictures of their leisure time than women
    - Men show more computer-related pictures than women

8. **Women and men both show good-quality pictures**

9. **The number of females using graphics is greater than the number of males**

10. **Women use more graphics than men**[6]

11. **Men and women differ in the type of graphics**

    - Women use more basic graphics
    - Women use more modern graphics
    - Women use trendier graphics
    - Women use more artistic graphics
    - Men use more comics graphics
    - Women and men do not differ regarding the use of computer-related graphics

2.2 The analysis of the numerical variables

In this section, we will analyze the variables which are not binary, that is to say:

- the number of pages, words per page, characters per word,
- the proportion of white spaces and the number of paragraphs per page,
- the number of fonts and colours for text and hypertext,
- the number of words for the main page,
- the number of colours for the background,
- the number of words for the self-description,
- the ratio of personal pages,

[5] see remark on previous page

[6] see remark on previous page

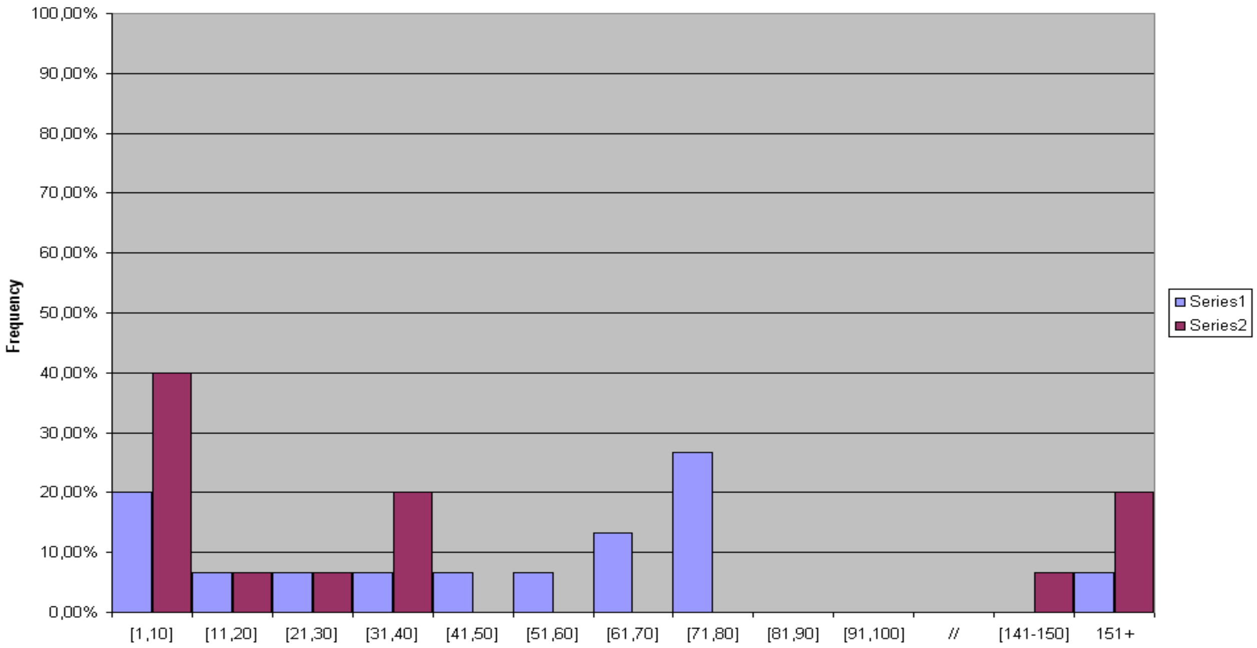


Figure 2.1: Distributions for the number of pages

- the number of links, non-personal and personal links,
- the number of self-photos and photos,
- and lastly the number of graphics.

In order to achieve this, you have to be introduced to the testing methods. According to the shape of the histograms for each non-binary variable, we cannot use statistics based on the normality hypothesis, even after applying variable transformations. Indeed, the distributions look bimodal as you can see in figure 2.1 for the number of pages[7], or can even be totally atypical. That is why we will apply non-parametric statistics we describe in the following subsection. We could have thought of running a $Student's\ t$ statistic on our sample. Indeed, the $t-distribution$ or $Student's\ distribution$ is a probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. It is the basis of the $Student's\ t-tests$ for the statistical significance of the difference between two sample means, and for confidence intervals for the difference between two population means.

Besides $t-test$ is any statistical hypothesis test in which the test statistic has a $Student's\ t-distribution$ if the null hypothesis is true. One of the most frequently used $t-tests$ is a statistical test of the null hypothesis that the means of two normally distributed populations are equal. If the $t\ value$ that is calculated is greater than the threshold chosen for statistical significance (alpha conventionally equal to 0,05), then the null hypothesis that the two groups

[7] Please note that *Series 1* represents the males and *Series 2* the females

do not differ is rejected in favor of the alternative hypothesis. The latter typically states that the groups do differ. But we have a small sample on which we can't apply the *Student's t* statistic because of the non-equality of the variances (equality of variances is required to compare means). After the section devoted to the non-parametric methods, you will find a description of each statistical method that has been chosen to conduct the analysis.

2.2.1 Classic parametric methods when assuming normality

Many classic statistical methods have been established on the normality condition in order to solve problems involving computations of means, variances and standard deviations as well as correlation and regression coefficients. These methods can be qualified as *parametric*, in contrast with *non-parametric methods* or *distribution-free methods.* If we go by the asymptotic normality property of the distribution of the mean sampling, the normality condition of the populations' distributions is not essential, in practice, in the case of confidence intervals and tests for equality of means.

But such a principle cannot be applied all the time. As such, methods related to variances and standard deviations (determination of confidence limits and tests for equality of variances and standard deviations) are definitely more sensitive to the non-normality of the populations, as well as some methods involving means. Each time that the normality condition is essential on a practical point of view, it is necessary to verify if this condition is effectively satisfied and if not, to try to adapt the data in consequence, for example by a variable transformation.

Regarding data collection, the normality condition does not involve the initial observations, but the deviations or residues in relation to a theoretical model. Here again, we have to verify the normality condition on the basis of these deviations or residues, and *not* the initial data, and choose a variable transformation according to these deviations or residues.

2.2.2 Non-parametric methods

Other statistical inference methods are, on the contrary, based on no particular assumption for the populations' distributions (normality, symmetry etc). These methods, called non-parametric or distribution-free methods, can be applied in general for a large variety of distributions. However, we have to notice that, regarding the comparisons between two or more populations, some non-parametric methods assume the compared distributions all belong to the same family. As such, comparison methods between means or medians, or simply between locations of two or more distributions, are based on a rank study. They thus assume the distributions only differ by their locations, and not by their shapes. Such restrictions are actually important limitations to the use of some non-parametric methods. An essential characteristic of non-parametric methods is their simplicity. This one results from the replacement of the observed values by ranks or binary variables. The median is then sometimes preferred to the mean, as a location parameter, and the amplitude is often used instead of the standard deviation or the variance as a dispersion parameter.

However, the replacement of the observed values by ranks or by binary variables involve some information loss. That is why non-parametric methods are generally less effective or less powerful than parametric methods. Thus, non-parametric methods are used when application conditions of other methods are not satisfied, even after having transformed the variables. On the other hand, the use of these methods is recommended when disadvantages due to effectiveness or power loss are offset by advantages of simplicity and computation rapidity. The cost or the time spent to data collection thus have to be reduced enough, by comparison with the computation cost or time.

2.2.3 The Wilcoxon-Mann-Whitney test

The Wilcoxon-Mann-Whitney test is one of the best-known non-parametric statistical significance tests. The test is appropriate to the case of two independent samples of observations that are measured at least at an ordinal level, i.e. we can at least say, of any two observations, which is the greater. The general philosophy of the test is similar to the one of the Wilcoxon signed rank test. To fully understand the Wilcoxon-Mann-Whitney test, you thus have to be introduced to the notion of rank and to the Wilcoxon signed rank test.

1. The notion of rank

In the theory of order relations, the rank of a single observation among a set is its ordinal number when the set is ordered according to some criterion.

2. The Wilcoxon signed rank test

To illustrate this test, suppose that 16 students in an introductory statistics course are presented with a number of questions concerning basic probabilities. In each instance, the question takes the form “What is the probability of such-and-such?”. However, the students are not allowed to perform calculations. Their answers must be immediate, based only on their raw intuitions. They are instructed to frame each answer in terms of a zero to 100 percent rating scale, with 0% corresponding to P=0.0, 27% corresponding to P=0.27, and so forth.

The instructor of the course is particularly interested in the student’s responses to two of the questions, which we will designate as question A and question B. He reasons that if students have developed a good, solid understanding of the basic concepts, they will tend to give higher probability ratings for question A than for question B; whereas, if they were sleeping through that portion of the course, their answers will be mere shots in the dark and there will be no overall tendency one way or the other. The instructor’s hypothesis is of course directional: he expects his students have mastered the concepts well enough to sense, if only intuitively, that the event described in question A has the higher probability. Table 2.1 shows the probability ratings of the 16 subjects for each of the two questions.

The mean difference resulting from this table equals +7.75. The observed results are consistent with the hypothesis. The probability ratings do on average end up higher for question A than for question B. Now we have to determine whether the degree of the observed

Subject	X_A	X_B	$X_A - X_B$
1	78	78	0
2	24	24	0
3	64	62	+2
4	45	48	-3
5	64	68	-4
6	52	56	-4
7	30	25	+5
8	50	44	+6
9	64	56	+8
10	50	40	+10
11	78	68	+10
12	22	36	-14
13	84	68	+16
14	40	20	+20
15	90	58	+32
16	72	32	+40

Table 2.1: Probability ratings for the 16 subjects for each question

difference reflects anything more than some lucky guessing. The Wilcoxon test begins by transforming each instance of $X_A - X_B$ into its absolute value, which is accomplished simply by removing all the positive and negative signs. Thus the entries in column 4 of table 2.2 become those of column 5.

In most applications of the Wilcoxon procedure, the cases in which there is zero difference between X_A and X_B are at this point eliminated from consideration, since they provide no useful information, and the remaining absolute differences are then ranked from lowest to highest, with tied ranks included where appropriate. The result of this step is shown in column 6 of table 2.2. The entries in column 7 of the same table will then give you the clue to why the Wilcoxon procedure is known as the signed-rank test. Here you see the same entries as in column 6, except now we have re-attached to each rank the positive or negative sign that was removed from the $X_A - X_B$ difference in the transition from column 4 to column 5.

The sum of the signed ranks in column 7 is a quantity symbolized as W, which for the present example is equal to 67. Two of the original 16 subjects were removed from consideration because of the zero difference they produced in columns 4 and 5, so our observed value of W is based on a sample of size N=14. The effect of replacing the original measures with ranks brings us to focus only on the ordinal relationships among the measures (“greater than,” “less than,” and “equal to”).

The sum of the N unsigned ranks in column 6 will be equal to:

$$\frac{N(N+1)}{2},$$

that is to say: $\frac{14(14+1)}{2}$ equalling 105. Thus the maximum possible positive value of W (in the case where all signs are positive) is W=+105, and the maximum possible negative value (in

Col.1	**Col.2**	**Col.3**	**Col.4**	**Col.5**	**Col.6**	**Col.7**
Subject	X_A	X_B	original $X_A - X_B$	absolute $X_A - X_B$	rank of $X_A - X_B$	signed rank
1	78	78	0	0	—	—
2	24	24	0	0	—	—
3	64	62	+2	2	1	1
4	45	48	-3	3	2	-2
5	64	68	-4	4	3.5	-3.5
6	52	56	-4	4	3.5	-3.5
7	30	25	+5	5	5	+5
8	50	44	+6	6	6	+6
9	64	56	+8	8	7	+7
10	50	40	+10	10	8.5	+8.5
11	78	68	+10	10	8.5	+8.5
12	22	36	-14	14	10	-10
13	84	68	+16	16	11	+11
14	40	20	+20	20	12	+12
15	90	58	+32	32	13	+13
16	72	32	+40	40	14	+14

Table 2.2: Transformations for applying the Wilcoxon signed rank test

the case where all signs are negative) is W=-105. For the present example, a preponderance of positive signs among the signed ranks would suggest that subjects tend to rate the probability higher for question A than for question B. A preponderance of negative signs would suggest the opposite. The null hypothesis is that there is no tendency in either direction, hence that the numbers of positive and negative signs will be approximately equal. In that event, we would expect the value of W to approximate zero.

3. The Wilcoxon-Mann-Whitney test

As said before, this test has much in common with the Wilcoxon signed rank test. Instead of distinguishing ranks by sign and summing those of identical sign, we consider the observations from both samples as being observations of a single sample. Then we rank these and finally, we sum the ranks associated with one of the two samples (in our case, the males or the females). If both samples come from the same population, we will have a fair mix of low-, medium- or high-ranking observations in each sample. If the alternative to a null hypothesis of identical populations is that our samples come from populations with distributions differing only in location (mean or median), then we can expect lower ranks to dominate in one population and higher ranks in the other. So, in summary, we compute the sum of the ranks and then we compute the following formula:

$$U_m = S_m \text{ - } \frac{1}{2}\ size_1(size_1 + 1)$$

with m representing the first sample, $size_1$ being the size of the sample and S_m being the sum of the ranks of the first sample. We can also choose to compute the formula corresponding to the second sample instead of the one corresponding to the first, that is to say:

$$U_n = S_n \text{ - } \frac{1}{2}\ size_2(size_2 + 1)$$

with n representing the second sample, $size_2$ being the size of the sample and S_n being the sum of the ranks of the second sample. There is no particular reason to compute U_m instead or U_n (or vice versa) since they are linked by the following formula:

$$U_m = size_1 * size_2 - U_n.$$

Regarding the critical values, these can be found in a table. This table contains the results of the computation of $P[U_m = k]$ for the different k's. If U_m or U_n is below the critical value, we **can reject** the null hypothesis stating there is no location difference between the male distribution and the female one.

2.2.4 The Hodges-Lehmann estimator

It isn't very useful to know there is a significant difference between two groups if we don't know how big this difference is. Fortunately, there are techniques for measuring the size of a difference that are insensitive to the shape of the distributions just like non-parametric statistical tests are. One class of such methods are called Hodges-Lehmann estimators.

The Hodges-Lehmann estimator $\hat{\Delta}$, for the difference between two groups provides a good illustration. If one group is the $x_i's$: x_1, x_2, x_3, ... , x_n and the other group is the $y_j's$: y_1, y_2, y_3, ... , y_n, the Hodges-Lehmann estimator for the difference between the x's and y's is determined as follows:

1. Calculate the difference between every possible pair of x's and y's:

   $$d_{i,j} = y_j - x_i.$$

   There will be a total of m times n such differences.

2. Rank the list of these differences in ascending order.

3. Pick the median (value of the variated dividing the total frequency into two halves) from this list:

   $$\hat{\Delta} = \text{median } [y_j - x_i].$$

Now, what does the Hodges-Lehmann estimator exactly do? It is the best unbiased estimator of the median of the distribution of possible differences between the median of x and the median of y. Since the Hodges-Lehmann estimator $\hat{\Delta}$ is the median of a distribution (of $d_{i,j}$'s), you can estimate the confidence limits for $\hat{\Delta}$ similarly to the way you estimate the confidence limits of the median of a set of measurements.

2.2.5 The squared rank test for variance

1. Notations and definitions

In order to fully understand the test, let's first denote the male distribution by X and the female distribution by Y. Their means are thus μ_x and μ_y and their variances are E$[(X-\mu_x)^2]$ and E$[(Y-\mu_y)^2]$. Let's remember that if μ_a= E(A) is the expected value (mean) of the random variable A, then the variance is:

$$var(A) = E[(A - \mu_a)^2].$$

The variance is thus the expected value of the square of the deviation of A from its own mean. In plain language, it can be expressed as “The average of the square of the distance of each data point from the mean”. It is thus the mean squared deviation.

2. Conover’s squared ranks test

The goal of this test is to test equality of variances between both populations, that is to say:

$$\mathrm{E}[(X - \mu_x)^2] =? \ \mathrm{E}[(Y - \mu_y)^2].$$

Conover (1980) proposes a test for equality of variance based on joint ranks of $(x_i - \mu_x)$, $(y_i - \mu_y)$. In practice it is unlikely that the population means will be known so it is reasonable to replace them by sample estimates $\overline{x}$ and $\overline{y}$.

How does this test actually work?

1. Compute $|x_i - \overline{x}|$ and $|y_i - \overline{y}|$. You obtain two sets of measures.

2. Consider these two sets of measures as one and classify all these according to an ascending order.

3. Rank the classified measures from the lowest value to the highest and compute the squares of these ranks.

4. Compute the sum T of the squared ranks corresponding to any of the two samples (i.e. choose the male distribution or the female distribution).

5. Compute the mean of the squares of all the ranks (i.e. the sum of all squared ranks divided by 30, that is to say the number of observations) and denote this mean by $\overline{sqrrank}$.

6. Compute the squares of the absolute deviations and denote them as $sqrdev_i(x)$ for the males and $sqrdev_i(y)$ for the females.

7. Compute S, the estimated standard deviation calculated from

$$S^2 = \frac{n^2\Big\{\sum_i sqrdev_i(x)^2 - (2n)\overline{sqrrank}^2\Big\}}{(2n)(2n-1)}$$

or from

$$S^2 = \frac{n^2\Big\{\sum_i sqrdev_i(y)^2 - (2n)\overline{sqrrank}^2\Big\}}{(2n)(2n-1)}$$

where n is the size of any of the two samples (i.e. 15).

8. Finally, compute Z thanks to the following formula:

$$Z = (T - n\ \overline{sqrrank})/S.$$

If the obtained value for Z is above the critical value, you **can reject** the null hypothesis stating the variances of the two samples are equal.

2.2.6 The Kolmogorov-Smirnov test for a common distribution

The Kolmogorov-Smirnov test is a two-sample test, that tests the H_0 hypothesis according to which two samples (numerical values) originate from the same (undetermined) distribution function F(x). It is based on the same principle as the (one sample) Kolmogorov test. Let's first describe the notion of goodness-of-fit tests.

1. The goodness-of-fit tests

Given a sample, one often has to formulate a hypothesis about which distribution generated that sample. One usually has a favorite candidate distribution, and a goodness-of-fit test (or "test of fit") will determine how likely it is that this distribution generated the sample. In this illustration, a sample of numerical values is pitted against two candidate normal distributions. Clearly, the fit with the first one (figure 2.2) is rather poor, whereas it is much better with the second distribution (figure 2.3).

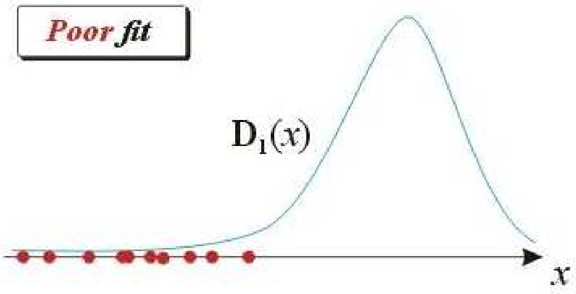


Figure 2.2: Poor fit

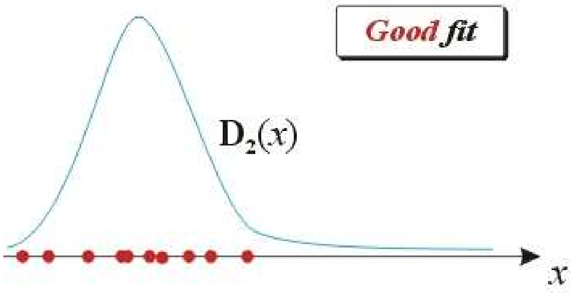


Figure 2.3: Best fit

2. The one-sample Kolmogorov test

It is one of the most important goodness-of-fit tests, together with the Chi-square goodness-of-fit test. Given a sample of observations on a numerical variable x and a completely determined distribution function $F(x)$, the Kolmogorov test will test the H_0 hypothesis according to which the sample originates from the reference distribution $F(x)$. For that purpose, it calculates a quantity D, the "Kolmogorov statistic" from the sample. D is a measure of the departure of the sample cumulated distribution function from $F(x)$. At this step, the theoretical distribution of D when H_0 is true is known. A large value of D is an indication that the sample distribution function departs substantially from $F(x)$, and leads to rejecting H_0. The Kolmogorov test is non-parametric (distribution-free).

3. The two-sample Kolmogorov test

The test first calculates $F_1(x)$ and $F_2(x)$, the respective cumulated distribution functions of the two samples. A quantity D, that measures the discrepancy between these two functions, is then calculated. At this stage, the theoretical distribution of D when H_0 is true is known. A large value of D is an indication that the two samples are too different for being reasonably believed to have been generated by the same underlying probability distribution, and leads to the rejection of H_0.

Let's apply these explanations on a small example. If we consider X with x_1=1, x_2=4, x_3=7, x_4=9 and x_5=10 and Y with y_1=2, y_2=3, y_3=5, y_4=6, y_5=8, we obtain the first two columns of table 2.3. Then we compute the cumulative distribution for each sample that results in adding columns 3 and 4 of the same table. As you can see, since the first numerical value belongs to X, we increase the cumulative distribution by $1/m$, m being the size of sample X. The cumulative distribution for Y remains at zero. The second value belongs to Y, thus we increase the cumulative distribution by $1/n$, n being the size of sample Y. And so on. We then calculate the distribution representing the difference between both cumulative distributions. We thus add the last column of table 2.3. Let's denote the critical value being

column 1	**column 2**	**column 3**	**column 4**	**column 5**
X	Y	*cumulative* X	*cumulative* Y	*difference*
1		$1/5 = 0.2$	0	0.2
	2	0.2	$1/5 = 0.2$	0
	3	0.2	$2/5 = 0.4$	0.2
4		$2/5 = 0.4$	0.4	0
	5	0.4	$3/5 = 0.6$	0.2
	6	0.4	$4/5 = 0.8$	0.4
7		$3/5 = 0.6$	0.8	0.2
	8	0.6	$5/5 = 1$	0.4
9		$4/5 = 0.8$	1	0.2
10		$5/5 = 1$	1	0

Table 2.3: Table for the example of the two-sample Kolmogorov test

M[8]. If the greatest difference (highest numerical value in final column) is greater than M, we **can reject** the null hypothesis stating the two samples come from the same distribution.

2.2.7 The analysis

In this subsection, you will be given the details for the number of pages. For the other variables, please see appendix C.

1. The number of pages

The problem here is to test if there is a location difference between the males and the females regarding the number of pages. So if we assume a location shift, the Wilcoxon-Mann-Whitney test is appropriate. In table 2.4, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 208 and for the females 257. U_m (the males) is 88 and U_n (the females) is 137. From Neave's table of critical values (appendix D), we can read that using a two-tail test at the 5 % level, the critical value is 64. So the conclusion can be easily drawn: since the lowest value (88) is situated above 64, we **cannot reject** the hypothesis that both distributions have the same location.

Value	1	1	2	2	4	4	5	6	10	14
Rank	1,5	1,5	3,5	3,5	5,5	5,5	7	8	9	10
Value	16	28	30	31	33	34	39	46	60	64
Rank	11	12	13	14	15	16	17	18	19	20
Value	66	73	75	78	80	148	172	359	659	813
Rank	21	22	23	24	25	26	27	28	29	30

Table 2.4: Wilcoxon-Mann-Whitney's table for the number of pages

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 96,4 pages whereas it is 100,5 pages for the females. In table 2.5, you can find the deviations, ranks and squares of these ranks. The sum

Deviation	16,4	18,4	21,4	23,4	30,4	32,4	36,4	47,5	50,4	61,5
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	65,4	66,5	67,5	68,4	70,5	71,5	80,4	86,4	86,5	92,4
Rank	11	12	13	14	15	16	17	18	19	20
Square	121	144	169	196	225	256	289	324	361	400
Deviation	94,4	94,5	95,5	96,5	98,5	99,5	99,5	258,5	558,5	716
Rank	21	22	23	24	25	26,5	26,5	28	29	30
Square	441	484	529	576	625	702,25	702,25	784	841	900

Table 2.5: Squared ranks table for the number of pages

[8] the critical value is obtained by reading a table corresponding to the statistical method we use

of the squared ranks for the females is T = 6562,5. The mean of the squared ranks for all thirty observations is 315,15. S^2 can be calculated using the formula above (see subsection "Squared rank test for variance") whence S = 769,96. Thus Z = (6562,5 - 15*315,15)/769,96 giving 2,38 as a result. Since Z is above 1,96 (the critical value), we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that both samples come from identically distributed populations. In table 2.6, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m).

Females	Males	S(f)	S(m)	S(f) - S(m)
1		0,0667	0	0,0667
1		0,1333	0	0,1333
	2	0,1333	0,0667	0,0667
2		0,2	0,0667	0,1333
	4	0,2	0,1333	0,0667
4		0,2667	0,1333	0,1333
5		0,3333	0,1333	0,2
6		0,4	0,1333	0,2667
	10	0,4	0,2	0,2
14		0,46667	0,2	0,2667
	16	0,4667	0,2667	0,2
	28	0,4667	0,3333	0,1333
30		0,5333	0,3333	0,2
	31	0,5333	0,4	0,1333
33		0,6	0,4	0,2
34		0,6667	0,4	0,2667
39		0,7333	0,4	0,3333
	46	0,7333	0,4667	0,2667
	60	0,7333	0,5333	0,2
	64	0,7333	0,6	0,1333
	66	0,7333	0,6667	0,0667
	73	0,7333	0,7333	0
	75	0,7333	0,8	0,0667
	78	0,7333	0,8667	0,0667
	80	0,7333	0,9333	0,2
148		0,8	0,9333	0,1333
172		0,8667	0,9333	0,0667
359		0,9333	0,9333	0
659		1	0,9333	0,0667
	813	1	1	0

Table 2.6: Smirnov's table for testing identically distributed populations

The difference of greatest magnitude is 0,3333 (final column). With a two-tail test at a nominal 5 % level, the critical value is 0,5333 (see table in appendix D). So we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

2.2.8 Summary of the results

In table 2.7, you will find the results of each test described above with the results associated to each numerical variable. First of all, let's denote $min(U_m, U_n)$ as MIN and the difference of greatest magnitude for the Smirnov test as DIFF.

The critical values and the null hypotheses

Regarding the MIN variable, the critical value is 64. That means that if U_m or U_n is below 64, we **can reject** the null hypothesis stating that both distributions have the same location (with a two-tail test at 5%). Let's now have a look at the Z variable. If this one is above 1,96, we **can reject** the null hypothesis stating the variances of both distributions are equal. Finally, for the difference of greatest magnitude, if this one is above 0,5333, we **can reject** the fact that both samples come from identically distributed populations (with a two-tail test at 5%).

The numbers of table 2.7 which are in bold are those which *do not allow* to reject H_0 but which are quite *close* to the critical value. Those in italics are the ones that are far from the critical value. Thus regarding the number of colours for (hyper)text, since the MIN variable is not far from the critical limit of 64, we could expect a significative difference between the medians with a larger sample. This hypothesis should obviously be tested again. Concerning the number of fonts, the MIN variable is so high compared with the critical value that we are pretty sure there is no location difference.

	# pages	# words/page	# char/word	% of white spaces
MIN	88	*108*	**79**	**75**
Z	2,38	**1,87**	*0,44*	*0,98*
DIFF	0,33	*0,1333*	**0,4**	0,33

	# paragraphs/page	# fonts	# colours(txt)	# words(main page)
MIN	94	*110*	**66,5**	*109,5*
Z	2,12	*0,79*	*1*	*0,88*
DIFF	*0,27*	**0,47**	**0,4**	*0,13*

	# colours(background)	# words(description)	ratio(personal content)	# links
MIN	85	*101*	91,5	87
Z	2,74	*0,63*	3,38	3,53
DIFF	*0,27*	0,33	**0,53**	0,33

	# non-personal links	# personal links	# self-photos	# photos
MIN	*107,5*	*106*	*111,5*	61,5
Z	3,59	3,12	3,42	2,48
DIFF	*0,13*	0,33	0,33	**0,47**

	# graphics
MIN	92,5
Z	4,20
DIFF	*0,27*

Table 2.7: Summary of the results

2.2.9 Conclusions

Let's now discuss the possible differences for the numerical variables between men's and women's websites. The conclusions will be illustrated by the male and female distributions for each numerical variable. **But you have to be very cautious when the charts indicate a difference or none since we use a small sample. Actually, the difference can fade or a difference can appear with a larger sample**. Before examining each hypothesis, we have to notice the Smirnov test for a common distribution is never significative since all values are below the critical limit of 0,5333. That is why we won't mention this one in the following descriptions of the conclusions.

1. Are men more expressive than women in terms of text use?

Let's remember that we have to find differences (in the males' favour) regarding the number of pages, words per page and characters per word in order to confirm this hypothesis.

- *The number of pages*

  From the results of table 2.7, we can see the test is not significative regarding the location but is significative for the variance. This means that, with our observations, we **cannot reject** the equality of medians. However we **can reject** the equality of variances. With an error of 5 %, we can thus conclude a variability difference for the number of pages between men's and women's websites. When looking at the chart (see figure 2.4), you can see there are only men in the middle (between 40 and 80 pages) whereas more females than males occupy the tails of their distributions (1-40 pages and more than 141 pages).

- *The number of words per page*

  Regarding this feature, we **cannot reject** the equality of medians nor the equality of variances. If we look at the distributions (figure 2.5), we can notice there is no difference between males and females, since they are more or less equally distributed regarding this feature.

- *The number of characters per word*

  Here again, the equality of medians and variances cannot be rejected. However the chart (figure 2.6) indicates there is a difference with more women belonging to the right half of the distributions and more men in the left half of the distributions. So, according to the chart, women would use longer words than men.

- In general...

  Since there is only a difference of variability for the number of pages, we can't conclude a gendered difference regarding expressiveness in terms of text use.

2. Are women's websites more spaced out than men's?

The necessary variables for this hypothesis are the proportion of white spaces and the number of paragraphs per page.

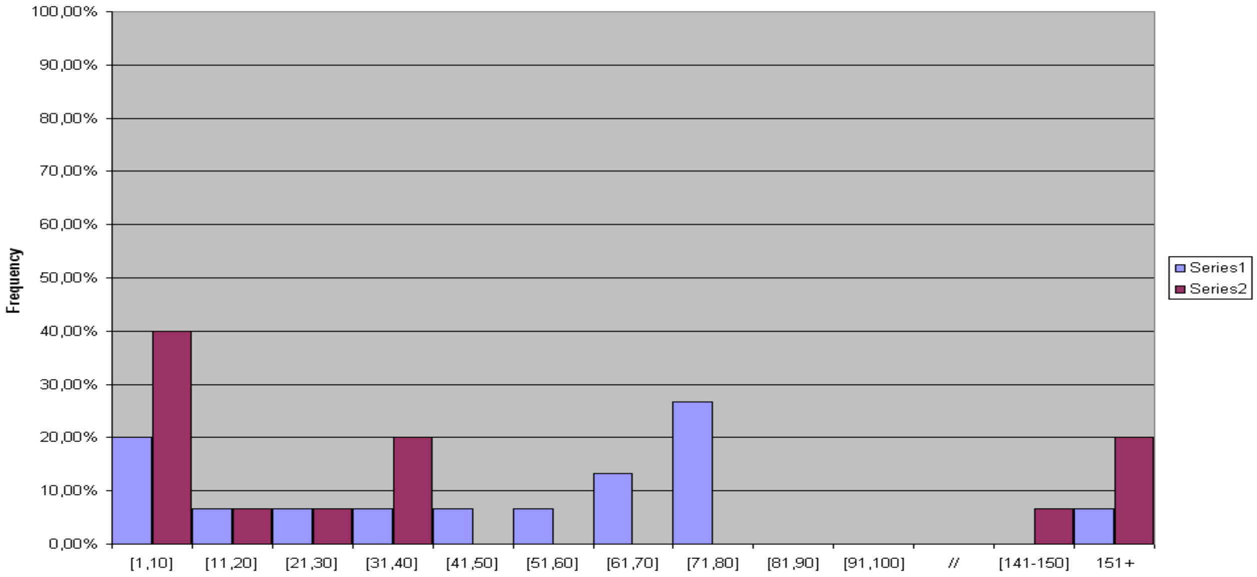


Figure 2.4: Page distributions

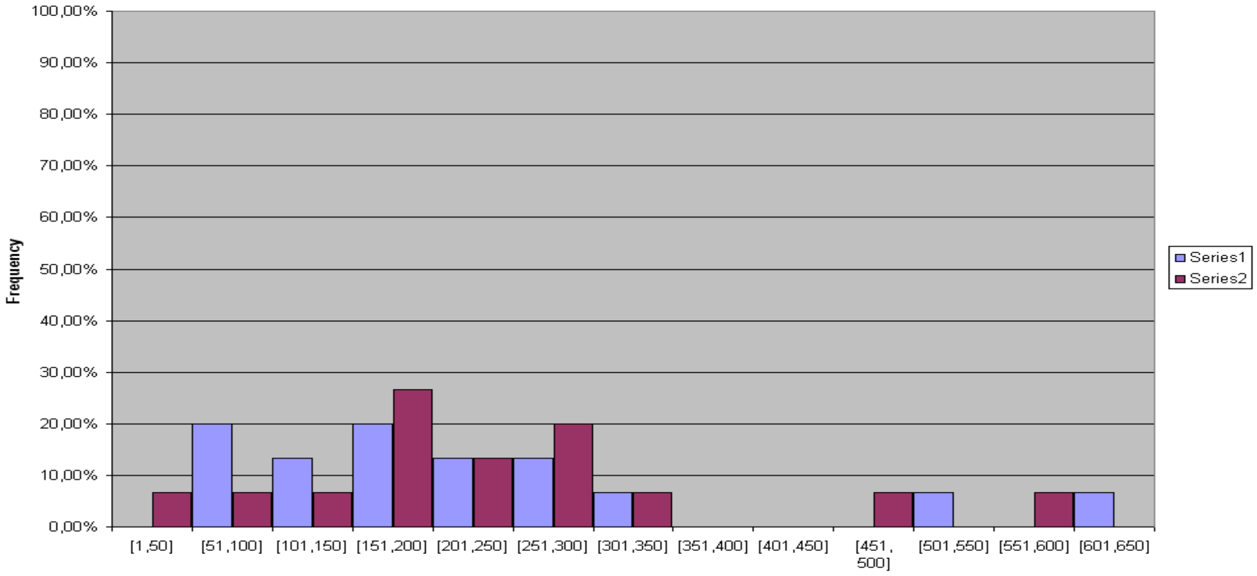


Figure 2.5: Words per page distributions

- *The proportion of white spaces*

  From the results of table 2.7, we can see the location and variance tests are not significative (even if the MIN value is not far from the critical value). Figure 2.7 shows us males and females are more or less equally distributed.

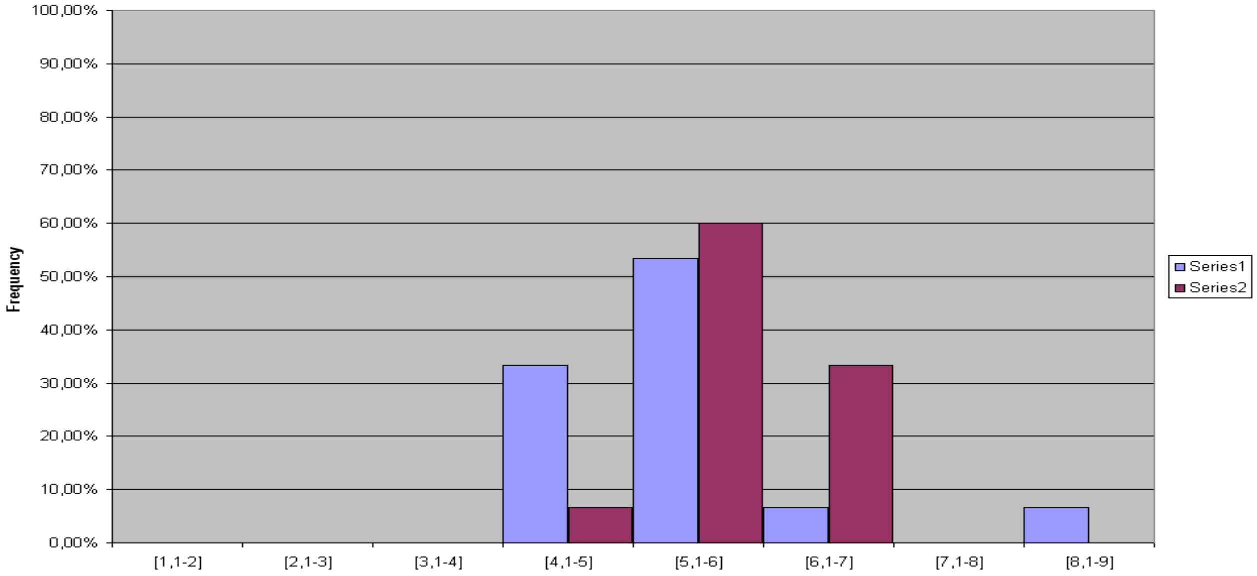


Figure 2.6: Distributions of the characters per word

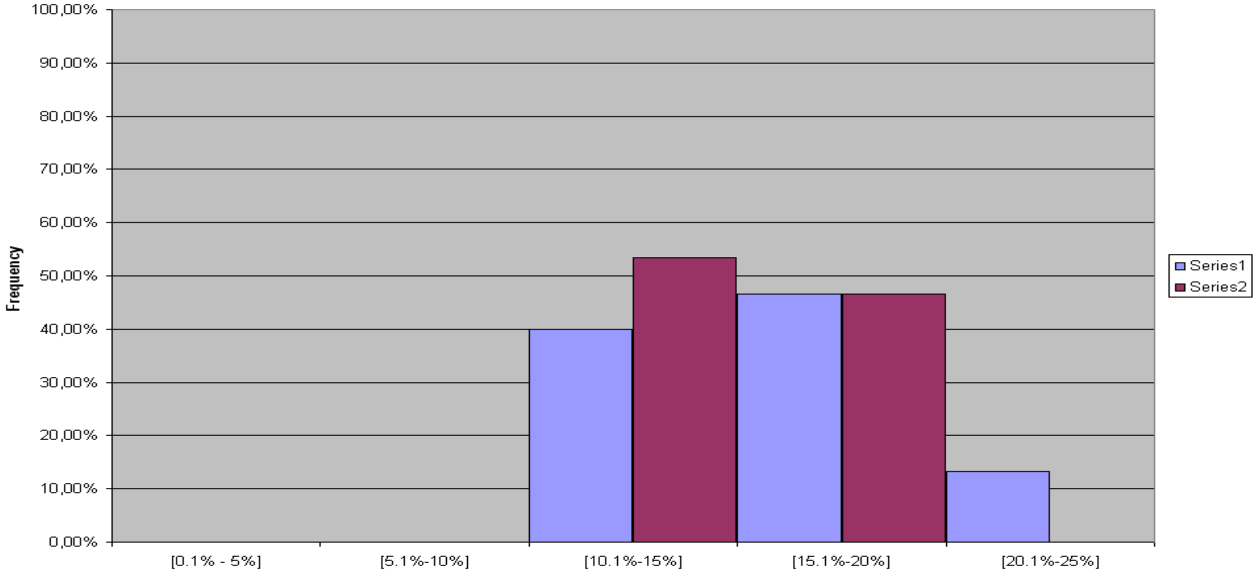


Figure 2.7: Distributions of the proportion of white spaces

- *The number of paragraphs per page*

  Here the location test is not significative. However we do have a significative test for the variances, meaning we have a variability difference between men's and women's websites for this feature (with an error of 5 %). The chart (figure 2.8) indicates there are almost exclusively females between 30 and 55 paragraphs per page. Men outscore

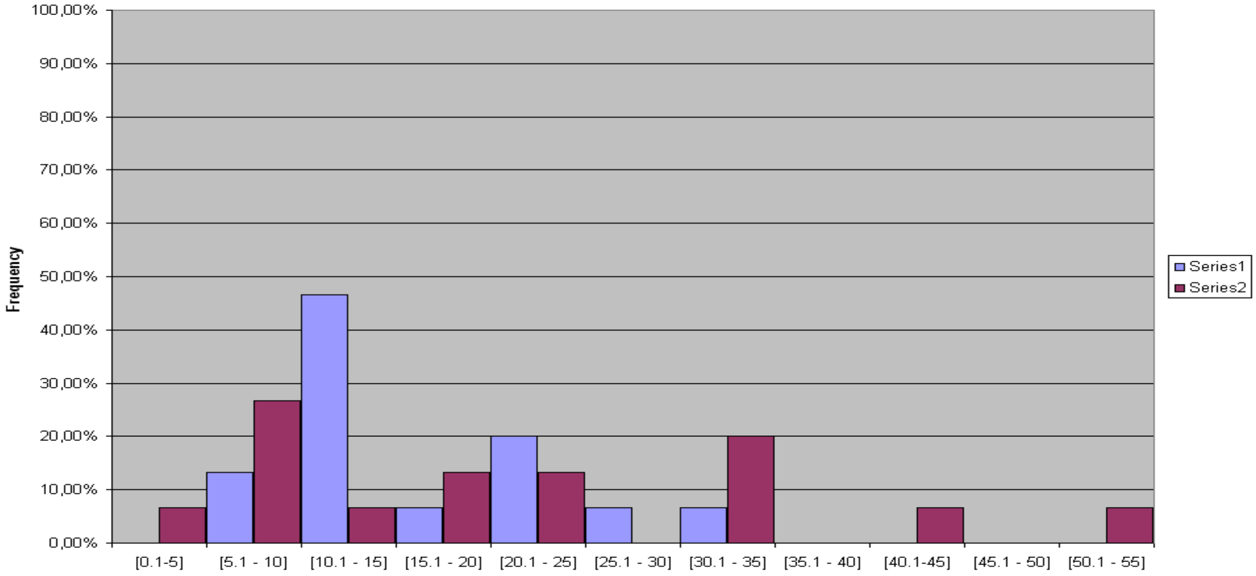


Figure 2.8: Distributions of the paragraphs per page

women regarding the range [10-15] paragraphs per page.

- In general...

  Apart from a variability difference for the proportion of white spaces, we can't draw any other conclusion.

3. Do women use more fonts than men?

Nor the location test or the variance test are significative. As you can see on figure 2.9, males and females are equally distributed.

4. Do women use more colours for text and hypertext?

Here again we do not have significative tests for location and variance, even if the MIN value is very close to the critical value. On the chart (figure 2.10), we can notice males and females have more or less the same distribution, apart from the fact there are more women using three colours and there are more men using four.

5. Do women have fewer words for the main page?

The same statement can be made again (no significative tests). The chart (figure 2.11) indicates both distributions have the same shape.

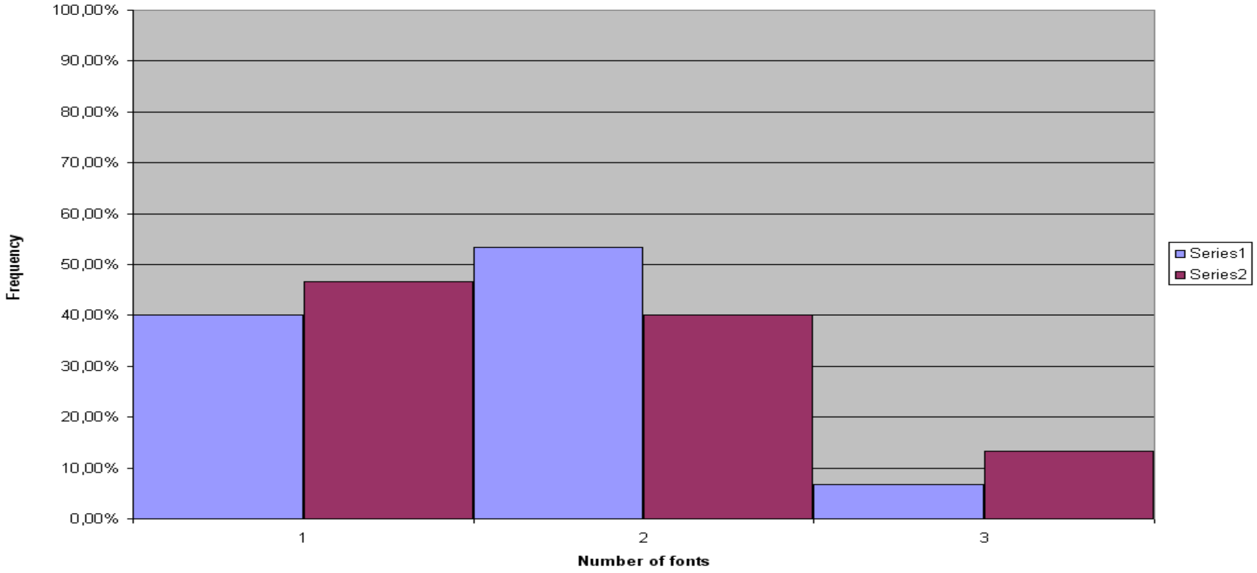


Figure 2.9: Fonts distributions

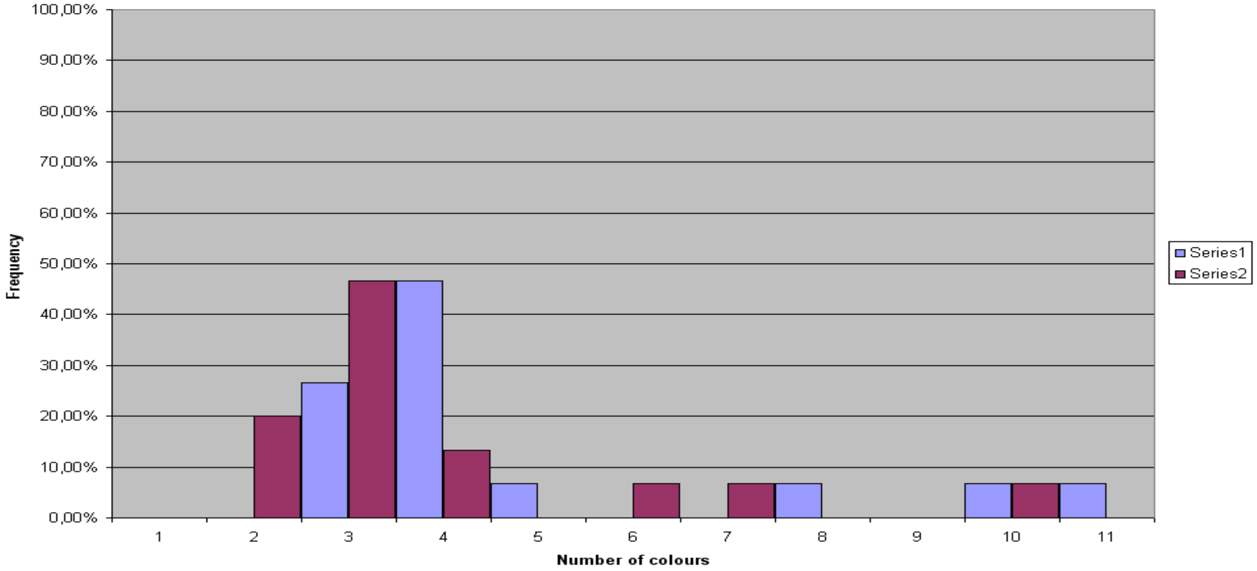


Figure 2.10: Distributions of the colours for (hyper)text

6. Do women's websites have a more colourful background?

Table 2.7 shows that we do have a variability difference between men's and women's websites regarding this feature (with an error of 5 %). According to the chart (figure 2.12), we can notice there are more women using only one colour. What is interesting is the fact

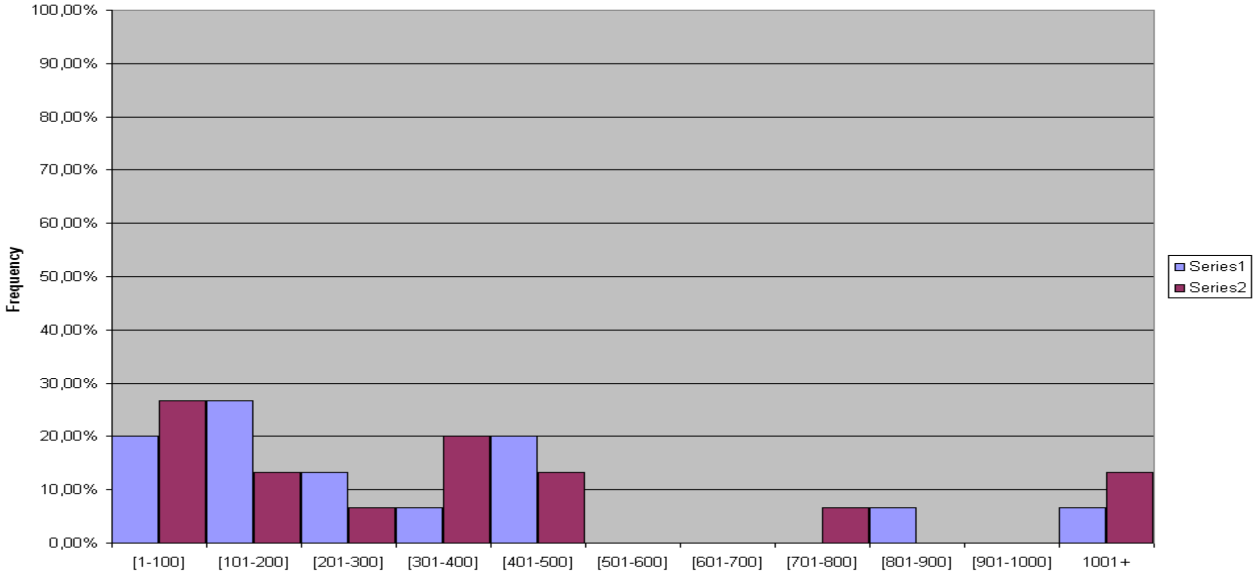


Figure 2.11: Distributions of the words for the main page

there are no women using more than 5 colours compared with men using from 6 up to 13 colours for their backgrounds.

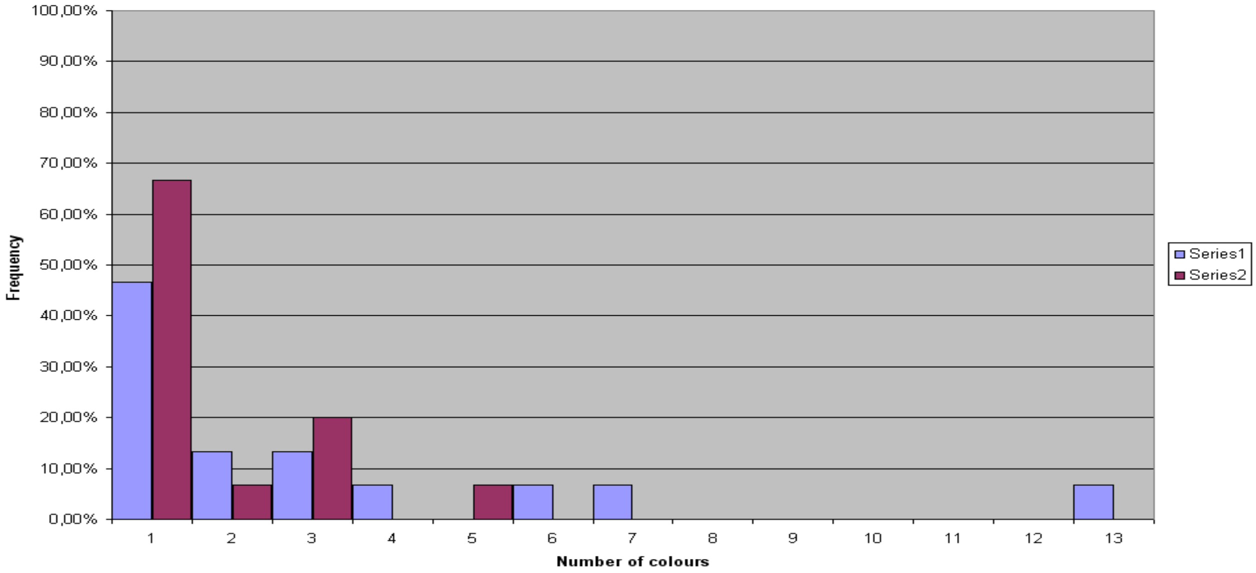


Figure 2.12: Distributions of the background colours

7. Do men describe themselves in a longer way than women do?

Concerning this feature, none of the tests is significative. On the chart (figure 2.13), we can see there are more women in the middles of the distributions, that is to say between 100 and 300 words to describe themselves. But there are more men in the tails of the distributions, that is to say men tend to use between 1 and 100 words and between 400 and 900 words. We can also notice there is a woman describing herself in a very long way (between 1600 and 1650 words).

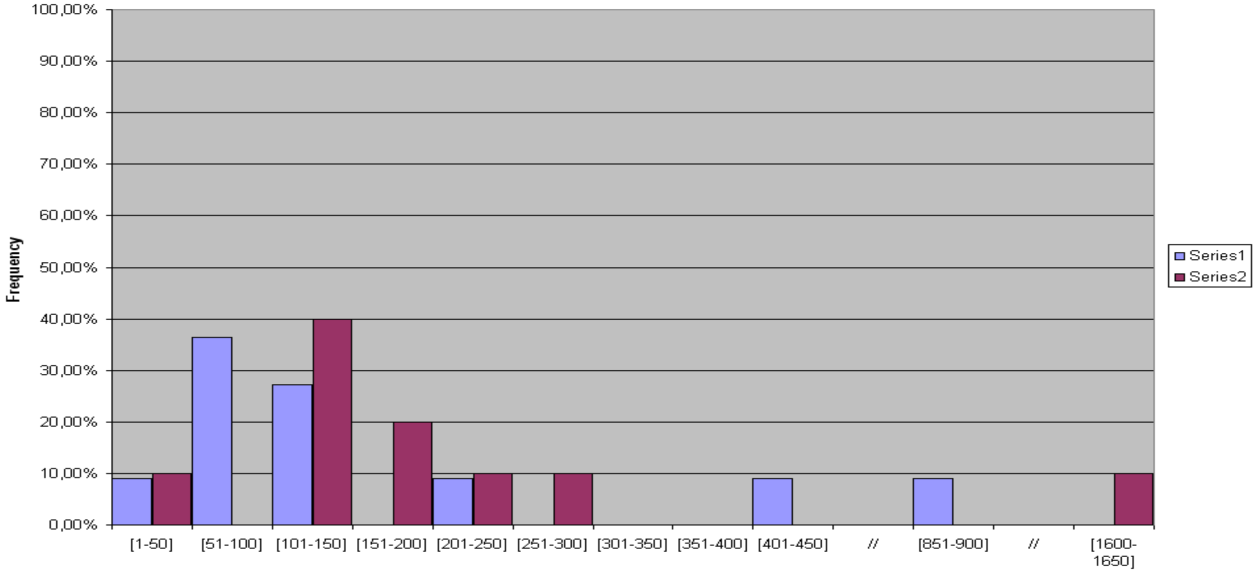


Figure 2.13: Distributions of the words for the self-description

8. Do men have a website with a higher percentage of personal content?

Regarding this feature, we have a variability difference between both genders. You can see on the chart (figure 2.14) there are more women having between 1 and 20 % of their pages devoted to personal content. The middles of the distributions (30 % - 70 %) are occupied by men whereas the tails of the distributions contain 20 % of the women having personal content on their website.

9. Is it true that men and women do not differ regarding the number of links in their website?

We can make the same statement as the previous one, that is to say we can only conclude a variability difference between genders regarding this feature. On the chart (figure 2.15), we can notice there are more women including between 1 and 100 links in their pages. The rest of the distributions is occupied by more men, especially in the middles of the distributions since there are no women between 300 and 800 links. However, we can see there are more

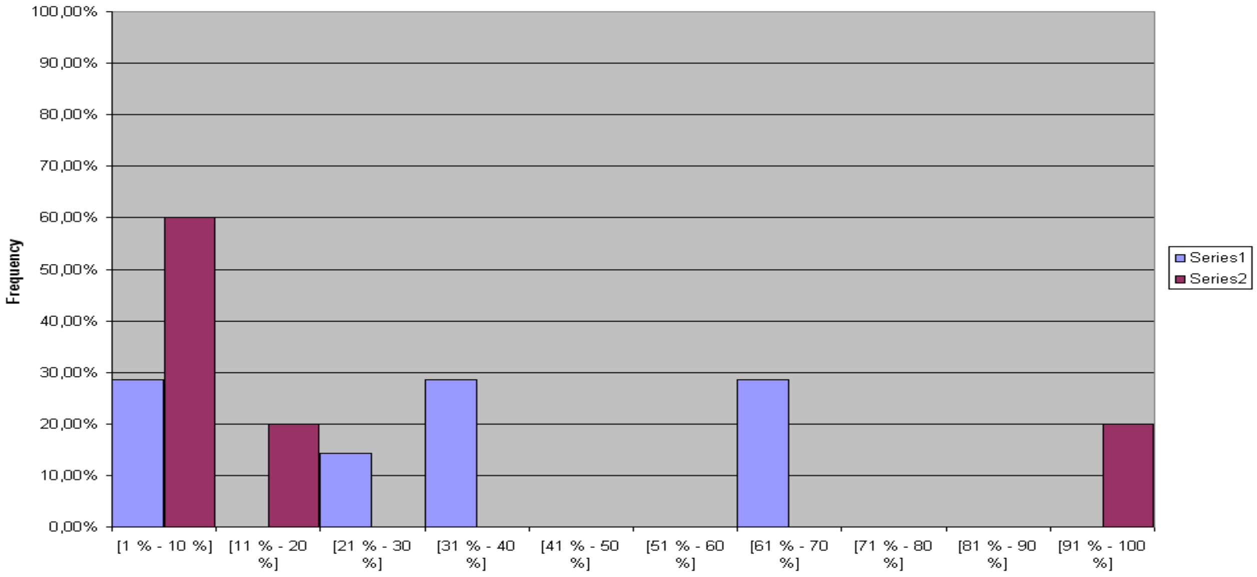


Figure 2.14: Distributions of the percentage of personal content (ratio)

women including more than 1000 links.

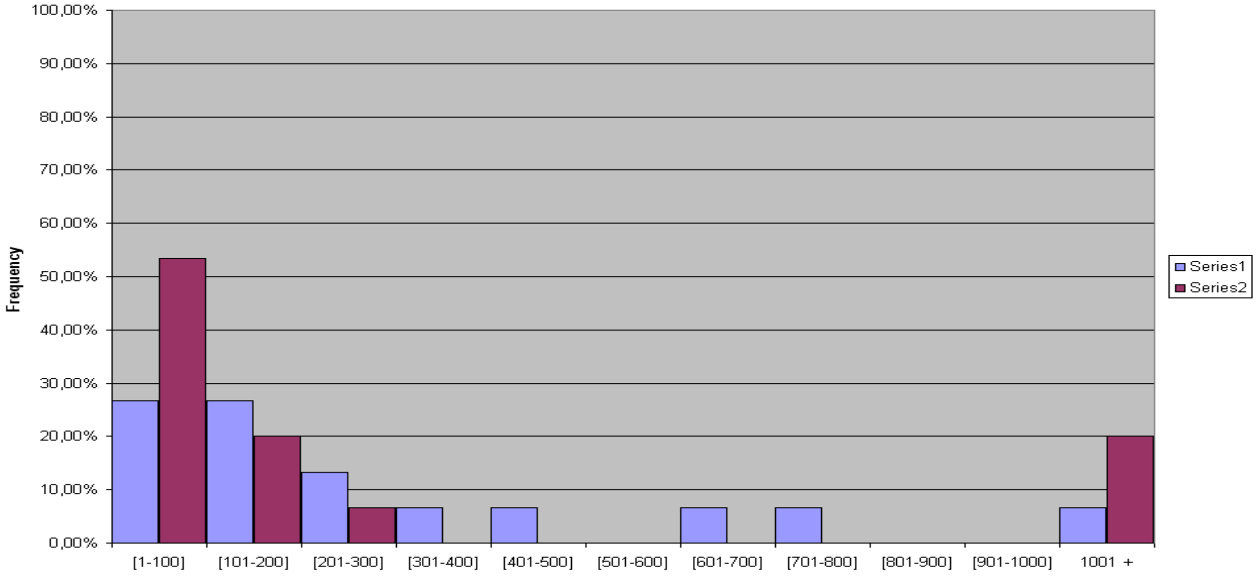


Figure 2.15: Links distributions

10. Do men and women differ in the type of links contained in their website?

- *Non-personal links*

  Here again we can conclude a variability difference. The chart (figure 2.16) shows males and females share the same shape regarding their distributions.

- *Personal links*

  We have a variability difference again regarding this second feature. On the chart (figure 2.17), it is visible there are more males in the range [21-30] links to other people's pages. Apart from that, there are only males between 80 and 100 links and only females between 160 and 440 links.

- In general...

  The only conclusion we can draw in both cases for the type of links is a variability difference between genders.

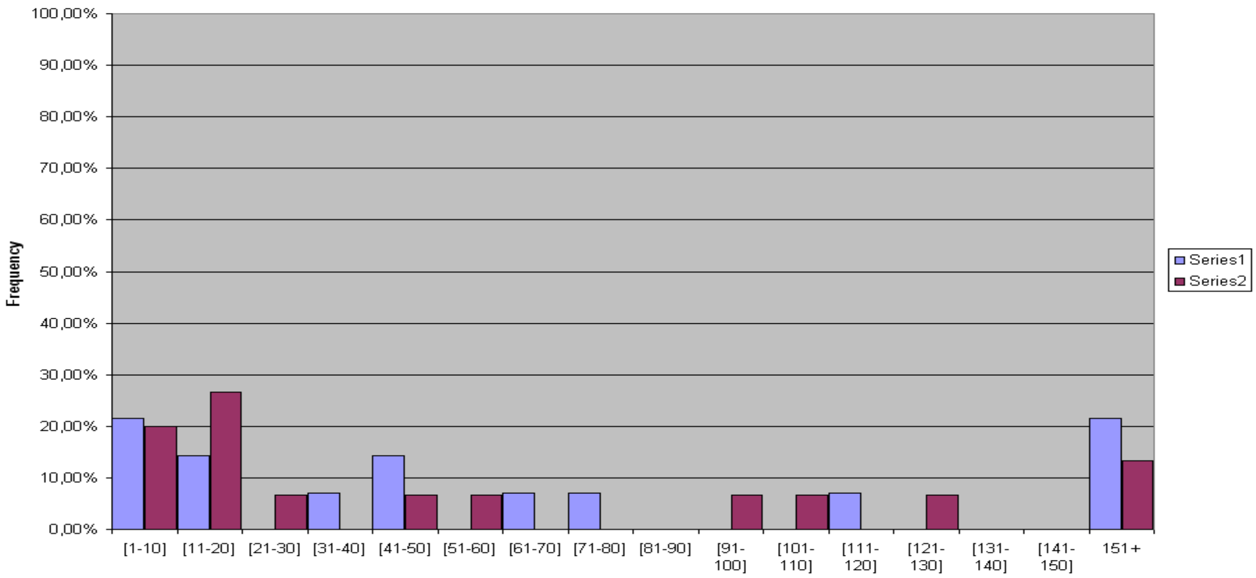


Figure 2.16: Distributions of the non-personal links

11. Do men show more self-photos than women?

We actually have a variability difference between males and females regarding the self-photos. The chart (figure 2.18) indicates there are more women showing two pictures of themselves. The middles of the distributions are occupied by males and females (with slightly more men) with one woman showing twelve pictures of herself.

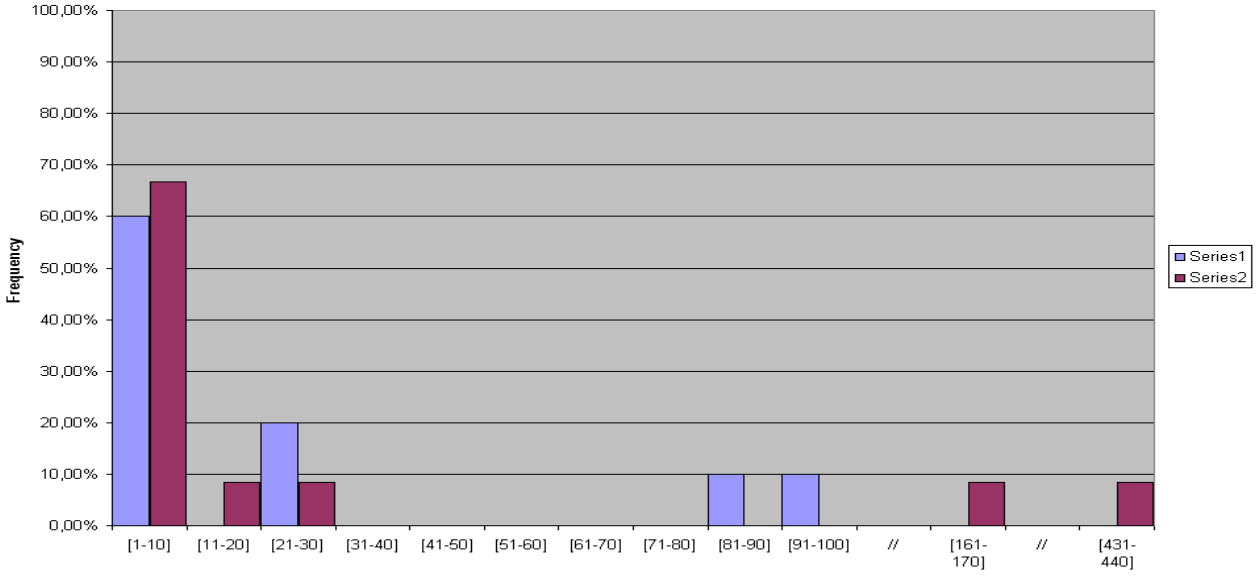


Figure 2.17: Distributions of the personal links

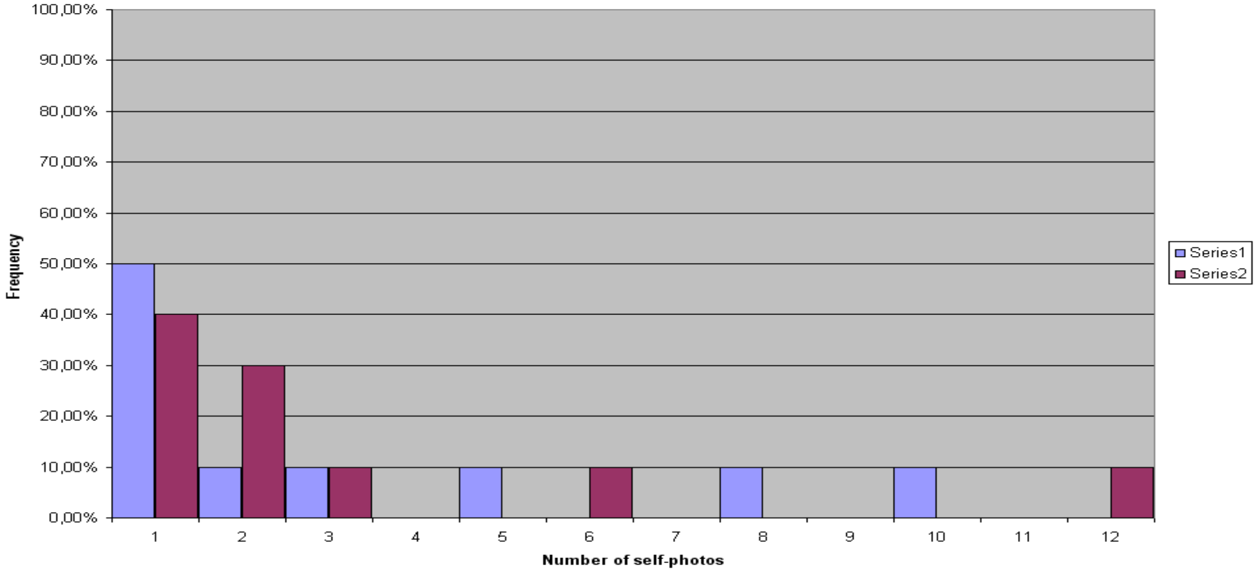


Figure 2.18: Self-photos distributions

12. Do men show more photos than women?

Here we can conclude a variability difference AND a location difference. However, the Hodges-Lehmann estimator indicates there is only a difference of one photo between men's and women's websites. As you can see on the chart (2.19), the middles of the distributions contain exclusively males (between 10 and 120 photos) whereas there is one woman (out of the two who put photos on their sites) showing more than 1080 pictures on her own site![9]

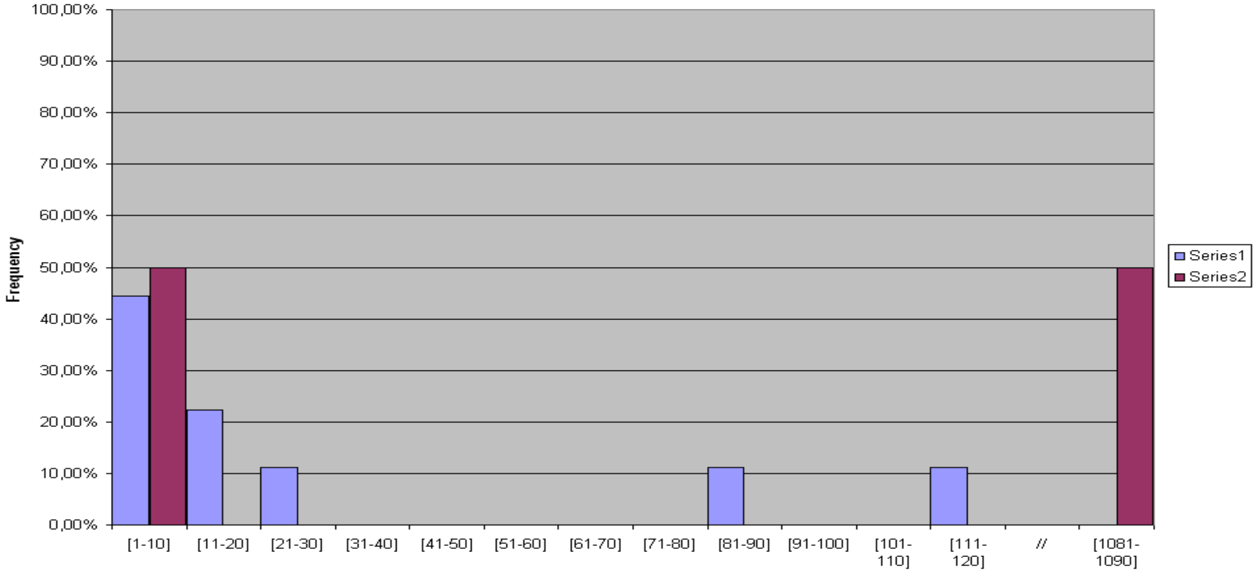


Figure 2.19: Photos distributions

13. Do men use more graphics than women?

Regarding this feature, we can conclude a variability difference. On the chart (figure 2.20) there are more women using 10 graphics than men. Regarding the middles of the distributions, there are almost exclusively males (between 40 and 180 graphics) whereas there are slightly more females using more than 200 graphics on their sites.

2.2.10 Classification of the academics

The purpose of this classification is to find genderless differences among the academics since we can only conclude a variability difference from the hypothesis tests. First of all, you will be introduced to the method used to achieve this purpose and then you will find the summary of the results.

1. K-means clustering

Let's first understand what is meant by "clustering" and then we will focus on the K-means method.

[9] Please note that men and women who do not show photos on their sites are not visible on the chart. This one only shows the distributions of the number of photos for the populations who do show photos on their sites.

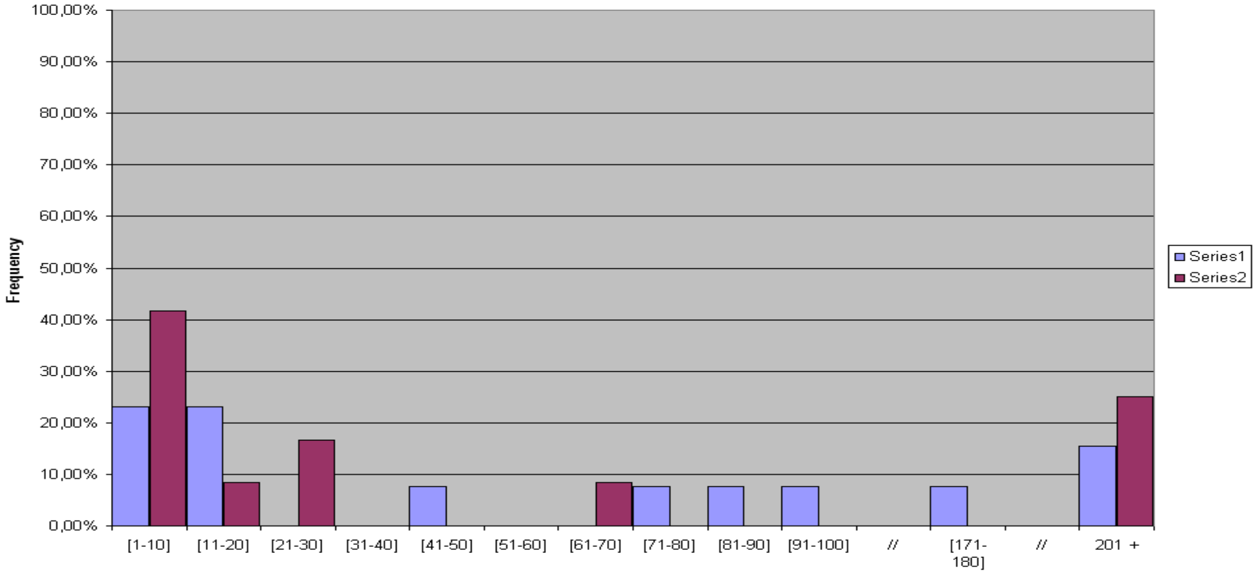


Figure 2.20: Graphics distributions

Definition of clustering

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a dataset into subsets (clusters), so that the data in each subset (ideally) share some common trait, often proximity according to some defined distance measure. Data clustering algorithms can be hierarchical or partitional. With hierarchical algorithms, successive clusters are found using previously established clusters, whereas partitional algorithms determine all clusters in one go.

The K-means clustering

The k-means algorithm assigns each point to the cluster whose center (or centroid) is nearest. The centroid is the point generated by computing the arithmetic mean for each dimension separately for all the points in the cluster. For example, if we consider a three-dimension data set with a cluster consisting of two points X = (x1, y1, z1) and Y = (x2, y2, z2), then the centroid Z becomes Z = (x3, y3, z3), where:

$$x3 = \frac{(x1 + x2)}{2}$$

and

$$y3 = \frac{(y1 + y2)}{2}$$

and

$$z3 = \frac{(z1 + z2)}{2}.$$

This is the basic structure of the algorithm (J. MacQueen, 1967):

- Randomly generate k clusters and determine the cluster centers or directly generate k seed points as cluster centers.
- Assign each point to the nearest cluster center.
- Recompute the new cluster centers.
- Repeat until some convergence criterion is met (usually that the assignment hasn't changed).

Let's insist on the fact the algorithm always converges to a solution which depends on the initial partitioning. Thus if the initial partitioning is not properly chosen, the solution won't be accurate.

The main advantages of this algorithm are its simplicity and speed, which allows it to run on large datasets. Yet it does not systematically yield the same result with each run of the algorithm. Rather, the resulting clusters depend on the initial assignments. The k-means algorithm maximizes inter-cluster (or minimizes intra-cluster) variance, but does not ensure that the given solution is not a local minimum of variance. So, in summary, the philosophy of the K-means algorithm consists of iterating until stable, that is to say until no object moves from group. The steps can be summarized as follows according to figure 2.21:

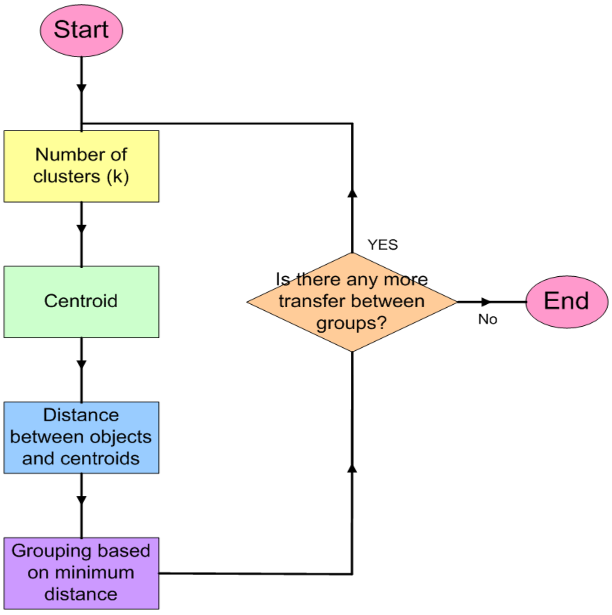

Start
Number of clusters (k)
Centroid
Distance between objects and centroids
Grouping based on minimum distance
Is there any more transfer between groups?
YES
No
End

Figure 2.21: K-means algorithm

1. Determine the centroids coordinates.
2. Determine the distance of each object to the centroids.
3. Group the objects on the basis of the minimum distance.

The numerical example below is given to understand this simple iteration. Suppose we have several objects (4 types of medicines) and each object has two attributes or features as shown in table 2.8. Our goal is to group these objects into K = 2 groups of medicines based on the two features (pH and weight index). Each medicine represents one point with two attributes (X, Y) that we can represent as coordinates in an attribute space as shown in figure 2.22.

Object	**attribute 1 (X): weight index**	**attribute 2 (Y): pH**
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Table 2.8: The four types of medicine with weight and pH index

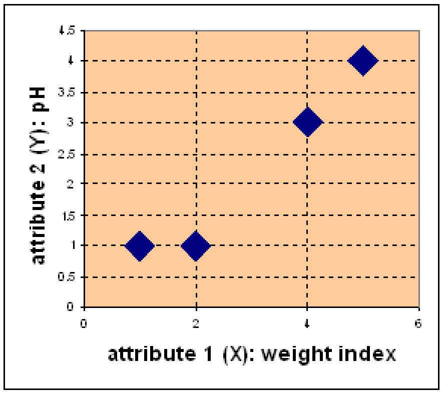


Figure 2.22: Attribute coordinates for medicines

1. Initial centroids values

Suppose we use medicine A and medicine B as the first centroids. Let c_1 and c_2 denote the coordinates of the centroids, thus $c_1 = (1, 1)$ and $c_2 = (2, 1)$, as shown in figure 2.23.

2. Distance between objects and centroids

We calculate the distance between the cluster centroids and each object. Let us use the Euclidean distance in order to obtain a distance matrix for iteration 0 (see figure 2.24).

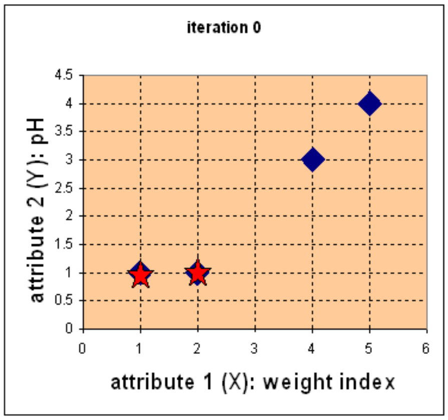


Figure 2.23: Centroids coordinates

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{array}$$

$$\begin{array}{cccc} A & B & C & D \end{array}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{array}{l} X \\ Y \end{array}$$

Figure 2.24: Distance matrix at iteration 0

Each column in the distance matrix symbolizes an object. The first row of the distance matrix corresponds to the distance of each object to the first centroid and the second row is the distance of each object to the second centroid. For example, the distance from medicine C = (4, 3) to first centroid $c_1 = (1, 1)$ is

$$\sqrt{(4-1)^2 + (3-1)^2} = 3,61$$

and its distance to second centroid $c_2 = (2, 1)$ is

$$\sqrt{(4-2)^2 + (3-1)^2} = 2,83$$

etc.

3. Objects clustering

We assign each object to a group based on the minimum distance. Thus, medicine A is assigned to group 1, medicines B, C and D to group 2. The element of the group matrix (see figure 2.25) corresponding to a particular object is 1 *if and only if* this object is assigned to that group.

$$\mathbf{G}^0 = \begin{array}{cc} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} & \begin{array}{l} group-1 \\ group-2 \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array}$$

Figure 2.25: Group matrix at iteration 0

4. Iteration-1, determining the new centroids

Knowing the members of each group, we now compute the new centroid for each group based on these new memberships. Group 1 only has one member thus the centroid remains $c_1 = (1, 1)$. Group 2 now has three members, thus the centroid is the average coordinates among the three members: $c_2 = (\frac{2+4+5}{3}, \frac{1+3+4}{3}) = (\frac{11}{3}, \frac{8}{3})$ as represented in figure 2.26.

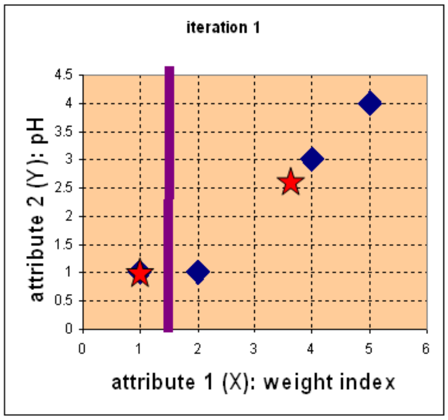


Figure 2.26: New centroids at iteration 1

5. Iteration-1, distances between objects and new centroids

The next step is to compute the distance of all the objects to the new centroids. As for step 2, we have a new distance matrix at iteration 1 that you can see in figure 2.27.

$$\mathbf{D}^1 = \begin{array}{cc} \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} & \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{array}{l} X \\ Y \end{array} \end{array}$$

Figure 2.27: Distance matrix at iteration 1

6. Iteration-1, objects clustering

As for step 3, we assign each object to a group based on the minimum distance. According to the new distance matrix, we move medicine B to group 1 while all the other objects remain in the group. The new group matrix is represented in figure 2.28.

$$\mathbf{G}^1 = \begin{array}{c} \\ \\ \\ \end{array}\begin{array}{cccc|l} 1 & 1 & 0 & 0 & group-1 \\ 0 & 0 & 1 & 1 & group-2 \\ A & B & C & D & \end{array}$$

Figure 2.28: Group matrix at iteration 1

7. Iteration-2, determining the new centroids

We now repeat step 4 to calculate the coordinates of the new centroids based on the clustering of previous iteration. Group 1 and group 2 both have two members, thus the new centroids are

$$c_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1.5, 1)$$

and

$$c_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4.5, 3.5)$$

as you can see in figure 2.29.

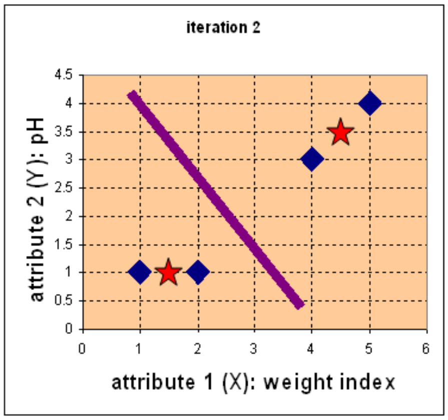


Figure 2.29: Attribute coordinates at iteration 2

8. Iteration-2, distances between objects and centroids

If we repeat step 2 again, we have a new distance matrix as shown in figure 2.30.

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \begin{array}{l} \mathbf{c}_1 = (1\tfrac{1}{2}, 1) \quad group-1 \\ \mathbf{c}_2 = (4\tfrac{1}{2}, 3\tfrac{1}{2}) \quad group-2 \end{array}$$

$$\begin{array}{cc} \begin{array}{cccc} A & B & C & D \end{array} & \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{array}{c} X \\ Y \end{array} \end{array}$$

Figure 2.30: Distance matrix at iteration 2

9. Iteration-2, objects clustering

Again, we assign each object to a group based on the minimum distance. We thus obtain the new group matrix you can see in figure 2.31. We can notice that $G^2 = G^1$. This means the objects haven't moved from group and won't anymore. Thus, the computation of the k-means clustering has reached its stability and no more iteration is needed. The final result is shown in table 2.9.

$$\mathbf{G}^2 = \begin{array}{cc} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{array}{l} group-1 \\ group-2 \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} & \end{array}$$

Figure 2.31: Group matrix at iteration 2

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Table 2.9: Final grouping of the subjects for the k-means example

2. The results

The analysis has been carried out by choosing two clusters. This choice has been made according to the fact that, by choosing three clusters, we couldn't get more than one observation in clusters 1 and 3. In table 2.10, you will find the mean and the intervals (centered around the cluster mean) characterizing each variable in each cluster.

Cluster 1 consists of only two females. The other academics belong to cluster 2. The first female of cluster 1 differs from the other academics by the number of pages (613), the ratio of personal content (99%), the number of links (5764) and non-personal links (802), the number of photos (1083) and graphics (2855). The second one differs by the number of pages (359), the number of links (6163), the number of links to other people's pages (439) and non-personal

1	# pages	509	[296,87 ; 721,13]
2	# pages	69,11	[0 ; 221,06]
1	# words/page	90	[0 ; 180,93]
2	# words/page	243,83	[96,10 ; 391,56]
1	# char/word	5,4	[5,26 ; 5,54]
2	# char/word	5,7	[4,89 ; 6,51]
1	% white spaces	15,38%	[15% ; 15,76%]
2	% white spaces	15,85%	[13,51% ; 18,19%]
1	# paragraphs/page	10,16	[0,38 ; 19,94]
2	# paragraphs/page	19,47	[7,86 ; 31,08]
1	# words for the self-description	148,5	[130,82 ; 166,18]
2	# words for the self-description	175,29	[0 ; 511,75]
1	ratio (personal content)	49,80%	[0% ; 100%]
2	ratio (personal content)	9,57%	[0% ; 28,94%]
1	# fonts	2	[2 ; 2]
2	# fonts	1,64	[0,96 ; 2,32]
1	# colours (txt)	4,5	[0,96 ; 8,04]
2	# colours (txt)	4,39	[1,98 ; 6,8]
1	# words for main page	441,5	[68,85 ; 814,15]
2	# words for main page	436,61	[0 ; 1008,41]
1	# links	5963,5	[5681,36 ; 6245,64]
2	# links	271,93	[0 ; 712,19]
1	# personal links	221,5	[0 ; 529,09]
2	# personal links	16,57	[0 ; 53,75]
1	# non-personal links	2166,5	[236,81 ; 4096,19]
2	# non-personal links	68,82	[0 ; 173,25]
1	# self-photos	47,5	[0 ; 106,20]
2	# self-photos	2,04	0 ; 6,15]
1	# photos	541,5	[0 ; 1307,3]
2	# photos	9,93	[0 ; 37,16]
1	# graphics	2036	[877,76 ; 3194,24]
2	# graphics	59,32	[0 ; 162,12]
1	# colours (bck)	2,5	[1,79 ; 3,21]
2	# colours (bck)	2,43	[0 ; 5,09]

Table 2.10: Results of the K-means clustering method

links (3531), the number of self-photos (89) and finally the number of graphics (1217).

2.3 The binary variables

Let's now analyse the binary variables, that is:

- the type of fonts (classic, girlish),

- the type of colours for text and hypertext (reddish, blueish, black, white, grey),
- if the website is technological,
- the type of colours for the background (soft, dark, reddish, blueish, black, white, grey),
- the type of background (classic, original),
- the presence of a self-description and if this one is professional or private,
- the presence of personal pages (denoted by *ratio*),
- if the academic focusses on his/her credentials,
- the presence of graphic accents,
- the type of links (non-personal, personal),
- the presence of self-photos and photos,
- the type of self-photos (official, non-official, family, friends, colleagues, pets, leisure time, computer-related),
- the type of photos (family, friends, colleagues, pets, leisure time, computer-related),
- the quality of self-photos and photos,
- the presence of graphics,
- and finally, the type of graphics (basic, modern, trendy, artistic, comics, computer-related).

First of all, you will be told about the different tests being used in this case. Then, you will find the analysis and the results.

2.3.1 The Fisher's test

To understand this test, let's consider the following 2*2 table:

	yes	no	
I	A	B	A+B
II	C	D	C+D
	A+C	B+D	A+B+C+D=N

with $\frac{A}{A+B} > \frac{C}{C+D}$

It is possible to exactly compute the probability of the configuration (A,B,C,D) of a table knowing the marginal totals (A+B, C+D, A+C, B+D). Actually, we are confronted with the hypergeometric distribution since the configuration (A,B,C,D) is determined by only one of the cell values, for example A, which probability is the one obtained when we have A successes after A+B exhaustive draws among N objects. So A+C are “good” and B+D “bad”.

$$P_A = \frac{C^A_{A+C}C^B_{B+D}}{C^{A+B}_N} = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

Let's suppose we observe (A,B,C,D). We can therefore test the hypothesis with $\frac{A}{A+B} = \frac{C}{C+D}$ considering the overrun probability:

OP = p { w': w more unfavourable than, as unfavourable as w }
= Pr { configurations more unfavourable to H than (A, B, C, D)
considering the marginal totals (A+B, C+D, A+C, B+D) }
= Σ Pr (A', B', C', D')

A' + B' = A + B
C' + D' = C + D
A' + C' = A + C
B' + D' = B + D
And (A', B', C', D') more unfavourable than (A, B, C, D)

For example, if we have the following 2*2 table:

10	1	11
4	5	9
14	6	20

Then the following table is more unfavourable:

11	0	11
3	6	9
14	6	20

so that OP = Pr(10,1,4,5) + Pr(11,0,3,6)
$= \frac{11!9!14!6!}{20!10!4!1!5!} + \frac{11!9!14!6!}{20!11!0!3!6!}$
= 0,03576 + 0,00216
= 0,03792
→ we **can reject** at a 5 % significance level

The table has $\hat{c}$ levels C = 4 at 0,05, C = 2 at 0,01

It is obvious the Fisher's test is hardly usable if the numbers in the cells differ a lot from 0. We can then assess the probabilities by using factorial log tables. The Fisher-Yates tables allow to carry out the test without having to explicitly compute the overrun probabilities if we use one of the four uncertainty levels of the table. The table gives the greatest value of C (or D) being significant at the uncertainty level 0,05, 0,025, 0,01, 0,005 [Sie88]. So, if C is greater than the significant value, we reject the homogeneity hypothesis. To consult the table, please see appendix D.

2.3.2 The binomial test

1. The binomial distribution

The binomial distribution is the distribution of the number of n successive draws in a dichotomy. The distribution is given by the formula:

$$Pr[X = k] = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$ and where

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

is the binomial coefficient "n choose k" (also denoted C(n, k)), whence the name of the distribution. The formula can be understood as follows: we want k successes (p^k) and $n - k$ failures ($(1-p)^{n-k}$). However, the k successes can occur anywhere among the n trials, and there are $C(n, k)$ different ways of distributing k successes in a sequence of n trials.

In statistics, the binomial test is an exact test of the statistical significance of deviations from a theoretically expected distribution of observations into two categories. For example, suppose a die is rolled 235 times, and 6 comes up 51 times. If the die is fair, we would expect 6 to come up 235/6 = 39.17 times. Is the proportion of 6s significantly higher than would be expected by chance, on the null hypothesis of a fair die? To find an answer to this question using the binomial test, we consult the binomial distribution B(235,1/6) to find out what the probability is of finding exactly 51 6s in a sample of 235 if the true probability of a 6 on each trial is 1/6. We then find the probability of finding exactly 52, exactly 53, and so on up to 235, and add all these probabilities together. That gives us the significance of the observed number of 6s.

The commonest use of the binomial test is in the case where the null hypothesis is that two categories are equally likely to occur. Tables are widely available to give the significance of the observed number of observations in the categories for this case. However, as the example above shows, the binomial test is not restricted to this case.

2. The binomial test in our context

Suppose we want to know a confidence interval for the probability of successes for a random variable that we will call X. This random variable is described by the following 2x2 table:

X	Successes	Failures	Total
Group 1	A	B	A+B
Group 2	C	D	C+D
Total	A+C	B+D	A+B+C+D

In order to achieve our goal, we have to carry out the following procedure:

1. Compute the relative frequencies for the successes, that is to say $\frac{A+C}{A+B+C+D}$.
2. Choose a level for the error risk, for example 0,025.
3. In the chart providing the confidence limits for a confidence coefficient of 1-2α = 0,95, choose the curve corresponding to the size of your sample, that is to say A+B+C+D.
4. On the top line (or the bottom one), locate the point corresponding to the computed relative frequency and draw a line to the chosen curve. You can then read p on the right (or left) line.
5. The limits of the confidence interval for the number of successes then consist of the computed relative frequency and p that you have found by reading the table. You can apply the same procedure for the failures.

2.3.3 The discriminant analysis

1. General purpose

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (1) to go to college, (2) to attend a trade or professional school, or (3) to seek no further training or education. For that purpose the researcher could collect data on numerous variables prior to students' graduation. After graduation, most students will naturally fall into one of the three categories. Discriminant analysis could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

A medical researcher may record different variables relating to patients' backgrounds in order to learn which variables best predict whether a patient is likely to recover completely (group 1), partially (group 2), or not at all (group 3). A biologist could record different characteristics of similar types (groups) of flowers, and then perform a discriminant function analysis to determine the set of characteristics that allows for the best discrimination between the types.

2. Computational approach

To understand the approach, let us consider a simple example. Suppose we measure height in a random sample of 50 males and 50 females. Females are, on the average, not as tall as males, and this difference will be reflected in the difference in means (for the variable $Height$). Therefore, variable $Height$ allows us to discriminate between males and females with a "better-than-chance" probability: if a person is tall, then he is likely to be a male, if a person is short, then she is likely to be a female.

We can generalize this reasoning to groups and variables that are less "trivial". For example, suppose we have two groups of high school graduates: those who choose to attend college after graduation and those who do not. We could have measured students' stated intention to continue on to college one year prior to graduation. If the means for the two groups (those who actually went to college and those who did not) are different, then we can say that intention to attend college as stated one year prior to graduation allows us to discriminate between those who are and are not college bound (and this information may be used by career counselors to provide the appropriate guidance to the respective students). To summarize the discussion so far, the basic idea underlying discriminant function analysis is to determine whether groups differ with regard to the mean of a variable, and then to use that variable to predict group membership (e.g., of new cases).

Analysis of variance

Stated in this manner, the discriminant function problem can be rephrased as a one-way analysis of variance (ANOVA) problem. Specifically, one can ask whether or not two or more groups are significantly different from each other with respect to the mean of a particular variable. It should be clear that, if the means for a variable are significantly different in different groups, then we can say that this variable discriminates between the groups.

3. Interpreting a two-group discriminant function

In the two-group case, discriminant function analysis can also be thought of as (and is analogous to) multiple regression. If we code the two groups in the analysis as 1 and 2, and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via discriminant analysis. In general, in the two-group case we fit a linear equation of the type:

$$Group = a + b_1 * x_1 + b_2 * x_2 + ... + b_m * x_m$$

where a is a constant and b_1 through b_m are regression coefficients. The interpretation of the results of a two-group problem is straightforward and closely follows the logic of multiple regression: those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

4. Classification

Another major purpose to which discriminant analysis is applied is the issue of predictive classification of cases. Once a model has been finalized and the discriminant functions have been derived, how well can we predict to which group a particular case belongs?

Classification functions

These are not to be confused with the discriminant functions. The classification functions can be used to determine to which group each case most likely belongs. Each function allows us to compute classification scores for each case for each group, by applying the formula:

$$S_i = c_i + w_{i1} * x_1 + w_{i2} * x_2 + ... + w_{im} * x_m.$$

In this formula, the subscript i denotes the respective group; the subscripts 1, 2, ..., m denote the m variables; c_i is a constant for the i'th group, w_{ij} is the weight for the j'th variable in the computation of the classification score for the i'th group; x_j is the observed value for the respective case for the j'th variable. S_i is the resultant classification score.

Classification of cases

Once we have computed the classification scores for a case, it is easy to decide how to classify the case: in general we classify the case as belonging to the group for which it has the highest classification score. Thus, if we were to study high school students' post-graduation career/educational choices (e.g., attending college, attending a professional or trade school, or getting a job) based on several variables assessed one year prior to graduation, we could use the classification functions to predict what each student is most likely to do after graduation.

Prediction of group membership

A common result that one looks at in order to determine how well the current classification functions predict group membership of cases is *the classification matrix.* The classification matrix shows the number of cases that were correctly classified and those that were misclassified.

5. Stepwise discriminant analysis

Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups. For example, an educational researcher interested in predicting high school graduates' choices for further education would probably include as many measures of personality, achievement motivation, academic performance, etc as possible in order to learn which one(s) offer the best prediction.

Model

Put another way, we want to build a "model" of how we can best predict to which group a case belongs. In the following discussion we will use the term "in the model" in order to refer to variables that are included in the prediction of group membership, and we will refer to variables as being "not in the model" if they are not included.

Forward stepwise analysis

In stepwise discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. The most contributing variable at that step is then included in the model (whence the name *forward analysis*) and the process starts again.

Backward stepwise analysis

One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful discriminant function analysis, one would only keep the "important" variables in the model, that is, those variables that contribute the most to the discrimination between groups.

F to enter, F to remove

The stepwise procedure is "guided" by the respective "F to enter" (for the *forward analysis*) and "F to remove" (for the *backward analysis*) values. The F value for a variable indicates its statistical significance in the discrimination between groups, that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership.

6. Conclusion

In general, discriminant analysis is a very useful tool for:

1. detecting the variables that allow the researcher to discriminate between different (naturally occurring) groups
2. classifying cases into different groups with a "better-than-chance" accuracy.

2.3.4 The segmentation method

1. Description

Decision-tree learning is a common method used for the segmentation method. A decision tree describes a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications. A decision tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. Splitting is done thanks to the computation of a distance measure. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the derived subset. Decision tree can also be described as the synergy of mathematical and computing techniques that aids on the description, categorisation and generalisation of a given set of data. Data comes in records of the form:

$$(x, y) = (x_1, x_2, x_3, ..., x_k, y).$$

The dependent variable, y, is the variable that we are trying to understand, classify or generalise. The other variables x_1, x_2, x_3 *etc* are the variables (the predictor attributes) that will help us on that job.

A decision tree has two other names:

1. *Regression tree*

   The regression tree approximate real-valued functions instead of being used for classification tasks (e.g. estimate the price of a house or a patient's length of stay in a hospital).

2. *Classification tree*

   If the y is a categorical variable like sex (male or female) or the result of a game (lose or win), we will use the term *classification tree.*

To fully understand the method, let's apply it on a small example[10]). Our friend David is the manager of a famous golf club. Sadly, he is having some trouble with his customers attendance. There are days that everyone wants to play golf and the staff of the club is not enough for them; on some other days for no apparent reason, no one plays golf; and the club has a high slack of employees. David's objective is to optimise the staff availability by trying to predict when people will play golf using the week forecast. To accomplish that, he needs to understand the reason why people decide not to play and if there is any explanation for that. So during two weeks he has been recording the following variables:

- the *Outlook*

  Was it sunny, clouded or raining on that day?

- the $Temperature$ in degrees Fahrenheit.

- the relative $Humidity$ in percent.

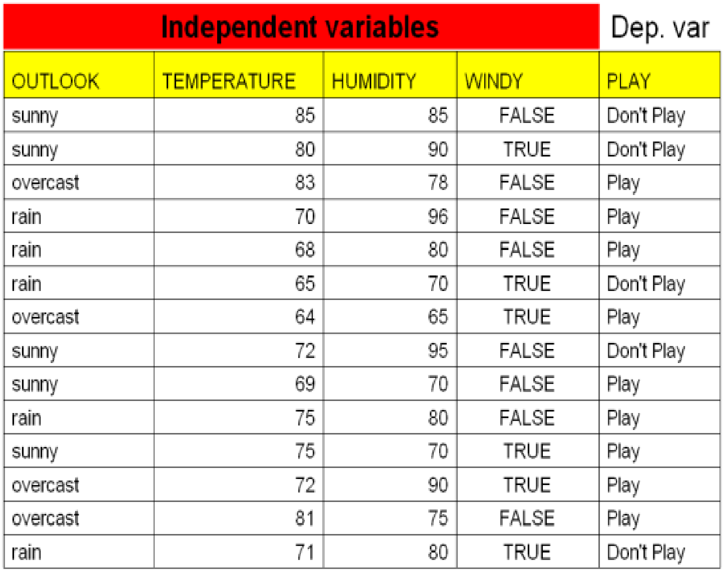
- whether it was $Windy$ or not.

- if people did *Play* on that day

He ended with the dataset of figure 2.32. The decision tree model of figure 2.33 is then proposed to solve David's problem. As you can see on the latter, the decision tree is a directed, acyclic graph in form of a tree. The top node represents all the data. The classification tree algorithm finds out that the best way to explain the dependent variable, *Play*, is by using the variable *Outlook*. Using the categories of the variable *Outlook*, three different groups were found: the group that plays golf when it is sunny, the group that plays when it is clouded and surprisingly we realise that when it is raining some people do play golf! Our first conclusion is: if the outlook is overcast, people always play golf and there are some fanatical people who play golf even in the rain. Then again we divide the sunny group in two groups. We realise that customers don't like to play golf if the humidity is higher than seventy percent. Finally we divide the rain category in two and find out that customers will not play golf if it is windy.

The short solution of the problem given by the classification-tree software is the following. David dismisses most of the staff on days that are sunny and humid or on rainy days that are windy because almost no one is going to play golf on those days. On the other days, when a lot of people will play golf, he can hire some temporary staff to help on the job. The conclusion is that decision tree helped us turn a complex data representation into a much easier structure (parsimonious).

[10] This example comes from `http://en.wikipedia.org`



Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

Figure 2.32: Play golf dataset

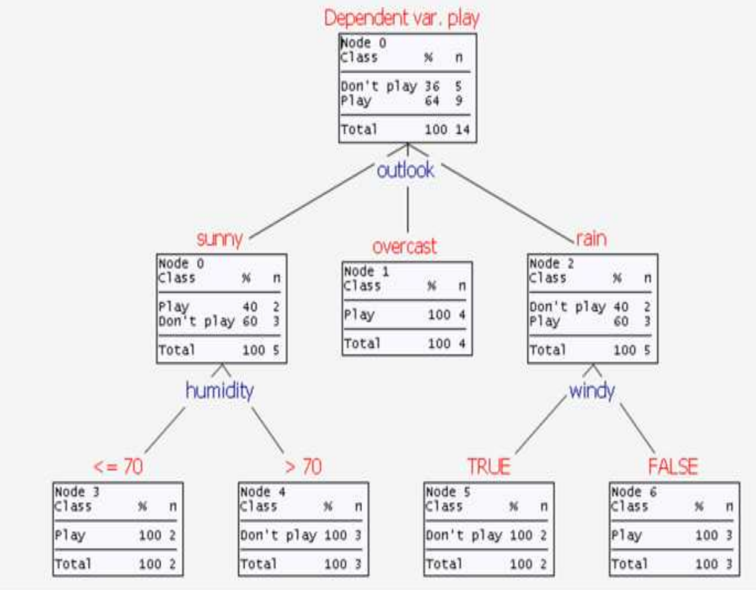


Figure 2.33: Decision tree model

Classification-tree analysis and discriminant analysis: the same method?

In order to understand the difference existing between both methods, let's consider the following example given by Breiman et al. (1984). When heart attack patients are admitted to a hospital, dozens of tests are often performed to obtain physiological measures such as heart rate, blood pressure, and so on. A wide variety of other information is also obtained, such as the patient's age and medical history. Patients subsequently can be tracked to see if they survive the heart attack, say, at least 30 days. It would be useful in developing treatments for heart attack patients, and in advancing medical theory on heart failure, if measurements taken soon after hospital admission could be used to identify high-risk patients (those who are not likely to survive at least 30 days).

One classification tree that Breiman et al. (1984) developed to address this problem was a simple, three-question decision tree. Verbally, the binary classification tree can be described by the statement: "If the patient's minimum systolic blood pressure over the initial 24 hour period is greater than 91, then if the patient's age is over 62.5 years, then if the patient displays sinus tachycardia, then and only then the patient is predicted not to survive for at least 30 days". It is easy to conjure up the image of a decision tree from such a statement. A hierarchy of questions are asked and the final decision that is made depends on the answers to all the previous questions. Similarly, the relationship of a leaf to the tree on which it grows can be described by the hierarchy of splits of branches (starting from the trunk) leading to the last branch from which the leaf hangs.

The hierarchical nature of classification trees is illustrated by a comparison to the decision-making procedure employed in low-risk discriminant analysis. A traditional linear discriminant analysis of the heart attack data would produce a set of coefficients defining the single linear combination of blood pressure, patient age, and sinus tachycardia measurements that best differentiate low-risk from high-risk patients. A score for each patient on the linear discriminant function would be computed as a composite of each patient's measurements on the three predictor variables, weighted by the respective discriminant function coefficients.

The predicted classification of each patient as a low-risk or a high-risk patient would be made by simultaneously considering the patient's scores on the three-predictor variables. Suppose

- P (minimum systolic blood $Pressure$ over the 24 hour period),
- A (Age in years),
- T (presence of sinus $Tachycardia$: 0 = not present; 1 = present)

are the predictor variables. Suppose p, a and t are the corresponding linear discriminant function coefficients, and c the "*cut* point" on the discriminant function for separating the two classes of heart attack patients. The decision equation for each patient would be of the form,

"if $pP + aA + tT$ - c is less than or equal to zero, the patient is low-risk, else the patient is high-risk".

In comparison, the decision tree developed by Breiman et al. (1984) would have the following hierarchical form, where p, a, and t would be -91, -62.5, and 0, respectively:

"If $p + P$ is less than or equal to zero, the patient is low-risk, else if $a + A$ is less than or equal to zero, the patient is low-risk, else if $t + T$ is less than or equal to zero, the patient is low-risk, else the patient is high-risk".

Superficially, the discriminant analysis and classification tree decision processes might appear similar, because both involve coefficients and decision equations. But the difference of the simultaneous decisions of discriminant analysis from the hierarchical decisions of classification trees cannot be emphasized enough. The distinction between the two approaches can perhaps be made most clear by considering how each analysis would be performed in *Regression*. Because risk in the example of Breiman et al. (1984) is a dichotomous dependent variable, the discriminant analysis predictions could be reproduced by a simultaneous multiple regression of risk on the three-predictor variables for all patients. The classification tree predictions could only be reproduced by three separate simple regression analyses, where risk is first regressed on P for all patients, then risk is regressed on A for patients not classified as low-risk in the first regression, and finally, risk is regressed on T for patients not classified as low-risk in the second regression. This clearly illustrates the simultaneous nature of discriminant analysis decisions as compared to the recursive, hierarchical nature of classification tree decisions.

2.3.5 The multiple correspondence analysis

In order to better understand this analysis, let's first have a look at other kinds of analyses.

1. The factorial analysis

The goal of such an analysis is to summarize and organize the information into a hierarchy, information which can be found in a matrix of *n rows* (the subjects) and *p columns* (the variables). The *n subjects* are described by a cloud of *p variables.* The information represented by this cloud is the dispersion of the n points. So, computing a summary of this information means projecting these points into a space which dimension is below p. The axes of this subspace are called "factors". Each variable p carries a part of original information and a part of information which is redundant with the other variables. The factorial summary will group this redundant information. Each factor is then the linear combination of the p variables. An a coefficient is associated with each variable. This a coefficient is proportional to the strength of the links between the variable and the factor. Since the factors are organized into a hierarchy, the first axis contains the maximum of information. This axis has the greatest dimension of the cloud. It is the best summary in a one-dimension space. But there still remains residue from information. The second axis contains the maximum of the remaining information and is orthogonal to the first one (by construction). This second axis is also the axis of greatest residual dimension of the cloud. The first and second axes form the best summary in a two-dimension space. But there is still residue. The third axis contains less information than the first two axes and is orthogonal to the first two axes (by construction as well). And so on.

2. The factorial analysis of the correspondences

This test can be applied on a table of *n subjects* and *p qualitative variables* (modality 0 for its absence, modality 1 for its presence). We can go one step further by carrying out what we call *a multiple correspondence analysis* on this table. The main goal of this method is to organize the information into a hierarchy as well. Unlike the *principal correspondence analysis*, the computation uses the χ^2 distance instead of the euclidian distance. Why don't we use the euclidian distance? Simply because the euclidian distance translates the mass differences between the subjects. To remove the mass effect, we can weight the distances. But then, the question is: “what is a weighted distance?”. Actually, it is an euclidian distance between the profiles (rows) of the subjects for which each term is weighted by the inverse of the relative weight of the corresponding variable (column). In the distance computations, this weighting strengthens the weight of low-mass variables and compensates the weight differences between variables. The formulation is the following:

$$d^2(i,i') = \sum_{j=1}^{p} \frac{\left[\frac{k_{ij}}{k_{i.}} - \frac{k_{i'j}}{k_{i'.}}\right]^2}{\frac{k_{.j}}{k_{..}}}$$

	j = 1	**2**	**.**	**j**	**j’**	**.**	**p**	Sum
i=1	k_{11}	k_{12}						$k_{1.}$
2	k_{21}							$k_{2.}$
.								
i				k_{ij}	$k_{ij'}$			$k_{i.}$
i’				$k_{i'j}$				$k_{i'.}$
.								
n								$k_{n.}$
Sum	$k_{.1}$	$k_{.2}$		$k_{.j}$	$k_{.j'}$			$k_{..}$

This weighted distance is called the χ^2 distance. As a result of the weighting symmetry applied to the contingency table (table which sum on the rows or on the columns has a meaning) by the χ^2 measure, this weighted distance can be applied to the rows (the subjects) as well as to the columns (the variables):

$$d^2(j,j') = \sum_{i=1}^{n} \frac{\left[\frac{k_{ij}}{k_{.j}} - \frac{k_{ij'}}{k_{.j'}}\right]^2}{\frac{k_{i.}}{k_{..}}}.$$

Each factorial axis is determined by a vector called **“eigenvector”**. The eigenvectors determine the different directions of the information cloud. The information part taken by each eigenvector is called the **“eigenvalue”**. It defines the hierarchy of the factorial axis. For the factorial analysis of the correspondences, the first eigenvalue is trivial and equals 1. Therefore it doesn’t help interprete. All the other eigenvalues are below 1. The greatest represents the variance of the first factorial axis. The second represents the variance of the second factorial axis and so on. The sum of all eigenvalues measures the total inertia of the cloud. The greater the differenciation degree involved by an axis, the greater the “eigenvalue”. As for the principal components analysis (description in chapter 3), the eigenvalue represents the variance part (of information) of the cloud taken into account by the axis. It is generally expressed with percentages.

The **contributions** to the factorial axes allow to figure out which variables (or subjects) are the most contributive for each axis. The sum of the contributions equals 1. They allow to identify the variables (or the subjects) which best define the axis. Unlike the principal components analysis, two variables (or subjects) can be projected at the same place along the axis, thus having the same coordinates but without having the same contributions since the mass of the variables (or the subjects) is taken into account during the computation of the contributions for the factorial analysis of the correspondences (see χ^2 measure).

The **quality** of representation on a factorial axis allows to characterize the variables (or the subjects) by the axes. It measures the distance between a variable (or a subject) and the center of gravity taken into account by an axis. The quality equals the cos^2 of the angle between the vector of the variable (or the subject) and the axis. $Cos^2(0°) = 1$ means the variable (or the subject) is on the axis. Therefore the variable is perfectly described by the axis. $Cos^2(90°) = 0$ means the variable (or the subject) is perpendicular to the axis. Thus, the variable is not described by the axis at all.

3. The notion of dispersion

As you have probably guessed, a very important notion in the context of the correspondence analysis is the concept of *dispersion* also known as *inertia.* To fully understand this notion, let's first define a few more concepts.

Primitive matrix

The original data matrix N(I,J), or contingency table, is called the primitive matrix or primitive table. The elements of this matrix are denoted by n_{ij}.

Profiles

While interpreting a cross-tabulation, it makes little sense to compare the actual frequencies in each cell. Each row and each column have a different number of respondents, called the *base* of respondents. For comparison, it is essential to reduce either the rows or columns to the same base.

Let's consider a contingency table N(I,J) with I rows (i=1, 2, ..., I) and J columns (j =1, 2, ... , J) having frequencies n_{ij}. Marginal frequencies are denoted by n_{i+} and n_{+j}:

$$n_{i+} = \sum_j n_{ij}$$

$$n_{+j} = \sum_i n_{ij}$$

Total frequency is given by:

$$n = \sum_j \sum_i n_{ij}$$

Row profiles

The profile of each row i is a vector of conditional densities:

$$profile_i = \frac{n_{ij}}{n_{i+}}$$

for $j = 1, 2, \ldots, J$. The complete set of the row profiles may be denoted by a matrix we will call $R(I\ x\ J)$ and is displayed in table 2.11.

Rows	Columns					Total
	1	2	3		J	
1.	n_{11}/n_{1+}	n_{12}/n_{1+}	n_{13}/n_{1+}	....	n_{1j}/n_{1+}	1
2.	n_{21}/n_{2+}	n_{22}/n_{2+}	n_{23}/n_{2+}	....	n_{2j}/n_{2+}	1
3.	n_{31}/n_{3+}	n_{32}/n_{3+}	n_{33}/n_{3+}	....	n_{3j}/n_{3+}	1
...	...	...	...	....	...	1
I	n_{i1}/n_{i+}	n_{i2}/n_{i+}	n_{i3}/n_{i+}	....	n_{ij}/n_{i+}	1
Column mass	n_{+1}/n_{++}	n_{+2}/n_{++}	n_{+3}/n_{++}	....	n_{+j}/n_{++}	1

Table 2.11: Matrix of row profiles

Column Profiles

The profile of each column j is a vector of conditional densities n_{ij}/n_{+j} for $i = 1, 2, \ldots, I$. The complete set of the column profiles may be denoted by a matrix we will call $C(I\ x\ J)$ and is displayed in table 2.12.

Rows	Columns					Row Mass
	1	2	3		J	
1.	n_{11}/n_{+1}	n_{12}/n_{+2}	n_{13}/n_{+3}	....	n_{1j}/n_{+j}	n_{+1}/n_{++}
2.	n_{21}/n_{+1}	n_{22}/n_{+2}	n_{23}/n_{+3}	....	n_{2j}/n_{+j}	n_{+2}/n_{++}
3.	n_{31}/n_{+1}	n_{32}/n_{+2}	n_{33}/n_{+3}	....	n_{3j}/n_{+j}	n_{+3}/n_{++}
...	...	...	...	....	...	...
I	n_{i1}/n_{+1}	n_{i2}/n_{+2}	n_{i3}/n_{+3}	....	n_{ij}/n_{+j}	n_{+i}/n_{++}
Total	1	1	1	....	1	1

Table 2.12: Matrix of column profiles

The average row profile is computed according to the following formula:

$$\overline{r} = \frac{n_{+j}}{N}$$

for $j = 1, 2, ..., J$.

The average column profile is computed according to the following formula:

$$\overline{c} = \frac{n_{i+}}{N}$$

for $i = 1, 2, ..., I$.

Masses

Another fundamental concept in correspondence analysis is the concept of mass. The mass of the i^{th} row equals:

Marginal frequency of the i^{th} row/total

that is,

$$\frac{n_{i+}}{n}.$$

Similarly the mass of the j^{th} column equals:

Marginal frequency of the j^{th} column/total

that is,

$$\frac{n_{j+}}{n}.$$

Correspondence Matrix

The correspondence matrix P is defined as the original table N divided by the total n, $P = (1/n)N$. Thus, each cell of the correspondence matrix is given by the cell frequency divided by the total. The correspondence matrix shows how one unit of *mass* is distributed across the cells. The row and column totals of the correspondence matrix are the row mass and column mass, respectively.

Clouds of Points N(I) and N(J)

The cloud of points $N(I)$ is the set of elements of points i, which coordinates are the components of the profile and which mass is $\frac{n_{+i}}{n_{++}}$. The cloud of points $N(J)$ is the set of elements of points j, which coordinates are the components of the profile and which mass is $\frac{n_{j+}}{n_{++}}$.

Inertia

Inertia is a term borrowed from the "moment of inertia" in mechanics. A physical object has a center of gravity (or centroid). Every particle of the object has a certain mass m and a certain distance d from the centroid. The moment of inertia of the object is the quantity md^2 summed over all the particles that constitute the object:

$$Moment\ of\ inertia\ = \sum md^2.$$

This concept has an analogy in correspondence analysis. There is a cloud of profile points with masses adding up to 1. These points have a centroid (i.e., the average profile) and a distance (Chi-square distance) between profile points. Each profile point contributes to the inertia of the whole cloud. The inertia of a profile point can be computed by the following formula:

$$\text{For the } i^{th} \text{ row profile, } inertia\ = \ m_i \textstyle\sum_j \frac{(r_{ij} - \overline{r}_j)^2}{\overline{r}_j}.$$

where r_{ij} is the ratio $\frac{n_w}{n_{i+}}$ and $\overline{r}_j$ is $\frac{n_{.j}}{n}$. The inertia of the j^{th} column profile is computed similarly. The total inertia of the contingency table is given by:

$$Total\ inertia\ = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

which is the Chi-square statistic divided by n.

2.3.6 The analysis

1. The Fisher's test

Here you will find the details of the test for a couple of binary variables. For the remaining variables, please see appendix E. Before the analysis, let's consider the cell situated at the intersection between the first row and the first column, denoted by $C(1,1)$, being A, $C(1,2)$ being B, $C(2,1)$ being C, $C(2,2)$ being D and H_0 stating both populations are homogeneous.

A. Description of some results

Self-description	yes	no	Total
M	11 (A)	4 (B)	15
F	10 (C)	5 (D)	15
Total	21	9	30

According to Fisher's tables (see appendix D), for A + B = 15, C + D = 15 and A = 11, the maximum value for C (above which we **cannot reject** H_0) is 5. Since our C (the one in the table) equals 10, we **cannot reject** H_0 with an error of 5 %.

Personal pages (ratio)	yes	no	total
M	7	8	15
F	5	10	15
Total	12	18	30

Since C is above 1, we **cannot reject** the homogeneity hypothesis at 0,05.

Focus on credentials	yes	no	total
M	4	11	15
F	6	9	15
Total	10	20	30

Since C is above 0, we **cannot reject** the homogeneity hypothesis at 0,05.

Reddish colours (txt)	yes	no	total
M	10	5	15
F	5	10	15
total	15	15	30

Since C is above 4, we **cannot reject** the homogeneity hypothesis at 0,05.

Blueish colours (txt)	yes	no	total
M	15	0	15
F	13	2	15
total	28	2	30

Since C is above 11, we **cannot reject** the homogeneity hypothesis at 0,05.

Graphic accents	yes	no	total
M	0	15	15
F	5	10	15
total	5	25	30

Since D is below 11, we **can reject** the homogeneity hypothesis at 0,05.

Photos	yes	no	total
M	9	6	15
F	2	13	15
total	11	19	30

Since C is below 3, we **can reject** the homogeneity hypothesis at 0,05.

B. Summary of the results

Only the variables “graphic accents” and “photos” allow to reject the homogeneity hypothesis. Thus for these variables, the males and the females have different behaviours: women use more graphic accents than men and men put more photos on their websites than women.

2. The binomial test

Since we can’t see any difference apart from the graphic accents and the photos between genders, let’s now carry out a binomial test to see if there are any genderless differences of behaviour. Using the binomial tables for a thirty-subject sample, we will work out the genderless proportions for each variable. To see the details of the test, please read appendix E. In table 2.13 you will find the variables for which the academics' behaviour can be determined. For example, regarding the self-description, since the confidence interval is [0,51 ; 0,7], we can state the academics describe themselves on the Internet.

Pr[Having a self-description]	[0,51 ; 0,7]
Pr[Not having a private description]	[0,65 ; 0,83]
Pr[Not having a technological website]	[0,62 ; 0,8]
Pr[Having classic fonts]	[0,84 ; 0,96]
Pr[Not having girlish fonts]	[0,84 ; 0,96]
Pr[Using reddish colours for (hyper)text]	[0,32 ; 0,5]
Pr[Using blueish colours for (hyper)text]	[0,78 ; 0,93]
Pr[Using black for (hyper)text]	[0,69 ; 0,86]
Pr[Not using white for (hyper)text]	[0,65 ; 0,83]
Pr[Not using grey for (hyper)text]	[0,69 ; 0,86]
Pr[Not showing graphic accents]	[0,65 ; 0,83]
Pr[Including links to other people's pages]	[0,57 ; 0,73]
Pr[Including non-personal links]	[0,84 ; 0,96]
Pr[Including graphics]	[0,65 ; 0,83]
Pr[Not having a dark background]	[0,65 ; 0,83]
Pr[Not having a blueish background]	[0,58 ; 0,76]
Pr[Not having a black background]	[0,69 ; 0,86]
Pr[Having a white background]	[0,58 ; 0,76]
Pr[Having a classic background]	[0,84 ; 0,96]
Pr[Not having an original background]	[0,65 ; 0,83]

Table 2.13: Binomial results

3. The discriminant analysis

As said previously, the goal is to find the variables among all the binary variables which best discriminate/separate the men from the women of our sample. Below, you will find four tables consisting of the results of the analysis.

Variables in the analysis		
Step	**Variable**	**Wilks' Lambda**
1	photos (yes/no)	/
2	photos (yes/no) graphic accents (yes/no)	0,800 0,766

At step 1, the variable representing the presence of photos was entered. At step 2, the presence of graphic accents was taken into account for the discrimination. As you can read in the table, the Wilks' Lambda value (which is a statistical criteria that is used to add or remove variables from the analysis) for the graphic accents is lower than for the photos. This means the variable "graphic accents" better separates the men from the women than the variable "photos".

Canonical discriminant function coefficients	
	Function 1
graphic accents (yes/no)	-2,041
photos (yes/no)	1,759
(constant)	-0,305

We can rewrite this function as

$$F(x) = -2,041g + 1,759p - 0,305$$

with g standing for "graphic accents" and p for "photos". Since the absolute value of the coefficient of the graphic accents is greater than the one for the photos, it confirms the fact graphic accents is the best discriminating variable.

Classification Function Coefficients		
	Gender 0 (F)	**Gender 1 (M)**
graphic accents (yes/no)	2,747	-0,317
photos (yes/no)	0,528	3,170
(constant)	-1,186	-1,644

We can rewrite these functions as follows:

$$F_0(x) = 2,747g + 0,528p - 1,186$$

and

$$F_1(x) = -0,317g + 3,170p - 1,644.$$

We clearly have a much higher coefficient for the females regarding the graphic accents, meaning these use more graphic accents than the males. But regarding the photos, we have the contrary. So, if we have an extra subject we want to classify as a male or a female, we will run both functions. The one giving the highest score will determine the subject's profile. For example, if $F_0(x)$ gives a higher result than $F_1(x)$, the subject will be considered as a female.

Classification results				
		PGM 0 (F)	**PGM 1 (M)**	**Total**
Original count	0 (F)	14	1	15
Original count	1 (M)	6	9	15
%	0 (F)	93,3	6,7	100
%	1 (M)	40	60	100

Let's notice 76,7% of original grouped cases are correctly classified. The classification results are computed on the basis of function $F_0(x)$ for the females and $F_1(x)$ for the males. Both functions are run on each subject. The function giving the highest score is the one allowing to know the subject's profile. For example, if $subject_i$ is a male, and that $F_0(x)$ gives the highest score, he will be assigned a female profile. This last table also shows there is one female with a profile corresponding to the male group whereas there are six males whose profiles correspond to the female group.

4. The segmentation tree

In order to know if we can get better results than the ones from the discriminant analysis, let's carry out a segmentation test. This one will show us the variables chosen in order to partition the men and the women and will give us new subsets of people. The segmentation test has been carried out by choosing the **entropy reduction**. So, you will be first introduced to this method before viewing the results.

1. The entropy reduction

The basic decision tree learning algorithm

Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees. In this section we present the underlying principles of this basic algorithm for decision tree learning. Our basic algorithm learns decision trees by constructing them top-down, beginning with the question "which attribute should be tested at the root of the tree?"

To answer this question, each instance attribute is evaluated using a statistical test to determine how well it classifies the training examples on its own. The best attribute is selected and used as the test to carry out at the root node of the tree. A descendant of the root note is then created for each possible value of this attribute, and the training examples are sorted to the appropriate descendant node (i.e., down the branch corresponding to the example's value for this attribute). The entire process is then repeated using the training examples associated with each descendant node to select the best attribute to test at that point of the tree. This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices.

Which attribute is the best classifier?

The central choice in the basic algorithm is selecting which attribute to test at each node of the tree. We would like to select the attribute that is most useful for classifying examples. We will define a statistical property, called $information\ gain$, that measures how well a given attribute separates the training examples according to their target classification. Our basic algorithm uses this information gain measure to select among the candidate attributes at each step while growing the tree.

Entropy measures homogeneity of examples

In order to define $information\ gain$ precisely, we begin by defining a measure commonly used in information theory, called *entropy*, that characterizes the (im)purity of an arbitrary collection of examples. Given a collection S, containing positive and negative examples of some target concept, the $entropy\ of\ S$ relative to this Boolean classification is:

$$Entropy(S) \equiv -p^{+}\ log_2 p^{+} - p^{-}\ log_2 p^{-} \quad (3.1)$$

where p^+ is the proportion of positive examples in S and p^- is the proportion of negative examples in S. In all calculations involving *entropy* we define 0 $log0$ to be 0.

To illustrate, suppose S is a collection of 14 examples (we adopt the notation [9+, 5-] to summarize such a sample of data). Then the *entropy* of S relative to this Boolean classification is:

$$Entropy([9+, 5-]) = -(9/14)\ log_2(9/14) - (5/14)\log_2(5/14) = 0,940 \quad (3.2)$$

Notice that the *entropy* is 0 if all members of S belong to the same class. For example, if all members are positive ($p^+ = 1$), then p^- is 0, and we thus have:

$$Entropy(S) = -1\ log_2(1) - 0\ log_2(0) = -1 * 0 - 0 * log_2(0) = 0.$$

Note the *entropy* is 1 when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the *entropy* is between 0 et 1. The figure below shows the form of the $entropy\ function$ relative to a Boolean classification, as p^+ varies between 0 and 1.

One interpretation of *entropy* from information theory is that it specifies the minimum number of bits of information needed to encode the classification of an arbitrary member of S (i.e., a member of S drawn at random with uniform probability). For example, if p^+ is 1, the receiver knows the drawn example will be positive, so no message needs to be sent, and the *entropy* is zero. On the other hand, if p^+ is 0,5, one bit is required to indicate whether the drawn is positive or negative. If p^+ is 0,8, then a collection of messages can be encoded using on average less than 1 bit per message by assigning shorter codes to collections of positive examples and longer codes to less likely negative examples. Thus far we have discussed *entropy* in the special case where the target classification is Boolean. More generally, if the target attribute can take on c different values, then the *entropy* of S relative to the c-wise classification is defined as

$$Entropy(S) \ \equiv \ \sum_{i=1}^{c} -p_i \log_2 p_i \quad (3.3)$$

where p_i is the proportion of S belonging to class i. Note the logarithm is still *base* 2 because *entropy* is a measure of the expected encoding length measured in *bits*. Note also that if the target attribute cans take on c possible values, the *entropy* can be as large as $log_2(c)$.

Information gain measures the expected reduction in entropy

Given *entropy* as a measure of the impurity in a collection of training examples, we can now define a measure of the effectiveness of an attribute in classifying the training data. The measure we will use, called *information gain*, is simply the expected reduction in *entropy* caused by partitioning the examples according to this attribute. More precisely, the information gain, $Gain(S, A)$ of an attribute A, relative to a collection of examples S, is defined as:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (3.4)$$

where $Values(A)$ is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S | A(s) = v\}$). Note the first term in Equation (3.4) is just the *entropy* of the original collection S, and the second term is the expected value of the *entropy* after S is partitioned using attribute A. The expected *entropy* described by this second term is simply the sum of the entropies of each subset S_v, weighted by the fraction of examples $\frac{|S_v|}{|S|}$ that belong to S_v. $Gain(S, A)$ is therefore the expected reduction in *entropy* caused by knowing the value of attribute A. Put another way, $Gain(S, A)$ is the information provided about the *target function value*, given the value of some other attribute A. The value of $Gain(S, A)$ is the number of bits saved when encoding the *target value* of an arbitrary member of S, by knowing the value of attribute A.

For example suppose S is a collection of training-example days described by attributes including $Wind$, which can have the values $Weak$ or $Strong$. As before, assume S is a collection containing 14 examples, [9+, 5-]. Of these 14 examples, suppose 6 of the positive and 2 of the negative examples have $Wind = Weak$, and the remainders have $Wind = Strong$. The information gain due to sorting the original 14 examples by attribute $Wind$ may then be calculated as

$$\begin{aligned} Values(Wind) &= Weak, Strong \\ S &= [9+, 5-] \\ S_{Weak} &\leftarrow [6+, 2-] \\ S_{Strong} &\leftarrow [3+, 3-] \end{aligned}$$

$$\begin{aligned} Gain(S, Wind) &= Entropy(S) - \textstyle\sum_{v \in \{Weak,\ Strong\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S) - (8/14)\ Entropy(S_{Weak}) - (6/14)\ Entropy(S_{Strong}) \\ &= 0,940 - (8/14)0,811 - (6/14)1,00 \\ &= 0,048 \end{aligned}$$

Information gain is precisely the measure used by our basic algorithm to select the best attribute at each step in growing the tree. The use of information gain to evaluate the relevance of attributes is summarized in the figure below.

In this figure, the information gain of two different attributes, $Humidity$ and $Wind$, is computed in order to determine which is the better attribute for classifying the training examples shown in table 2.14. From this figure, we thus learn $Humidity$ provides greater information gain than $Wind$, relative to the target classification. Here, E stands for $Entropy$ and S for the original collection of examples. Given an initial collection S of 9 positive and 5 negative examples, [9+,5-], sorting these by their $Humidity$ produces collections of [3+,4-]($Humidity = High$) and [6+,1-]($Humidity = Normal$). The information gained by this partitioning is 0,151, compared to a gain of only 0,048 for the attributed $Wind$.

An Illustrative Example

To illustrate the operation of our basic algorithm, consider the learning task represented by the training examples of table 2.14. Here the target attribute *PlayTennis*, which can have values *yes* or *no* for different Saturday mornings, is to be predicted based on other at-

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Table 2.14: Training examples for the target concept *PlayTennis*

tributes of the morning in question. Consider the first step through the algorithm, in which the topmost node of the decision tree is created. Which attribute should be tested first in the tree? Our basic algorithm determines the information gain for each candidate attribute (i.e., *Outlook*, *Temperature*, *Humidity*, and *Wind*), then selects the one with highest information gain. The computation of information gain for two of these attributes is shown in the figure of page 80.

The information gain values for all four attributes are

$$Gain(S, Outlook) = 0,246$$

$$Gain(S, Humidity) = 0,151$$

$$Gain(S, Wind) = 0,048$$

$$Gain(S, Temperature) = 0,029$$

where S denotes the collection of training examples from table 2.14.

According to the information gain measure, the *Outlook* attribute provides the best prediction of the target attribute, *PlayTennis*, over the training examples. Therefore, *Outlook* is selected as the decision attribute for the root node, and branches are created below the root for each of its possible values (i.e., *Sunny*, *Overcast*, and *Rain*). The resulting partial decision tree is shown in the figure below, along with the training examples sorted to each new descendant node.

Note that every example for which $Outlook = Overcast$ is also a positive example of $PlayTennis$. Therefore, this node of the tree becomes a leaf node with the classification $PlayTennis = Yes$. In contrast, the descendants corresponding to $Outlook = Sunny$ and $Outlook = Rain$ still have nonzero *entropy*, and the decision tree will be further elaborated below these nodes.

The process of selecting a new attribute and partitioning the training example is now repeated for each nonterminal descendant node, this time using only the training examples associated with that node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either of two conditions is met:

1. every attribute has already been included along this path through the tree,

2. or the training examples associated with this leaf node all have the same target attribute value (i.e., their *entropy* is zero).

The figure above illustrates the computations of information gain for the next step in growing the decision tree. The final decision tree learned by our basic algorithm from the 14 training examples of table 2.14 is shown in the figure below.

2. The results

Let's remember the goal of this analysis is to divide the males and the females of our sample into groups which are as homogeneous as possible. In our context, the set of attributes is composed of the binary variables. As you can see in figure 2.34, the first discriminant variable is the **presence of graphic accents**. The group using these only consists of women. The group who does not use them consists of more men than women. The latter is then divided into two groups: the one **showing photos** and the one **showing no photos**. The first division only consists of men (apart from one woman). The second division consists of more women than men. This second group is divided again into two groups: the academics **using comics graphics** and the other ones. What we are interested in is the group of academics who do not use comics graphics for which we have more women than men. We can summarize the description of the tree with table 2.15.

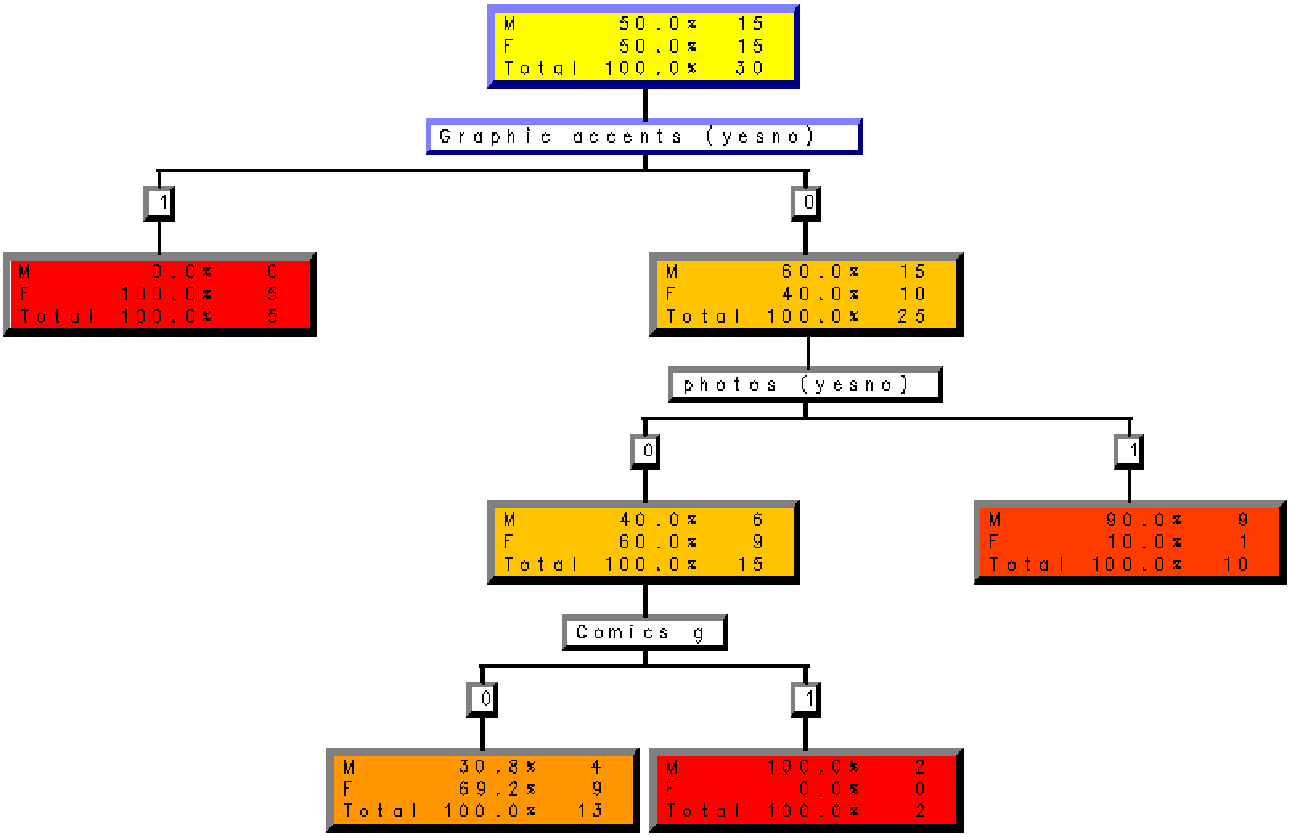


Figure 2.34: Tree using the entropy reduction method

Type of group	**Feature**
Exclusively women	Using graphic accents
More women than men	Using no graphic accents and showing no photos and no comics graphics
Almost exlusively men	Using no graphic accents and showing photos

Table 2.15: Summary of results for the entropy reduction

2.3.7 Conclusions for the binary variables

Let's first remember all the hypotheses we had formulated regarding the binary variables.

1. **Men and women both use classic fonts**

2. **Men and women don't use girlish fonts**

3. **Men and women differ in the type of colours used for text and hypertext**

   - Women tend to use more reddish colours

   - Men tend to use more blueish colours

- Both use black
- Men tend to use white more than women
- Men tend to use grey more than women

4. **Men tend to have more technological websites**

5. **Women and men differ regarding the type of colours used for their backgrounds**

    - Women use more soft colours
    - Men use more dark colours
    - Women use more reddish colours
    - Men use more blueish colours
    - Men use black more often
    - Men and women do not differ regarding the use of white
    - Men use grey more often

6. **Men and women differ regarding the type of background**

    - There are more women's websites with a classic background compared with men's sites.
    - There are more men's websites with an original background compared with women's sites.

7. **There are more men describing themselves on their websites than women**

8. **Men and women differ in the way they describe themselves**

    - Men tend to describe themselves in a private way
    - Women tend to describe themselves in a professional way

9. **The number of male academics having personal content in their website is greater than the number of female academics**

10. **Men and women do not differ regarding the focus on credentials**

11. **There are more women using graphic accents than men**

12. **Men and women differ in the type of links they have on their websites**

    - Men's websites include more links to non-personal pages
    - Women's websites include more links to personal pages

13. **There are more men showing self-photos than women**
14. **Women and men differ regarding the type of self-photos**
    - Women and men both show the official picture
    - Men show more non-official pictures of themselves than women
    - Men show more family pictures with themselves than women
    - Men show more pictures of themselves with friends than women
    - Men and women do not differ regarding pictures of themselves with colleagues
    - Men and women do not differ regarding pictures of themselves with their pets
    - Men show more pictures of themselves in their leisure time than women
    - Men show more computer-related pictures with themselves
15. **Men and women both show good-quality self-photos**
16. **There are more men showing photos apart from self-photos than women**
17. **Women and men differ regarding the type of photos**
    - Men show more pictures of their families than women
    - Men and women do not differ regarding pictures showing their friends
    - Women show more pictures of their colleagues than men
    - Men show more pictures of their pets than women
    - Men show more pictures of their leisure time than women
    - Men show more computer-related pictures than women
18. **Women and men both show good-quality pictures**
19. **The number of females using graphics is greater than the number of males**
20. **Men and women differ in the type of graphics**
    - Women use more basic graphics
    - Women use more modern graphics
    - Women use trendier graphics
    - Women use more artistic graphics
    - Men use more comics graphics
    - Women and men do not differ regarding the use of computer-related graphics

Most hypotheses we had formulated at the beginning cannot be validated, in terms of differences between men and women. But of course, this can change with a larger sample. The only characteristics for which we have a gendered difference in our context are the *presence of graphic accents*, which use corresponds to a female profile, and the *presence of photos*, with men putting up more photos on their websites. However, we can take advantage of our observations to highlight the most common behaviours regarding the layout of the websites:

- Between 84 % and 96 % of the academics use classic fonts
- Between 84 % and 96 % do not use girlish fonts
- Between 84 % and 96 % include non-personal links in their websites
- Between 84 % and 96 % use a classic background
- Between 78 % and 93 % use blueish colours for (hyper)text
- Between 69 % and 86 % use black for (hyper)text
- Between 69 % and 86 % do not use grey for (hyper)text
- Between 69 % and 86 % do not have a black background
- Between 65 % and 83 % do not describe themselves in a private way
- Between 65 % and 83 % do not use white for (hyper)text
- Between 65 % and 83 % do not show graphic accents
- Between 65 % and 83 % include graphics in their websites
- Between 65 % and 83 % do not have a dark background
- Between 65 % and 83 % do not have an original background
- Between 62 % and 80 % do not have a technological website
- Between 58 % and 76 % do not have a blueish background
- Between 58 % and 76 % have a white background
- Between 57 % and 73 % include links to other people's pages
- Between 51 % and 70 % describe themselves on their websites
- Between 32 % and 50 % use reddish colours for (hyper)text

2.3.8 Application of the multiple correspondence analysis

1. The observations

Since we can't learn more from the previous analyses, let's carry out a multiple correspondence test in order to find variable patterns, but this time without considering gender. In this case, we will focus on the first ten factors since they explain 71 % of the variance. For the complete description of the ten axes, please see appendix E. In the series of tables on next page, you will find a summary of the results.

For each axis, we will consider the four most important variables for the negative part and the four most important variables for the positive one. For example, if we look at axis 1 (the one explaining the most the cloud dispersion), we can see that the variable describing the best the **negative part** of the axis is the *absence of photos*, then it is the *absence of self-photos*, then the *absence of leisure time self-photos* and finally the *absence of personal content* on the website. Regarding the **positive part** of axis 1, the variable that best describes it is the *presence of photos*, then the *presence of self-photos*, then the *presence of leisure time self-photos* and finally the *presence of personal content* on the website. This means the category of academics showing photos and self-photos and particularly leisure time self-photos and having personal content on his/her website is opposed to the category of academics showing no photos, no leisure time self-photos and no self-photos in general and having no personal content on his/her website.

	Axis 1	
Variable	**Negative part**	**Positive part**
Photos	no	yes
Self-photos	no	yes
Leisure time self-photos	no	yes
Personal content (denoted by ratio)	no	yes

	Axis 2	
Variable	**Negative part**	**Positive part**
Artistig graphics	no	yes
Blueish background	no	yes
Computer-related background	no	yes
Family self-photos	yes	no

	Axis 3	
Variable	**Negative part**	**Positive part**
Girlish fonts	yes	no
Classic fonts	no	yes
Not official self-photos	yes	no
Blueish colours (text)	no	yes

	Axis 4	
Variable	**Negative part**	**Positive part**
Original background	no	yes
Trendy graphics	no	yes
White background	yes	no
Leisure time photos	yes	no

	Axis 5	
Variable	**Negative part**	**Positive part**
Comics graphics	no	yes
Colleagues self-photos	no	yes
Grey (text)	no	yes
Original background	yes	no

	Axis 6	
Variable	**Negative part**	**Positive part**
Dark colours for background	no	yes
Private self-description	no	yes
Reddish colours (text)	no	yes
Grey for background	yes	no

	Axis 7	
Variable	**Negative part**	**Positive part**
Highly-technological website	no	yes
White (text)	yes	no
Modern graphics	yes	no
Colleagues photos	no	yes

	Axis 8	
Variable	**Negative part**	**Positive part**
Pets photos	yes	no
Dark colours for background	yes	no
Soft colours for background	no	yes
Black for background	yes	no

	Axis 9	
Variable	**Negative part**	**Positive part**
Family photos	no	yes
Black (text)	yes	no

	Axis 10	
Variable	**Negative part**	**Positive part**
Classic background	no	yes
Pets photos	no	yes

Now, we are going to see if there are any differences between men and women according to their situations in the different factorial plans. From figure 2.35 to figure 2.38, you will find a couple of plans.

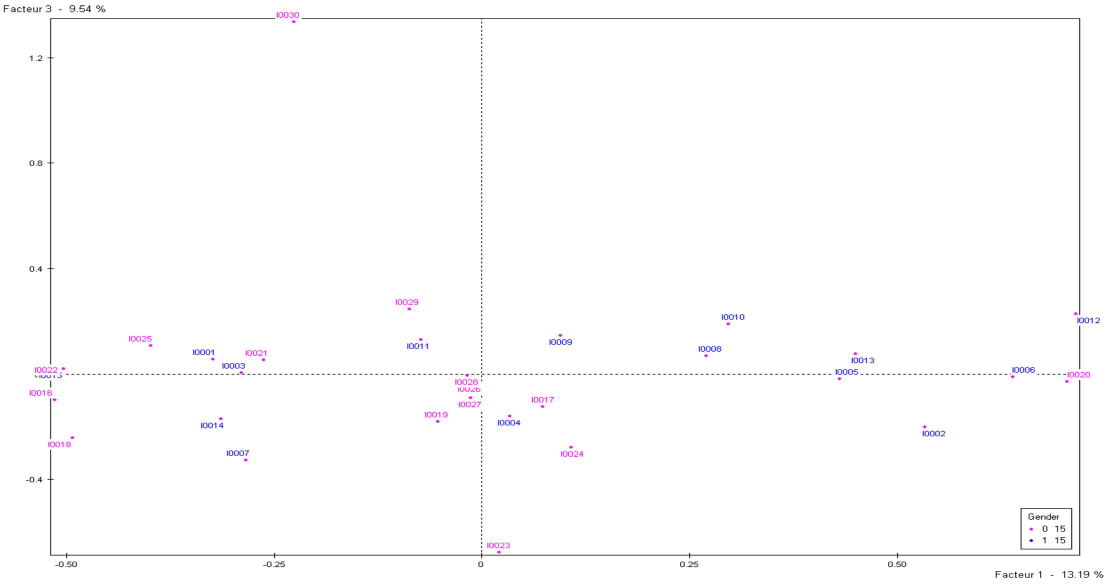


Figure 2.35: MCA for BN: axis 1 vs axis 3

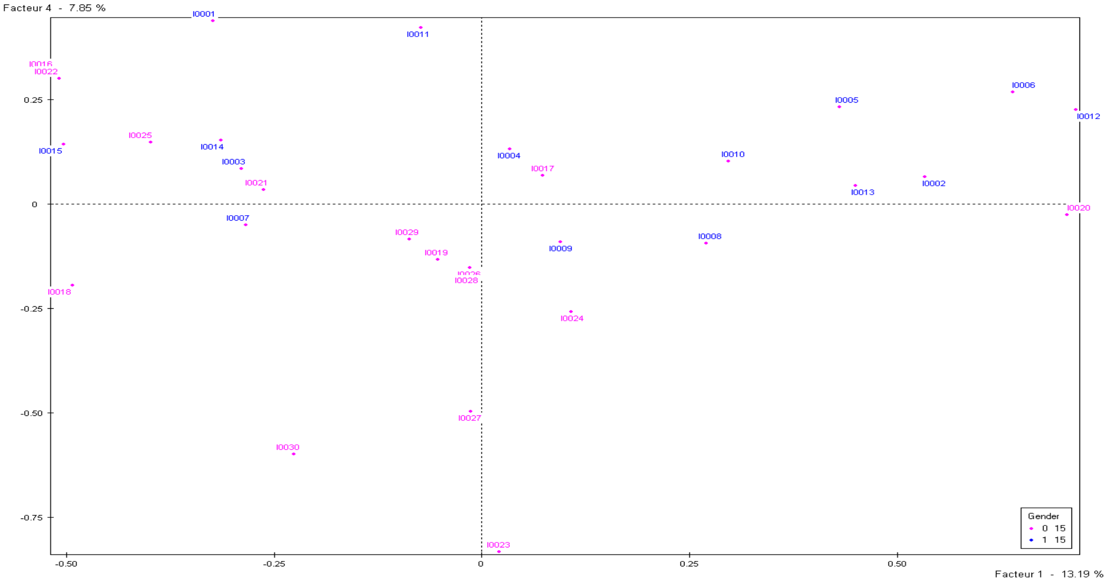


Figure 2.36: MCA for BN: axis 1 vs axis 4

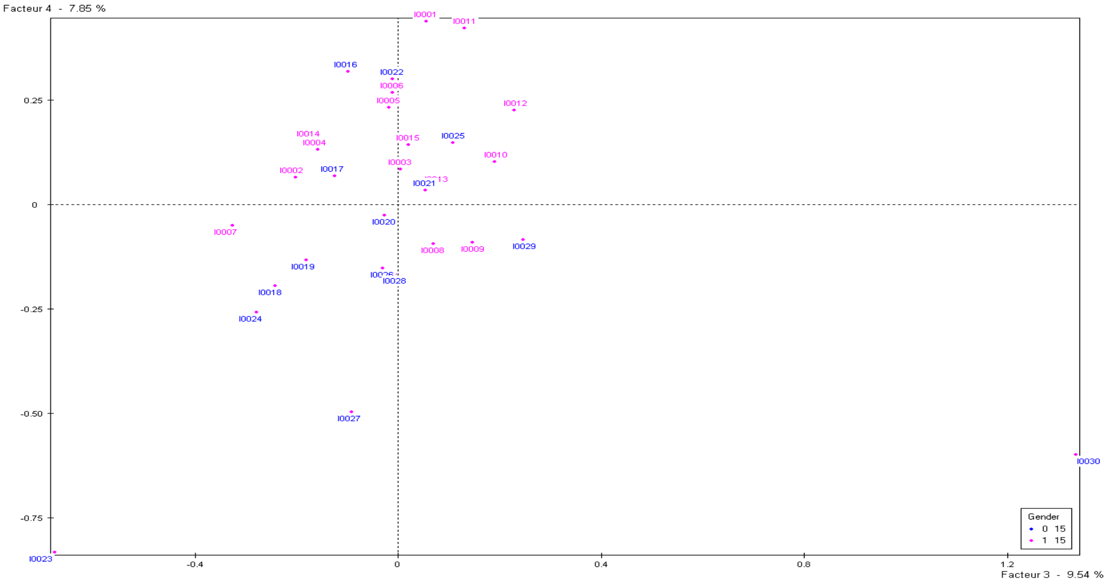


Figure 2.37: MCA for BN: axis 3 vs axis 4

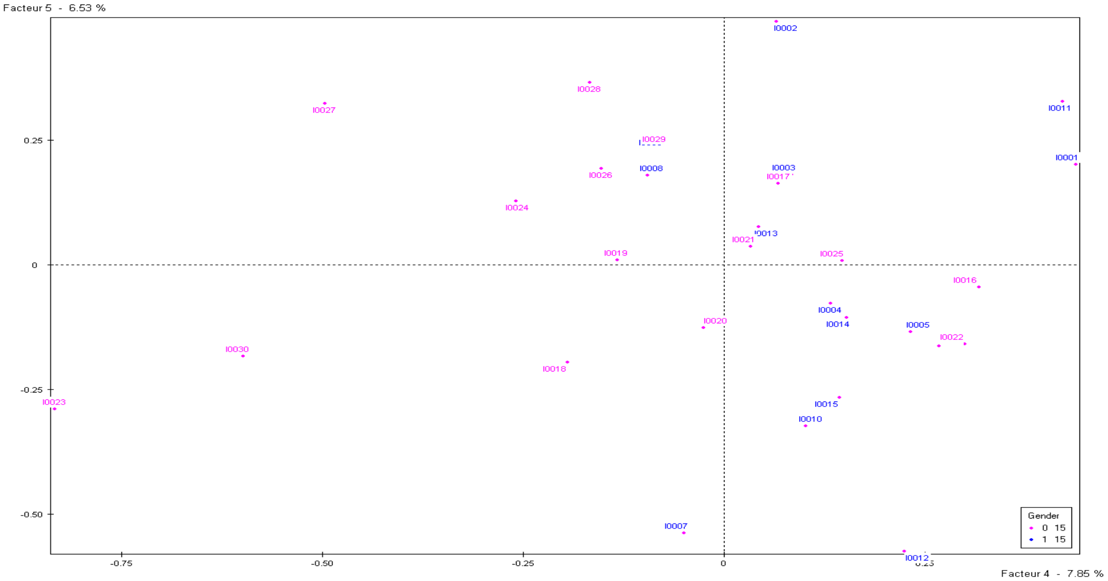


Figure 2.38: MCA for BN: axis 4 vs axis 5

Regarding the first factorial plan (axes 1 and 3, see figure 2.35), we can notice there are **only three females** in the right half of the chart, meaning these **prefer not** to show any **photos** or **self-photos** (and particularly **leisure time self-photos**) and **not** to include **personal content** on their website. **Men** can be found in **both** parts of the plan, if we divide this one according to the first axis.

As you can see in the factorial plan involving the first and fourth axes (figure 2.36), we have the same behaviour as in the previous plan since axis 1 separates males and females again.

Regarding the third factorial plan, with axis 3 and axis 4 (figure 2.37), we can see there are **only three males** in the **bottom part** of the chart, meaning **males** tend to have an **original background** that is not white, with trendy graphics. **Women** do **not seem** to have a **particular preference** since they can be found in **both** parts of the plan.

For the last plan (axis 4 and axis 5, see figure 2.38), there are **more men** in the **right half** of the graphic, that is, if we divide the factorial plan according to axis 4. So, we can make the same observation as for the third factorial plan. For the other combinations of axes, we can't really distinguish any gender pattern.

However, it is interesting to look at the variables (without considering gender) combined with their modalities to know which ones are part of the *common* profile and which ones are part of *aberrant* cases. In table 2.16, you will find the variables being far away from the cloud center (aberrant cases) and table 2.17 shows the ones close to the cloud center (common profile). For example, if we consider table 2.16, the variables defining the best the aberrant cases are the *absence of a classic background*, the *absence of classic fonts*, the *presence of girlish fonts*, the *absence of non-personal links*, the *presence of colleagues and computer-related self-photos*.

variables	**modality**	**DISTO**
classic background	no	29
classic fonts	no	29
girlish fonts	yes	29
non-personal links	no	29
colleagues self-photos	yes	29
computer-related self-photos	yes	29
pets photos	yes	14
blueish colours (text)	no	14
modern graphics	yes	9
computer-related photos	yes	9
family photos	yes	9
friends photos	yes	9
colleagues photos	yes	9
friends self-photos	yes	9
non-official self-photos	yes	9
grey (text)	yes	6,5
trendy graphics	yes	6,5
black (text)	no	6,5

Table 2.16: Aberrant profile

variables	modality	DISTO
classic background	yes	0,03
computer-related self-photos	no	0,03
colleagues self-photos	no	0,03
non-personal links	yes	0,03
girlish fonts	no	0,03
classic fonts	yes	0,03
blueish colours (text)	yes	0,07
pets photos	no	0,07
non-official self-photos	no	0,11
friends self-photos	no	0,11
family photos	no	0,11
friends photos	no	0,11
computer-related photos	no	0,11
colleagues photos	no	0,11
modern graphics	no	0,11
trendy graphics	no	0,15
black (text)	yes	0,15
black (background)	no	0,15

Table 2.17: Common profile

2. The conclusions

The multiple correspondence analysis shows **males** are not as **reserved** as **females**. Indeed, females tend not to put photos and self-photos (particularly leisure time self-photos) on their websites. They don't like including personal content either. The boys don't show any tendency in one way or the other for these features. Maybe that one of the reasons for this observation would be that males feel more confident about themselves than females. But this assumption should be verified by further investigation. We can also note that **males** show a tendency to have an original background, whereas **females** are divided between classic and original backgrounds.

If we consider the whole sample, that is to say males and females put together without any gender distinction, we can bring out **two** tendencies. The first one represents the less common behaviour among our academics, which is defined by the fact of having a website with:

- an original background,
- girlish fonts,
- no non-personal links,
- colleagues and computer-related self-photos,
- pets photos and
- no blueish colours for text and hypertext.

The second tendency represents the most common behaviour among the professors, defined by the fact of having a website with:

- a classic background,
- no colleagues and computer-related self-photos,
- non-personal links,
- classic fonts,
- blueish colours for text and hypertext and
- no pets photos.

Chapter 3

The qualitative analysis

In this chapter, you will learn about the results of a survey conducted among students of a web design class. These students had to answer fifty questions about their design preferences. Ninety questionnaires were collected: fifty-five from male students and thirty-five from female students. First of all, you will get an explanation about the *principal components analysis* since this test is conducted in this chapter. Then we will discuss the results of the different analyses that have been carried out.

3.1 The principal components analysis

This analysis focusses on quantitative features that can be expressed in the same unit or in different units at the same time. They are grouped in a measure table consisting of *n statistical subjects* and *p quantitative variables.* Like the other factorial analyses, the principal components analysis (PCA) focusses on the analysis of the columns (variables) of the information table in order to analyze relationships between variables and highlight more or less systematic combinations between these by simplifying the original information.

3.1.1 The different steps of a PCA

The computation, based on the euclidian distance, consists of three steps:

1. **Creating an information matrix**

   Here, the n subjects must form a coherent set (no aberrant subjects) and the p variables can be heterogeneous.

2. **Altering the original data by centring and reduction of the data (standardization)**

   The information matrix then becomes a correlation matrix between variables. In general, we carry out a *normalized* PCA. That means the variables are standardized, the projection is orthogonal and the adjusting criterion is the least squares method. The correlation matrix is a square matrix of the p^{th} order (“p” rows and “p” columns) with a diagonal equalling 1.

3. **Computing the factorial axes**

   This computation is carried out on the inertia matrix. Each factorial axis is defined by an *eigenvector*. These vectors determine the different directions of the information cloud. The information part taken into account by each eigenvector is the *eigenvalue*. It defines the hierarchy of the factorial axis. The eigenvalue is the variance part (of information) of the cloud taken into account by the axis.

3.1.2 The results for the variables

These results are generally symmetric to the results for the subjects. The **coordinates (or saturation)** of a variable on a factorial axis is the *correlation coefficient* between the variable and the axis. These coefficients vary between +1 and -1 and can all be on the same side of the axis. We then call this axis an “intensity axis”.

The **contributions** (CTR) of the variables to the factorial axes measure the relative role of each variable in the computation (the characterization) of the factorial axes. They allow to figure out which variables are the most contributive per axis. The contributions allow to identify the variables defining best the different axes, that is to say the most contributive variables. They also allow to isolate the variables having an aberrant behaviour.

The **quality** (QLT or Cos^2) of the representation of a variable on a factorial axis is given by its square coordinate on this axis. It measures the part of the variable explained by the axis.

3.1.3 The results for the subjects

The **coordinates (or scores)** of the subjects on the factorial axes allow to situate the subjects along the axes and can be positive or negative (highlighting contrasts).

The **contributions** (CTR) of the subjects to the factorial axes indicate how the subjects contribute to the computation (the characterization) of the factorial axes. The sum of the contributions equals 1. The contributions allow to identify the subjects defining the best the axes. They also allow to isolate the subjects having an aberrant behaviour.

The **quality** (QLT or Cos^2) of the representation of the subjects on the factorial axes allow to characterize the subjects by the axes. They also allow to measure the distance taken into account by an axis between a subject and the centre of gravity. QLT equals Cos^2 of the angle between the subject vector and the axis. $Cos^2(O°) = 1$ means the subject is on the axis (perfect description of the subject by the axis). $Cos^2(90°) = 0$ means the subject is perpendicular to the axis (null description of the subject by the axis). The closer to the cloud center, the less concerned the variables by the definition of the axis and by the contrast between variables or the worse the quality of their representation (they are far from the factor/axis).

Scores and **saturations** are not expressed in the same measure units. However, the subject vectors and the variable vectors have the same directions. These ones can thus be stacked up in the same space and thus go through the cloud center (0,0).

3.1.4 The correlation matrix and the test values

The matrix of the test values allow to test whether the correlation coefficients are relevant or not. If the value is greater than 2, the correlation coefficient is significative with an error of 5 %. The greater the value (in absolute value), the more significative the link between the variables. So, the test values allow to classify the links between the variables into a hierarchy.

3.1.5 Interpretation of the results

How many axes do we have to keep? If all the variables are strongly correlated, only a few axes are sufficient. In a normalized PCA, the cos^2 equal the square coordinates of the variables. Regarding the contributions, we have

$$CTR \;=\; (former\ unit\ axes)^2.$$

When we project the variables into a factorial plan, we can notice the variables forming an acute angle correspond to a highly positive correlation coefficient. Two variables with a highly negative correlation coefficient are diametrically opposed. Two independent variables will have a null correlation coefficient and will form a right angle. The DISTO column gives the square distance between each subject and the gravity centre of the cloud. It allows to find the most common subjects, that is to say the closest to the centre of gravity, and the most original subjects, that is to say the furthermost from the centre of gravity.

3.2 The questions

The first thirty-two questions are related to the students' preferences in web design. Then, the questions focus on the way they would design their own web homepages. For each question (apart from questions 1 and 2), the students could answer “I strongly disagree”, “I disagree”, “I am neutral” (no preference), “I agree” and “I strongly agree”.

1. Are you a male or a female?

2. What nationality are you?

3. I prefer when there are many pages in a site.

4. I prefer when there is a lot of text on the same page.

5. I prefer a page for which I have to scroll down in order to see all text than a page in which all text is cluttered.

6. I prefer using menus to having to navigate to find my way by clicking through the website.

7. I prefer pull-down menus.
8. I prefer menus you have to click through in order to achieve my goal.
9. I prefer when a variety of fonts are used for text and hypertext.
10. I prefer soft colours like pastel colours to dark colours like dark blue or black.
11. I prefer reddish colours (red, yellow, pink, orange etc) to blueish colours (blue, green, purple etc).
12. I prefer when many colours are used for text and hypertext.
13. I prefer when the site's designer shows awareness to the user by using words like "you" etc.
14. I prefer when hypertext is used a lot.
15. I prefer websites in which there are many white spaces between the elements of the site (images, text etc).
16. I prefer when there are a lot of white spaces inside a text.
17. It doesn't matter to me if the page is not well structured.
18. It doesn't matter to me if the site is not well structured.
19. I prefer sites with technological tools like search engines.
20. I prefer when there are a lot of links to other websites.
21. I like websites in which you have many links to people's web homepages to find out who they are.
22. I prefer sites in which you can subscribe to a forum or an online community in order to talk to other people.
23. I prefer when there are many static images in the site.
24. I prefer when there are many graphic animations on a page.
25. I prefer trendy graphics to basic graphics.
26. I prefer comics graphics to basic graphics.
27. I prefer computer-related graphics to basic graphics.
28. I prefer jokey graphics to basic graphics.
29. I prefer a site in which the background is colourful.
30. I prefer a site in which the pages do not look similar. For instance, if there are 10 pages on the site, I prefer when they don't have the same background, the same fonts etc.
31. I prefer a background with motifs than a plain background.
32. I prefer when there are guiding tools that could help me navigate through the site.

If I had to design my own web homepage...

33. I would describe myself so that the reader can know who I am.

34. I would write a SHORT description of myself.

35. I would write a *long* description of myself.

36. I would write a jokey description of myself.

37. I would write a serious description of myself.

38. I would put my self-description on the main page.

39. I would try to include as much information as I can about me (hobbies, personal interests, pictures of my pets etc).

40. I would include a guestbook to let the reader sign it.

41. I would include a counter to count the number of people having visited my site.

42. I would put a lot of information on the main page and not just the link to enter the site or the links to navigate through the site.

43. I would put a picture of myself.

44. I would put *many* pictures of myself.

45. I would put jokey pictures of myself.

46. I would put pictures showing myself on the main page.

47. I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site.

48. I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site.

49. I would put pictures representing my private life on the main page.

50. If I had to include graphics in my web homepage, I would try to make these jokey.

3.3 The analysis

3.3.1 The discriminant analysis

The first test to be carried out is a discriminant analysis in order to know which questions discriminate (separate) the best the male and the female students. Below you will find four tables with the results.

Variables entered/removed		
Step	Entered	Wilks' Lambda Statistic
1	Q32	0,916
2	Q33	0,868
3	Q8	0,824

Please note that at each step, the variable minimizing the overall Wilks' Lambda is entered. So, as we can see in the table, the first variable to be entered is **question 32**, that is to say "I prefer when there are guiding tools that could help me navigate through the site". The second one is **question 33** which is "If I had to design my own web homepage, I would describe myself so that the reader can know who I am". Finally, **question 8** is entered, that is to say "I prefer menus you have to click through in order to achieve my goal".

Variables in the analysis		
Step	**Question**	**Wilks'Lambda**
1	Q32	
2	Q32	0,983
	Q33	0,916
3	Q32	0,943
	Q33	0,889
	Q8	0,868

This table indicates the hierarchy of the questions taken into account in the analysis. Since the Wilks' Lambda for question 8 is the **lowest value**, we can say the statement "I prefer menus you have to click through in order to achieve my goal" is the **most discriminant** feature between male and female students. The most discriminant variable after question 8 is question 33 and thus the **less discriminant variable** is question 32 with the guiding tools.

Classification Function Coefficients		
	Gender 0 (F)	Gender 1 (M)
Q8	4,959	4,317
Q32	3,867	2,979
Q33	1,846	2,595
(Constant)	-19,605	-16,816

The functions described in the above table can be rewritten as follows:

$$F_0(x) = 4,959\ Q_8 + 3,867\ Q_{32} + 1,846\ Q_{33} - 19,605$$

and

$$F_1(x) = 4,317\ Q_8 + 2,979\ Q_{32} + 2,595\ Q_{33} - 16,816.$$

So, if we have a new subject, we will be able to say if this person has a male or a female profile according to the score resulting from the computation of these two functions. For example, if $F_1(x)$ gives a higher score than $F_0(x)$, this will mean the person is likely to have a masculine profile.

From this table, we can also see the coefficient for **question 8** is much **higher for the females** than for the males. This means the girls have a greater preference for menus you have to click through than the boys. Let's remember that in [WBT02], 52 % of women said they preferred pull-down menus rather than navigating through the site. For **question 32**, we also have a much **higher coefficient for the females** than for the males. This means the girls have a greater preference for the presence of guiding tools in the website than the boys. As said in [Khu04], girls ask for help as soon as they get stuck while boys keep navigating through the environment until they find their way around. Once again, our result confirms the fact women don't like exploring without help. Regarding **question 33**, we have a **higher coefficient for the boys** than for the girls, meaning the boys have a greater tendency to describe themselves on their web homepage than the girls. This finding hasn't been highlighted in previous research (see chapter 1) and is very interesting since no difference between male and female adults was found in [AM99a] apart from the focus on credentials. To visualize these statements, let's have a look at the frequency charts for each of the three questions. Please remember that *Series 1* represents the males and *Series 2* the females.

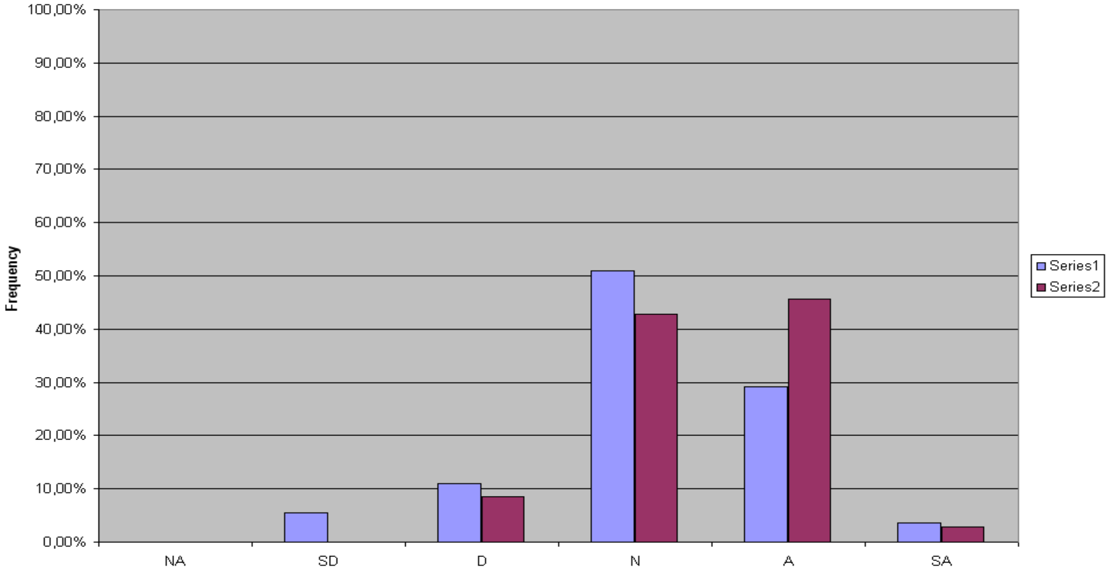


Figure 3.1: Distributions of the males and the females for Q8

As you can see in figure 3.1, girls are neutral or agree compared with boys who answered they were neutral or they disagreed. From figure 3.2, we can notice boys answered they were neutral or they agreed compared with girls who answered they were neutral or they disagreed. Figure 3.3 shows girls were neutral, agreed or strongly agreed with this question whereas boys were neutral or disagreed with the statement. What we can observe from these three charts

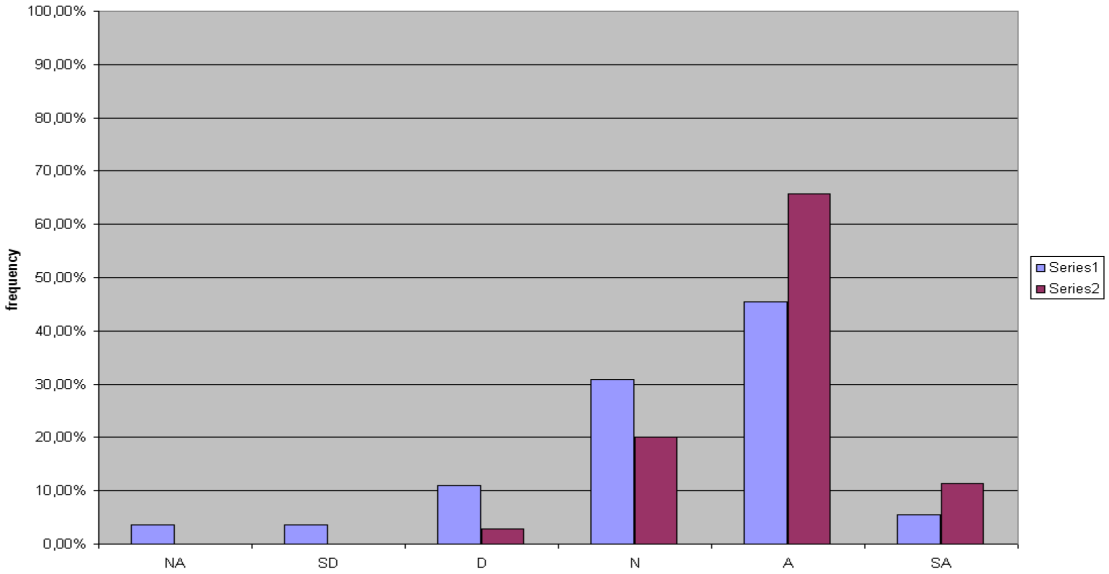


Figure 3.2: Distributions of the males and the females for Q32

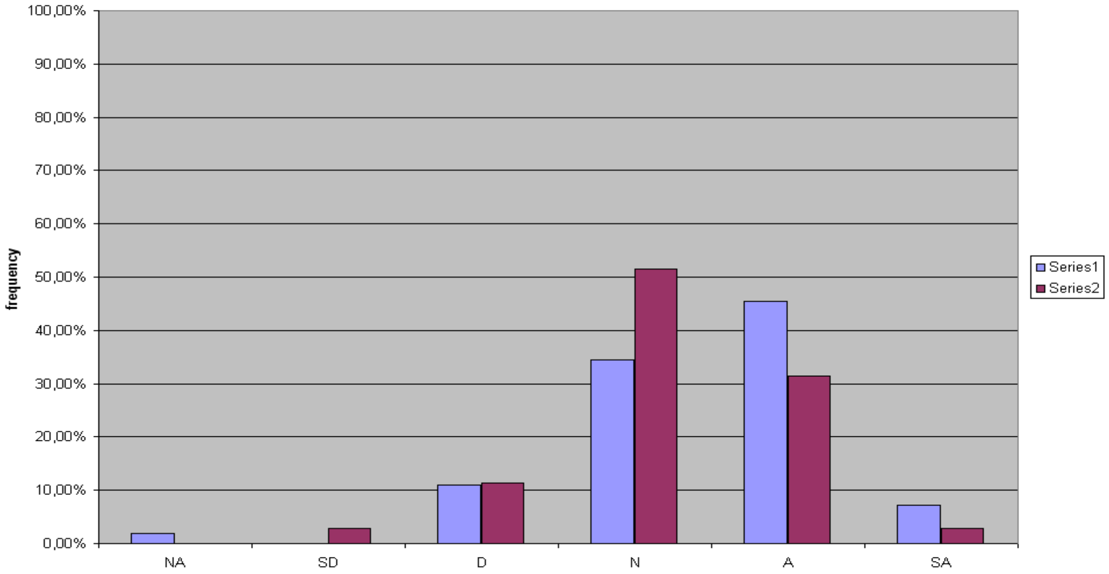


Figure 3.3: Distributions of the males and the females for Q33

thus goes along with the findings from the discriminant analysis.

Classification results				
		PGM 0 (F)	**PGM 1 (M)**	**Total**
Original count	0 (F)	23	12	35
Original count	1 (M)	17	38	55
%	0 (F)	65,7	34,3	100
%	1 (M)	30,9	69,1	100

In order to classify our subjects, the two functions described above were computed for each student. If this one got a higher score with F_0, he/she was considered as a girl, otherwise as a boy. Let's notice 67,8% of original grouped cases are correctly classified. This last table shows 12 females (34,3% of the feminine group) are not classified properly since they have a more masculine profile. Regarding the males, 17 (30,9%) belong to the feminine group since they have a more feminine profile.

3.3.2 The principal components analysis

1. The correlation matrix

The correlation matrix is very useful to figure out the strength of the links between the different questions. Since the number of axes we have to take into account in order to explain 70% of the cloud inertia is high (29 axes), we already know the links won't be strong. That is why we will consider the questions having a correlation coefficient of **0,5** or **greater**. In table 3.1 you will find a couple of questions linked with each other, their correlation coefficients and the corresponding test values (the higher the test value, the more significative the correlation coefficient).

Linked questions	Correlation coefficient	Test value
Q44-Q48	0,79	10,06
Q47-Q48	0,74	9,12
Q46-Q47	0,71	8,44
Q48-Q49	0,69	7,98
Q43-Q44	0,68	7,87
Q44-Q46	0,65	7,34
Q44-Q47	0,64	7,26
Q46-Q48	0,64	7,26
Q40-Q41	0,64	7,24
Q43-Q48	0,64	7,12
Q43-Q46	0,63	7,01
Q35-Q49	0,63	7,00
Q44-Q49	0,61	6,78
Q43-Q47	0,6	6,53
Q39-Q44	0,59	6,49
Q45-Q47	0,59	6,45
Q45-Q48	0,58	6,22
Q46-Q49	0,56	6,01
Q9-Q12	0,56	5,95
Q35-Q44	0,55	5,85

Second part of table on next page...

Linked questions	Correlation coefficient	Test value
Q35-Q48	0,54	5,75
Q36-Q45	0,54	5,75
Q47-Q49	0,53	5,63
Q43-Q49	0,53	5,61
Q24-Q29	0,53	5,54
Q45-Q50	0,51	5,29
Q45-Q49	0,50	5,17

Table 3.1: The most strongly linked questions

Let's give a couple of details about this table. The explanations are classified into a hierarchy according to the correlation coefficient.

1. **Q44 and Q48**

   The strongest link combines "I would put *many* pictures of myself in my own site" and "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site".

2. **Q47 and Q48**

   The second strongest link concerns "I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site" and "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site".

3. **Q46 and Q47**

   Here we have a combination between "I would put pictures showing myself on the main page" and "I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site".

4. **Q48 and Q49**

   "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site" and "I would put pictures representing my private life on the main page" are concerned by this link.

5. **Q43 and Q44**

   Here the questions being linked are "I would put a picture of myself in my own site" and "I would put *many* pictures of myself in my own site".

6. **Q44 and Q46**

   The sixth strongest link combines "I would put *many* pictures of myself in my own site" and "I would put pictures showing myself on the main page".

7. **Q44 and Q47**

   The seventh strongest link concerns "I would put *many* pictures of myself in my own site" and "I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site".

8. **Q46 and Q48**

   Here we have a link between "I would put pictures showing myself on the main page" and "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site".

9. **Q40 and Q41**

   This combination concerns "I would include a guestbook to let the reader sign it" and "I would include a counter to count the number of people having visited my site".

10. **Q43 and Q48**

    Here we have a correlation between "I would put a picture of myself in my own site" and "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site".

11. **Q43 and Q46**

    The eleventh combination links "I would put a picture of myself in my own site" and "I would put pictures showing myself on the main page".

12. **Q35 and Q49**

    Here we have a link between "I would write a *long* description of myself" and "I would put pictures representing my private life on the main page".

13. **Q44 and Q49**

    The thirteenth link combines "I would put *many* pictures of myself in my own site" and "I would put pictures representing my private life on the main page".

14. **Q43 and Q47**

    The fourteenth correlation concerns "I would put a picture of myself in my own site" and "I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site".

15. **Q39 and Q44**

    Here we have a link between "I would try to include as much information as I can about me (hobbies, personal interests, pictures of my pets etc)" and "I would put *many* pictures of myself in my own site".

16. **Q45 and Q47**

    Here we have a correlation between "I would put jokey pictures of myself in my own site" and "I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site".

17. **Q45 and Q48**

    Here "I would put jokey pictures of myself in my own site" and "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site" are linked.

18. **Q46 and Q49**

    Here the questions being linked are "I would put pictures showing myself on the main page" and "I would put pictures representing my private life on the main page".

19. **Q9 and Q12**

    The nineteenth link combines "I prefer when a variety of fonts are used for text and hypertext" and "I prefer when *many* colours are used for text and hypertext".

20. **Q35 and Q44**

    Here "I would write a *long* description of myself" and "I would put *many* pictures of myself in my own site" are correlated.

21. **Q35 and Q48**

    This combination concerns “I would write a *long* description of myself” and “I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site”.

22. **Q36 and Q45**

    The twenty-second correlation combines “I would write a jokey description of myself” with “I would put jokey pictures of myself in my own site”.

23. **Q47 and Q49**

    Here the questions being linked are “I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site” and “I would put pictures representing my private life on the main page”.

24. **Q43 and Q49**

    This link concerns “I would put a picture of myself in my own site” and “I would put pictures representing my private life on the main page”.

25. **Q24 and Q29**

    The twenty-fifth link combines “I prefer when there are many graphic animations on the page” and “I prefer a site in which the background is colourful”.

26. **Q45 and Q50**

    Here the questions being linked are “I would put jokey pictures of myself in my own site” and “If I had to include graphics in my web homepage, I would try to make these jokey”.

27. **Q45 and Q49**

    The last correlation combines “I would put jokey pictures of myself in my own site” and “I would put pictures representing my private life on the main page”.

We can classify these links into different groups in order to bring out tendencies. The first group (see figure 3.4) consists of **link 1 up to link 8, link 10 up to link 15, link 18, links 20 and 21, and finally links 23 and 24**. From this group, we can state the student is likely to put private information on his/her site, that is to say he/she would include self-pictures, pictures of his/her private life and that he/she would put all these pictures on the main page. The main page then becomes a spot where private information is displayed. According to this first classification, the student would try to include as much information as he/she can about him/her, that is to say self-pictures, pictures of his/her private life, textual information about his/her hobbies, personal interests, pictures of pets etc. That is probably why findings in previous research highlight the fact women do not dare to put their self-pictures on the main page, since it is a spot where private life is displayed and they dare not upload much private information about themselves. In this first group, we can also bring out the fact the student would write a long description of himself/herself, thus considering a long self-description as being part of the private sphere.

The next group (see figure 3.5) consists of **links 16, 17, 22, 26 and 27**. This classification highlights a self-mockery tendency, that is to say the student would write a jokey self-description. Moreover, if he/she put pictures of himself/herself on the Internet, he/she would put jokey pictures, and these would be considered as private elements. The student would also make the graphics of the pages jokey. This might be the reason why it was found in previous investigations that women remained serious on their websites since jokey items are correlated with private sphere.

The third group (see figure 3.6) consists of **links 19 and 25**. This is the group of “fluffy” websites, that is to say the student prefers websites in which there is a variety of fonts with many different

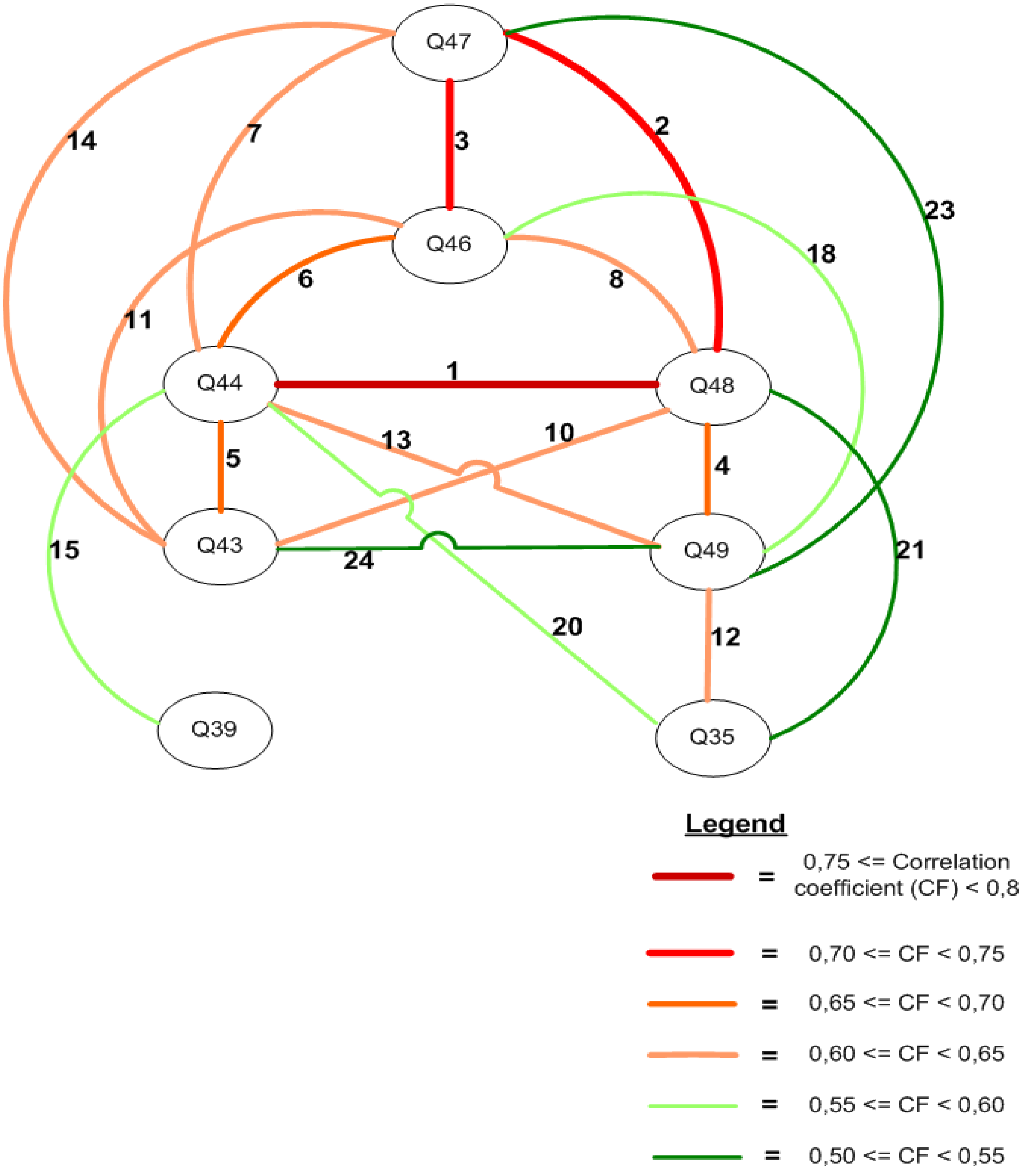


Figure 3.4: Relationships between questions - first group

colours for text (and hypertext) combined with many graphic animations (thus much movement) on a colourful background. This concurs with the definition of a “fluffy feminine” site consisting of many different fonts, many text and background colours, much movement etc.

The last group consists of the remaining link, that is to say **link 9**. Here the student would show

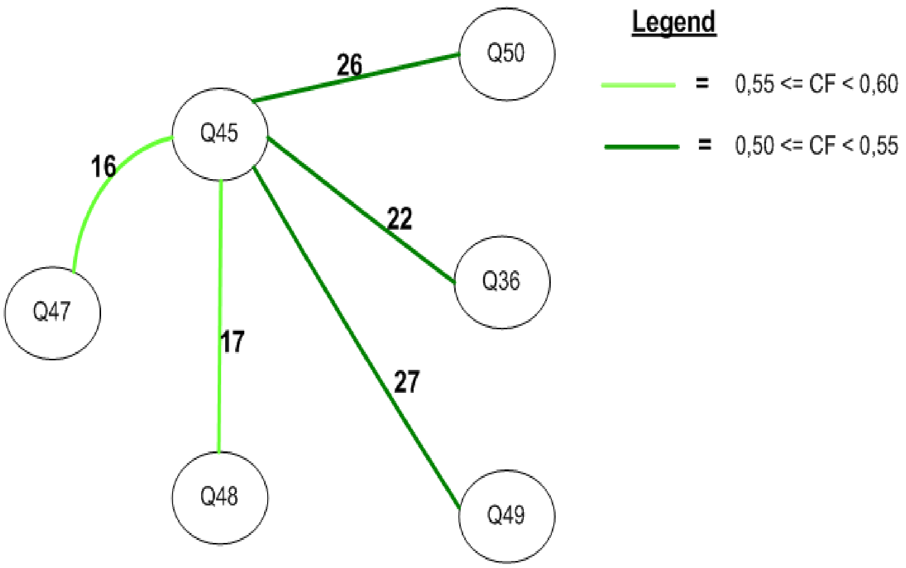


Figure 3.5: Relationships between questions - second group

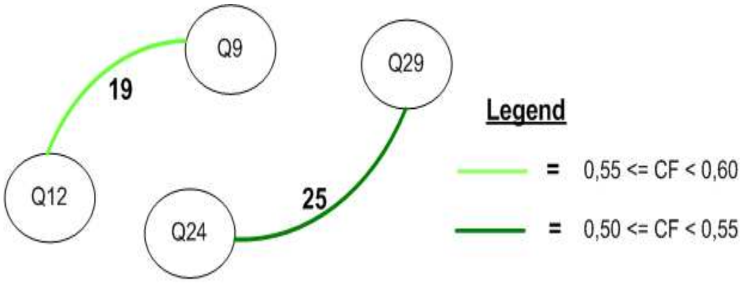


Figure 3.6: Relationships between questions - third group

awareness to the reader by including a guestbook to let him/her sign it combined with a counter to let the reader know how often the site is visited.

2. The factors

According to the results of the test, we have to keep fourteen new variables, combination of different questions, in order to explain 70% of the cloud inertia. In appendix F, you will find the complete description of the fourteen axes. Below you will find the four most important questions for the **negative part** and for the **positive one** of each axis. For example, if we look at axis 2, we can see the questions describing the best the **negative part** of the axis are Q34, Q10, Q32 and Q40. The questions describing the best the **positive part** of axis 2 are Q18, Q17, Q35 and Q48.

Description of axis 1	
Negative part	**Positive part**
/	Q47
/	Q48
/	Q44
/	Q46

This intensity axis concurs with the tendency we have already brought out (see correlation matrix). This table thus confirms the student tends to put (many) pictures of his/her private life with many self-pictures he/she would put on the main page. As we have already said, self-pictures are thus considered as being part of the private sphere and the main page is the preferred spot to upload these private elements.

Description of axis 2	
Negative part	**Positive part**
Q34	Q18
Q10	Q17
Q32	Q35
Q40	Q48

For axis 2, we have an opposition between the **negative part** of the axis and the **positive one**. Actually, the **negative part** combines the short self-description, the preference for soft colours, the guiding tools and the guestbook. For the **positive part**, we have a correlation between the fact it is not important if the site/the page is not well structured, the long self-description and the upload of many pictures of the student's private life.

Description of axis 3	
Negative part	**Positive part**
Q23	Q14
Q25	Q9
Q22	Q5
Q43	Q3

The **negative part** of axis 3 combines a preference for sites with many static images and in which you can subscribe to a forum, the fact the student prefers trendy graphics to basic graphics and finally the fact he/she would put a self-picture on his/her site. The **positive part** consists of a preference for an intensive use of hypertext with a variety of fonts for text and for sites with many pages for which you have to scroll down in order to see all text than pages in which all text is cluttered.

Description of axis 4	
Negative part	**Positive part**
Q44	Q28
Q43	Q27
Q4	Q12
Q3	Q26

The **negative part** of axis 4 consists of a preference for websites with many pages on which there is a lot of text combined with the fact the student would put many self-pictures in his/her own site. The **positive part** is composed of the preference for jokey, computer-related and comics graphics combined with many colours for text and hypertext.

Description of axis 5	
Negative part	**Positive part**
Q29	Q36
Q24	Q33
Q41	Q13
Q40	Q22

For the **negative part** of axis 5, we have a combination between the preference for a colourful background with many graphic animations on it, and the fact the student's homepage would include a guestbook and a counter. Regarding the **positive part**, we have a combination between the self-description to let the reader know who the student is, self-description that would be jokey, the preference for many colours for text and hypertext and for the web designer's awareness to the reader.

Description of axis 6	
Negative part	**Positive part**
Q30	Q14
Q33	Q31
Q21	Q37
Q11	Q35

The **negative part** of axis 6 highlights the relationship between the preference for a site in which the pages do not look similar and in which you can find many links to other people's homepages, the preference for reddish colours to blueish ones and the fact the student would write a self-description to let the reader know who he/she is. The **positive part** consists of the preference for an intensive use of hypertext with a background with motifs rather than a plain background and the fact the student would write a long and serious self-description.

Description of axis 7	
Negative part	**Positive part**
Q23	Q8
Q4	Q21
Q45	Q27
Q15	Q20

The **negative part** of axis 7 links the preference for a lot of text on the same page with a spaced-out website, the preference for many static images and the fact the student would put jokey self-pictures. The **positive part** combines the preference for menus you have to click through with sites in which you can find many links in general and especially many to people's web homepages and a preference for computer-related graphics to basic graphics.

Description of axis 8	
Negative part	**Positive part**
Q11	Q42
Q22	Q47
Q25	Q19
Q4	Q15

For the **negative part** of axis 8, we have a correlation between the preference for pages on which there is a lot of text, the preference for reddish colours to blueish colours, the preference for sites in which you can subscribe to a forum and the preference for trendy graphics to basic graphics. Regarding the **positive part**, we have a combination between the preference for spaced-out websites and in which you have technological tools, the fact the student would put a lot of information on the main page and not just the link to enter the site and the fact the student would insert pictures of his/her private life.

Description of axis 9	
Negative part	**Positive part**
Q20	Q10
Q7	Q5
Q14	Q31
Q23	Q18

For the **negative part** of axis 9, we have a link between the preference for pull-down menus, for an intensive use of hypertext, for websites in which there are many links to other sites and with many static images. The **positive part** combines the preference for pages for which you have to scroll down in order to see all text, the preference for soft colours, the fact it is not important if the site is not well structured and the preference for a background with motifs.

Description of axis 10	
Negative part	**Positive part**
Q37	Q8
Q30	Q19
Q32	Q7
Q38	Q28

Regarding the **negative part** of axis 10, the correlation concerns the preference for sites which do not look similar, with guiding tools, the fact the student would write a serious self-description he/she would put on the main page. The **positive part** links the preference for pull-down menus as well as for menus you have to click through, the preference for sites with technological tools and lastly the preference for jokey graphics.

Description of axis 11	
Negative part	**Positive part**
Q42	Q16
Q13	Q15
Q23	Q12
Q19	Q34

For the **negative part** of axis 11, the link concerns the preference for the designer's awareness to the reader, for sites with technological tools and many static images and finally the fact the student would put a lot of information on his/her main page and not just the link to enter the site or the links to navigate through it. The **positive part** combines the preference for an intensive use of colours for text and hypertext, with a spaced-out website and spaced-out text and the fact the student would write a short self-description.

Description of axis 12	
Negative part	**Positive part**
Q16	Q7
Q24	Q40
Q19	Q41
Q20	Q18

Regarding the **negative part** of axis 12, the relationship concerns the preference for spaced-out text, for sites with technological tools and with a lot of links to other websites and with many graphic animations. The **positive part** combines the preference for pull-down menus and the fact it is not important if the page is not well structured, the fact the student would include a guestbook and a counter.

Description of axis 13	
Negative part	**Positive part**
Q28	Q8
Q32	Q33
Q12	Q25
Q16	Q4

The **negative part** of axis 13 combines the preference for many colours for text and hypertext with spaced-out text, jokey graphics and guiding tools. Regarding the **positive part**, the correlation concerns the preference for pages on which there is a lot of text, menus you have to click through, trendy graphics and the fact the student would describe himself/herself to let the reader know who he/she is.

Description of axis 14	
Negative part	**Positive part**
Q15	Q17
Q41	Q39
Q11	Q32
Q37	Q34

Regarding the **negative part** of axis 14, we have a combination between the preference for reddish colours, for spaced-out websites, the fact the student would write a serious self-description and would include a counter. The **positive part** links the fact it is not important if the page is not well structured, the preference for guiding tools, the fact the student would write a short self-description and would include as much information as he/she can about him/her.

3. The factorial plans

Let's project our subjects in a few plans in order to know how they are situated on the different axes we have described above.

- **Axis 1 vs axis 2**

If we refer to the investigations described in chapter 1, we could say we have an **opposition** between a majority of girls and a majority of boys. According to [WBT02], girls prefer to get help in order to navigate through a site. Guiding tools are thus a way to provide them with help. If we refer to [MM98a], we can say the presence of a guestbook is a feminine feature since the authors of this article found more guestbooks on females' pages than on males'.

Regarding the boys, it is stated in [Khu04] these can more easily navigate even if there is no structure in a page or in a site in general.

As you can see in figure 3.7, approximately **two thirds** of the girls are situated in the bottom part of the chart (when dividing this one according to axis 2). It thus **validates** our assumptions and the findings from the articles we have just mentioned.

Regarding the boys, they are more or less **equally** distributed in the top and bottom parts of figure 3.7. So, according to our sample, these do not show any preference regarding the questions describing axis 2 (see *The factors* above) and thus we **can't validate** the masculine character of the positive part of axis 2.

- **Axis 1 vs axis 4**

If we refer to [MM98a], we can put forward a male profile since men are more likely to put jokey and comics graphics in their sites compared with women.

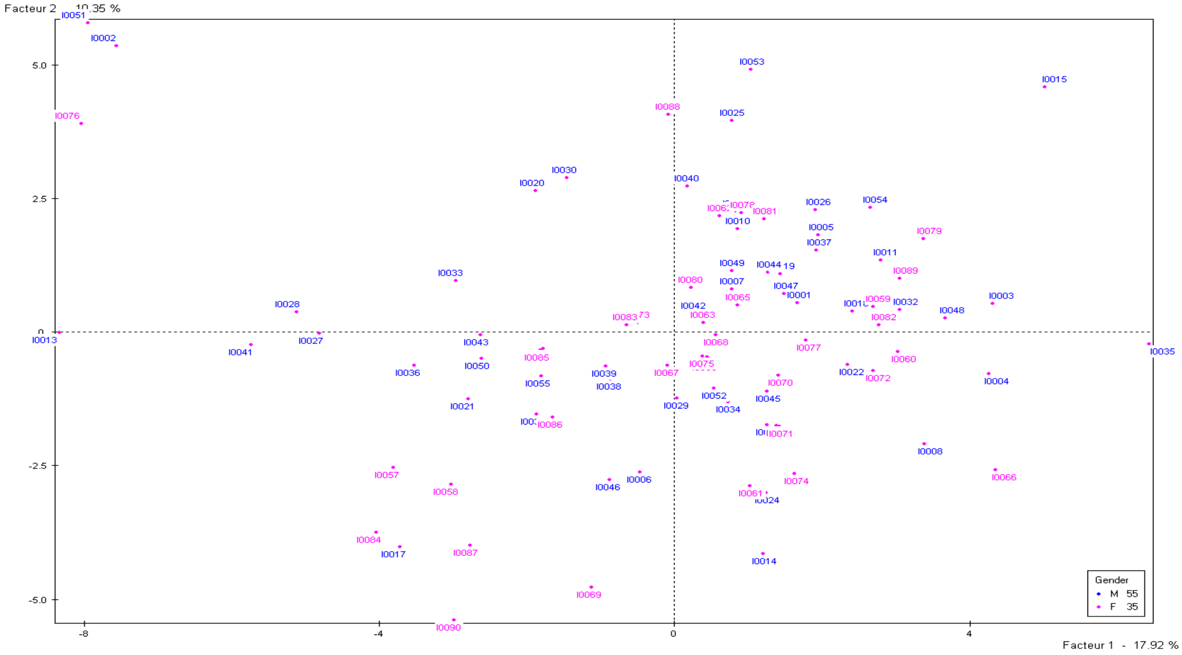


Figure 3.7: PCA for the questions: axis 1 vs axis 2

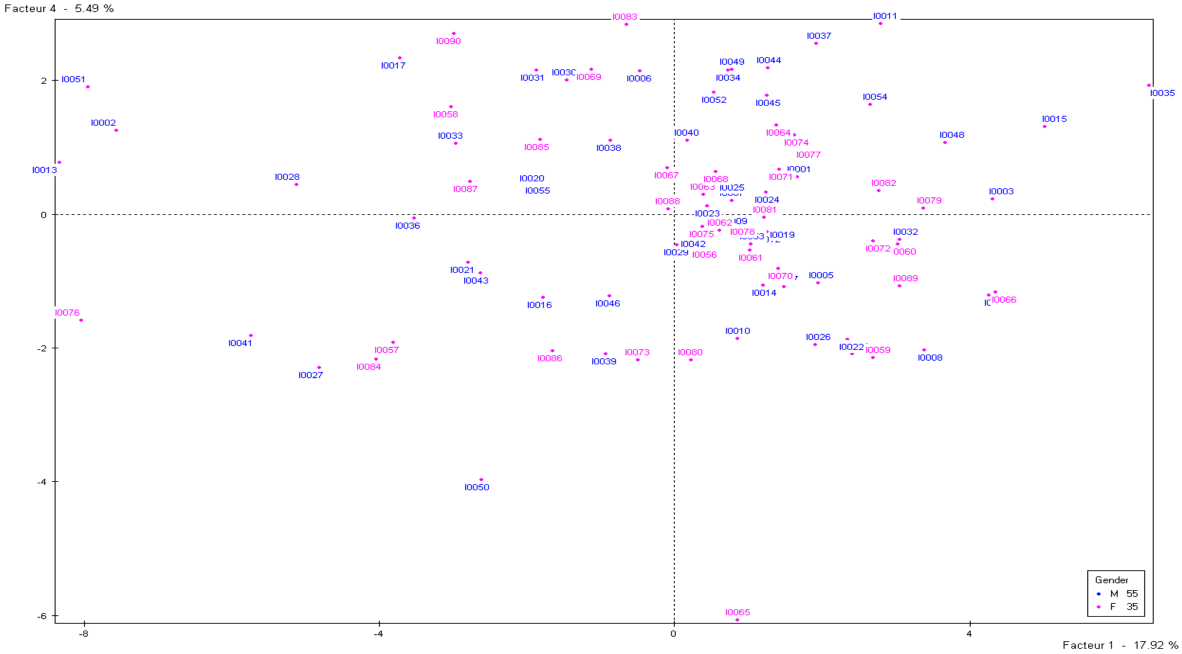


Figure 3.8: PCA for the questions: axis 1 vs axis 4

But in figure 3.8, we can notice boys are more or less **equally** distributed in the top and bottom parts of the chart (when dividing this one according to axis 4). So, according to our observations, these do **not** show any tendency in one way or the other.

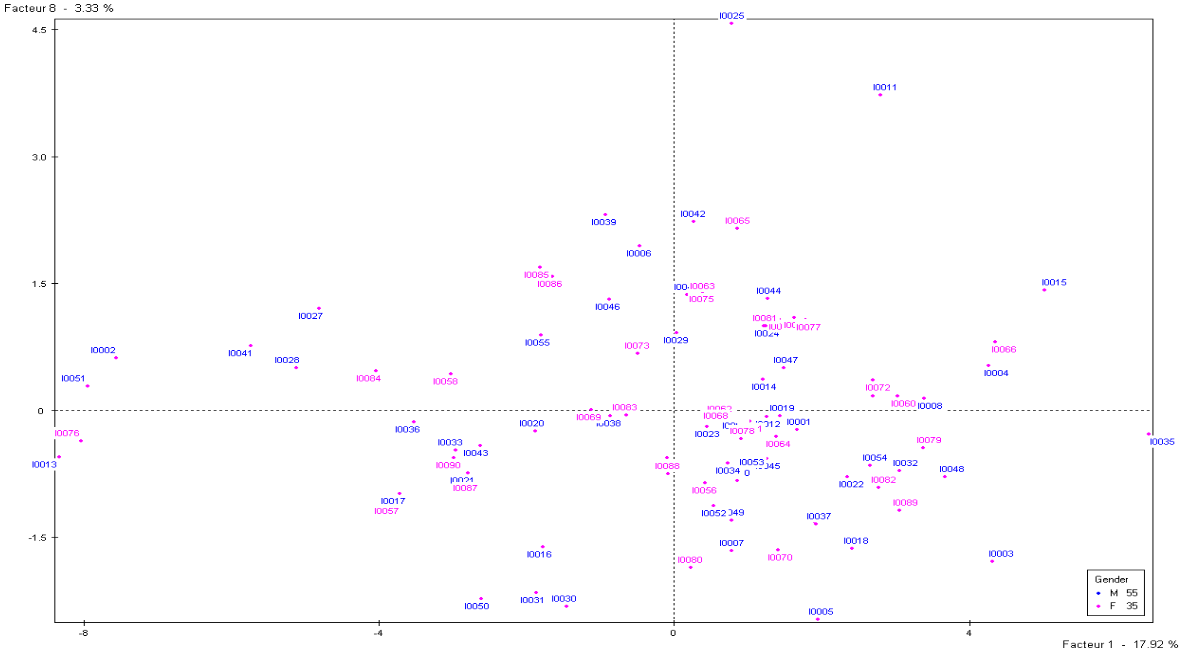


Figure 3.9: PCA for the questions: axis 1 vs axis 8

- **Axis 1 vs axis 8**

Here again, if we refer to previous research, we could say we are confronted with a feminine profile according to the preference for reddish colours and forums. Indeed, [Kho] highlights women prefer red to blue and [WBT02] says women place more value upon interpersonal communication, thus subscribing more often to forums in order to communicate with other people.

But figure 3.9 shows girls are **equally** distributed in the top and bottom parts of the chart. So, there is **no** "feminine" tendency as we could first reckon.

- **Axis 1 vs axis 11**

On the basis of chapter 1, we are in presence of a feminine profile again. Indeed, in [AM99a], girls emphasised their preference for the use of colours and in the same gender study, they overwhelmingly (84%) prefer sites that are less cluttered.

Figure 3.10 indicates approximately **two thirds** of the girls are situated in the top part, **strengthening** the findings from the previous article we have just mentioned.

- **Axis 1 vs axis 13**

We are confronted with a feminine tendency as well if we still refer to literature. Actually, the preference for many colours in text and hypertext corresponds to the finding of [WBT02] in which it was reported girls emphasised a preference for the use of colours. The preference for spaced-out text corresponds

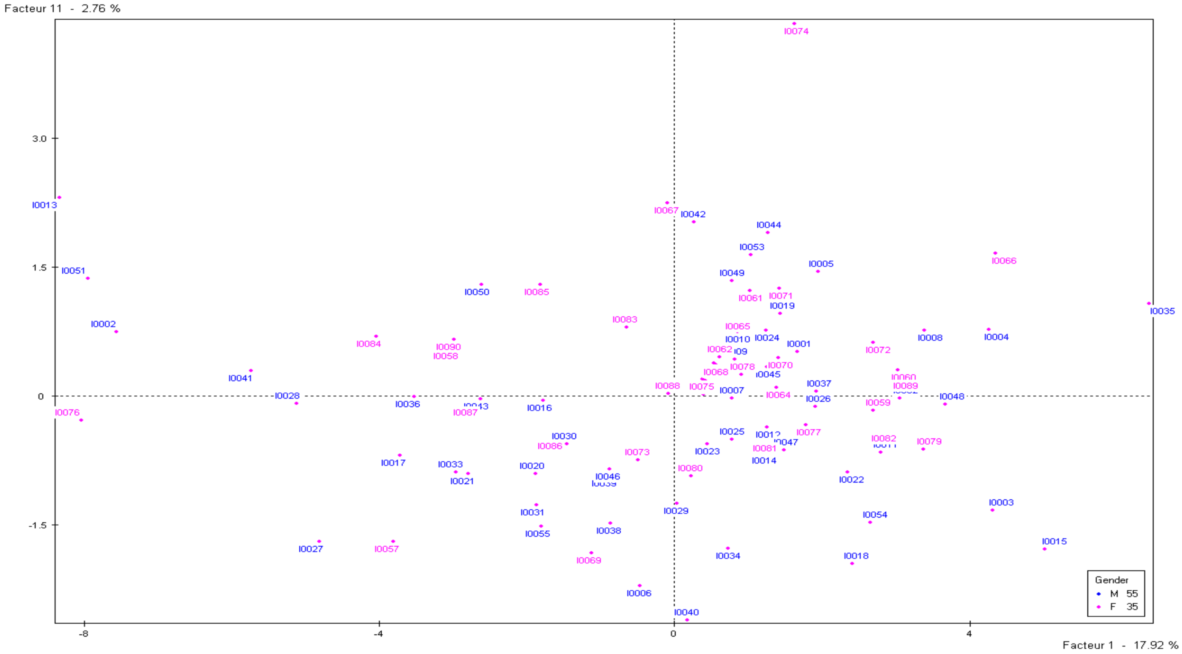


Figure 3.10: PCA for the questions: axis 1 vs axis 11

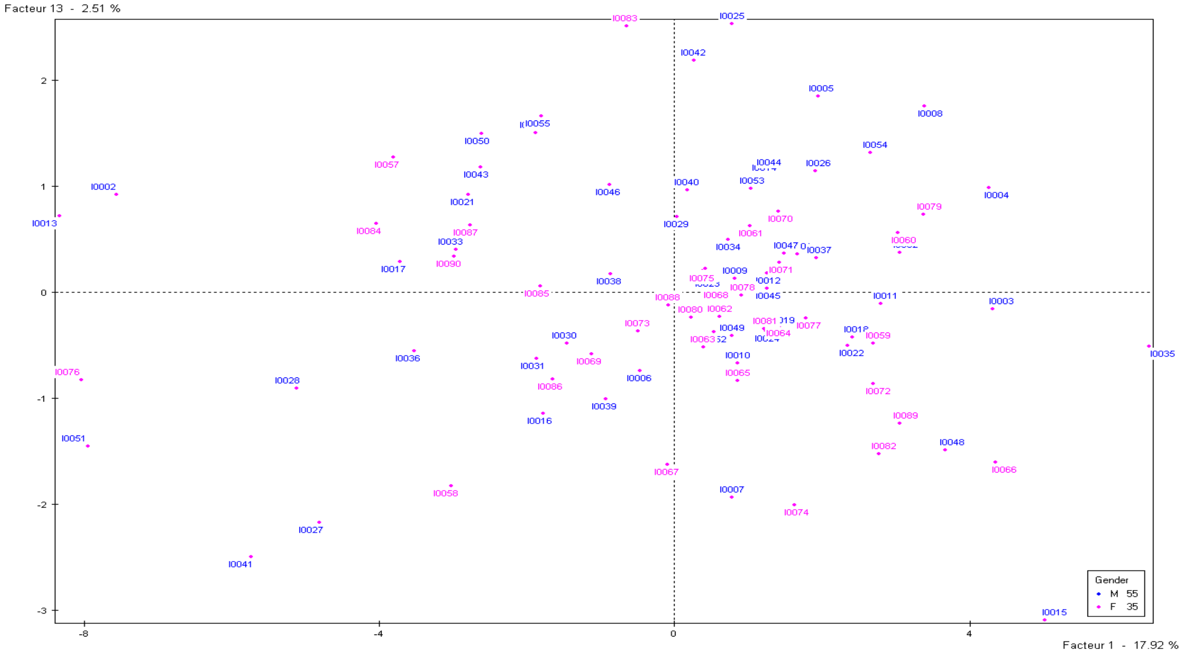


Figure 3.11: PCA for the questions: axis 1 vs axis 13

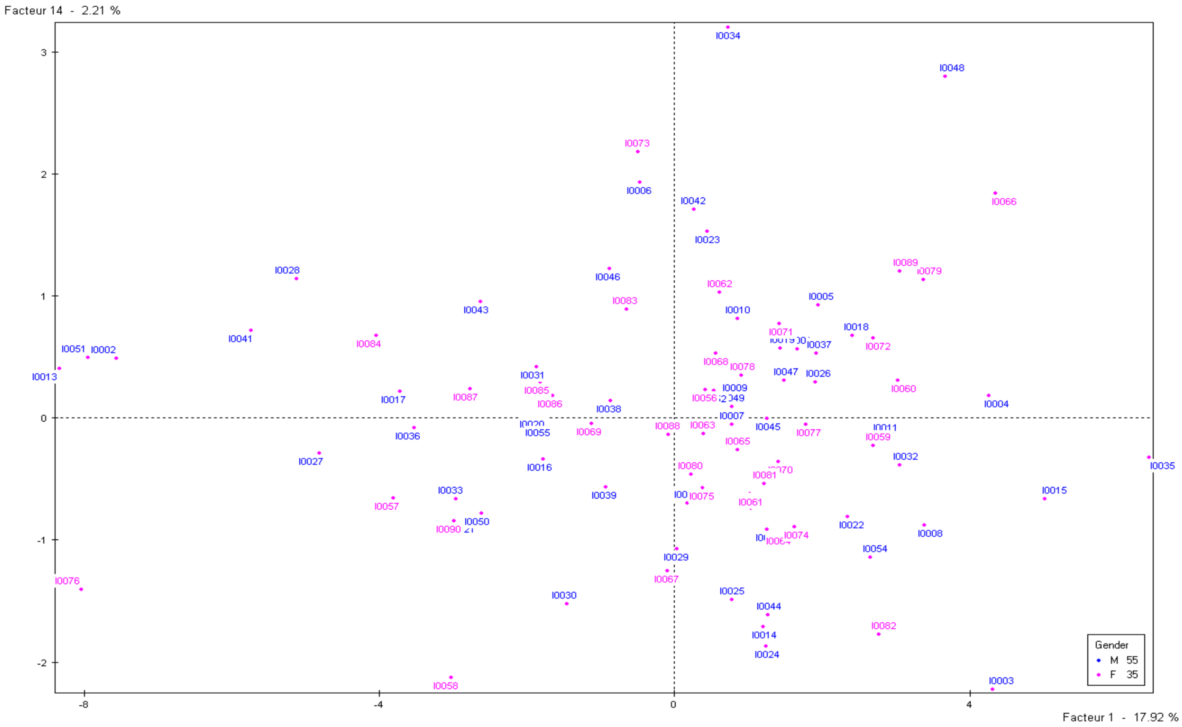


Figure 3.12: PCA for the questions: axis 1 vs axis 14

again to a finding of [WBT02] in which females indicate they overwhelmingly (84 %) prefer sites that are less cluttered. Regarding the question related to the use of guiding tools, we can mention again the finding of [Khu04] in which it was stated girls prefer to get help in order to navigate through a site and this can be put in practice with guiding tools.

When looking at figure 3.11, we can clearly see there are **more women** in the bottom part of the chart than in the top one. So, our observations **strengthen** the findings from previous research.

- **Axis 1 vs axis 14**

Again, we are in the presence of a feminine profile if we refer to [Kho] for women's preference for reddish colours and to [WBT02] for their preference for non-cluttered sites. Regarding the presence of guiding tools, this refers again to [Khu04]. For the serious self-description, let's remember the authors of [AM99a] said female academics included a full CV, list of honours etc when describing themselves (seriously then).

But in figure 3.12, we can't notice any tendency since the girls are **equally** distributed in the top and bottom parts of the chart. So, according to our sample, the girls do not show any tendency in one way or the other.

3.3.3 The segmentation tree

To find other possible gender differences, let's conduct a segmentation test. Let's remember that our goal is to divide the group of students according to their gender. Thus we try to get homogeneous groups of boys and homogeneous groups of girls. The method we will use to do so will be the *entropy reduction* again. Before running the test, a grouping had to be carried out in order to have enough subjects in the leaves of the tree to be able to analyse it. Thus the subjects having answered “I strongly disagree” and “I disagree” have been grouped in **category 1**. The ones being neutral form **category 2** and finally the students having answered “I agree” or “I strongly agree” were grouped in **category 3**. After running the test, we obtain the tree you can see in figure 3.13.

As you can see, the 90 subjects were first divided according to **question 40** which is “I would include a guestbook to let the reader sign it”. The first group disagrees with the statement and only consists of males (apart from one girl). The neutral group is mixed as well as the group having agreed. The neutral group can be divided into three groups according to **question 15** which is “I prefer websites in which there are many white spaces between the elements of the site (images, text etc)”. The first group disagrees with this statement and only consists of males (apart from one girl). The neutral group consists of more females than males (8 girls and 3 boys). The last group is a mixed group. If we go back to the group agreeing with question 40, we have a division into two other groups according to **question 17** which is “It doesn’t matter to me if the page is not well structured”. The first group disagrees and is a mixed group. The second agrees and only consists of males (apart from one female). The group disagreeing with question 17 is divided again into three groups according to **question 5** which is “I prefer a page for which I have to scroll down in order to see all text than a page in which all text is cluttered”. The first group disagrees and consists of more males than females (10 boys for 3 girls). The neutral group is a feminine group since there are 8 girls and 3 boys. The last group, the one agreeing, is a feminine group again with 9 girls and 3 boys. So, as you can see in figure 3.13, the boys have a more coherent behaviour since the non-mixed groups consist of males and we can’t see any of these non-mixed groups with females only.

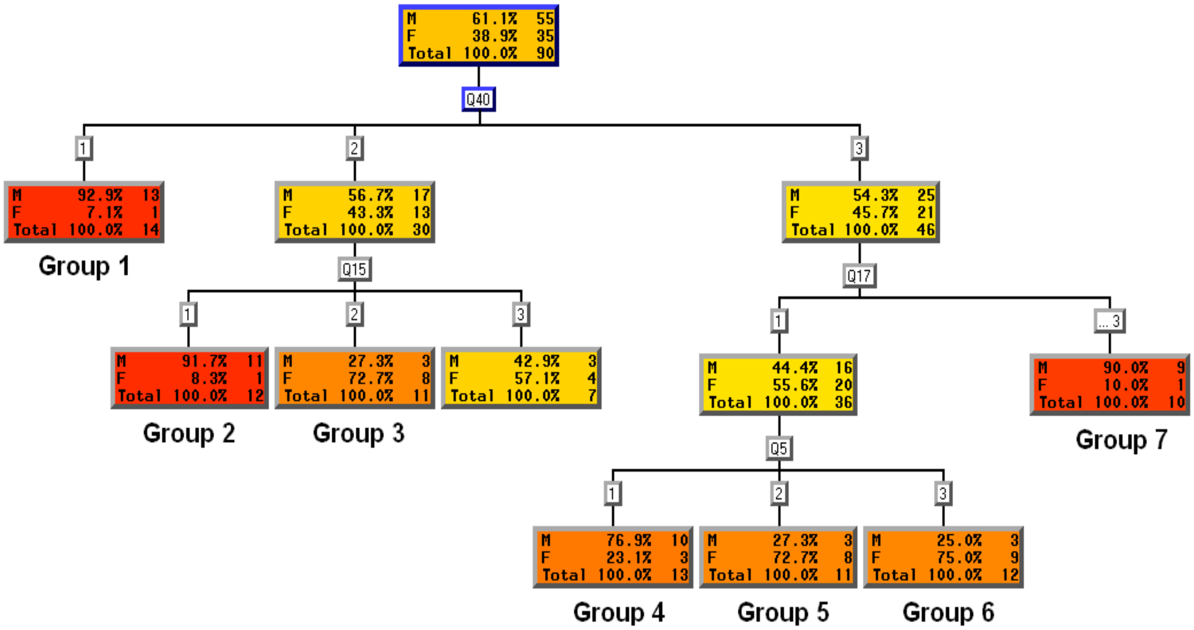


Figure 3.13: Classification tree - entropy method

In order to learn more about the subjects forming the leaves of the tree, we will analyse their answers to other significant questions, that is to say to **Q8** “I prefer menus you have to click through in order to achieve my goal", **Q12** "I prefer when many colours are used for text", **Q32** "I prefer when there are guiding tools that could help me navigate through the site", **Q33** "I would describe myself so that the reader can know who I am", **Q41** "I would include a counter to count the number of people having visited my site", **Q43** "I would put a picture of myself in my own site", **Q44** "I would put *many* pictures of myself in my own site", **Q45** "I would put jokey pictures of myself in my own site", **Q46** "I would put the pictures showing myself on the main page", **Q47** "I would insert pictures of my private life (family, pets, friends, leisure time etc) in my own site", **Q48** "I would insert *many* pictures of my private life (family, pets, friends, leisure time etc) in my own site" and finally **Q49** "I would put the pictures representing my private life on the main page".

- **Group 1's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
31 %	31 %	38 %	38,5 %	38,5 %	23 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
23 %	23 %	38 %	15,5 %	38,5 %	46 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
46 %	31 %	23 %	46 %	31 %	23 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
69 %	15,5 %	15,5 %	69 %	31 %	0 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
54 %	15 %	31 %	54 %	23 %	23 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
54 %	23 %	23 %	54 %	31 %	15 %

As you can notice in the table, the males of group 1 are quite equally distributed according to the different questions we focus on. They overall do not show any particular preference for the different questions apart from Q44 and Q45 for which they mostly (69 %) disagree. They also have a slight tendency (54 %) to disagree with Q46, Q47, Q48 and Q49. Regarding Q33, there are only 15 % to disagree, the rest being neutral or agreeing. We can summarize these observations with the following table (including the questions of figure 3.13):

Disagree	Q40, Q44, Q45, Q46, Q47, Q48, Q49	Don't disagree	Q33
Neutral	/	Are not neutral	/
Agree	/	Don't agree	/

In summary, the males of group 1 **won't include** a guestbook in their sites, guestook that is considered as a feminine feature in [MM98a]. They **won't put** many self-pictures in their own web homepage if they design it. If they put self-pictures, these **won't be** jokey as we could have thought by referring to [AM99a] in which the authors say they have not found any women's pages that use jokey pictures of themselves, as some men do. They **won't put** their self-pictures on the main page as we could have thought by reading the same article. Indeed, it is stated in the latter men are able to be confident about the way they present themselves and thus we could guess they wouldn't be reluctant to put their pictures on the main page (females wouldn't since they tend to be afraid to be judged by their physical appearance). We can notice the same for the pictures of their private lives. Indeed, the males of group 1 **won't insert pictures** of their private lives in their sites, and if they do so, there **won't be** many and these **won't appear** on the main page.

On the other hand, these boys **don't disagree** with the fact they would describe themselves so that the reader know who they are.

- **Group 2's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
18 %	55 %	27 %	64 %	18 %	18 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
27 %	18 %	55 %	9 %	36 %	45 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
18 %	18 %	64 %	18 %	45,5 %	36,5 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
73 %	18 %	9 %	27 %	64 %	9 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
45,5 %	36,5 %	18 %	45,5 %	36,5 %	18 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
36,5 %	36,5 %	27 %	36,5 %	36,5 %	27 %

A great majority (73 %) of the males of group 2 disagree with Q44, as the males of group 1. We can also notice 64 % disagree with Q12, are neutral with Q45 and agree with Q41. They have a slight tendency to be neutral with Q8 and to agree with Q32. Regarding Q33, few males disagree as the males of group 1, as well as for Q43. For Q46 and Q47, few agree. For the remaining questions, they are more or less equally distributed, thus showing no tendency for a particular answer. Let's summarize these observations in the following table:

Disagree	Q12, Q15, Q44	Don't disagree	Q33, Q43
Neutral	Q8, Q40, Q45	Are not neutral	/
Agree	Q32, Q41	Don't agree	Q46, Q47

In summary, the males of group 2 **do not like** when many colours are used for text or hypertext. They **don't like** spaced-out websites either. If they design their own web homepages, they **won't put** many pictures of themselves on their sites as the males of group 1.

However, they **don't care** about including a guestook or not in their sites, about menus you have to click through and about putting jokey pictures of themselves on their own sites.

On the other hand, they **do prefer** when there are guiding tools on a site to help them navigate. We could have thought they would give the opposite answer since getting help to navigate is considered as a feminine feature in [Khu04]. These boys **would also include** a counter to count the number of people having visited their sites. Again, we could have thought they would answer the opposite since counters are considered as a feminine feature in [MM98a].

As the males of group 1, they **don't disagree** with the fact they would describe themselves so that the reader know who they are. They **don't disagree** either with the fact they would put a self-picture. However, they show the same tendency as the males of group 1 regarding the the fact they **wouldn't put** their self-pictures on the main page nor **would they insert** pictures of their private lives.

- **Group 3's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
12,5 %	50 %	37,5 %	37,5 %	62,5 %	0 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
12,5 %	50 %	37,5 %	12,5 %	62,5 %	25 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
0 %	75 %	25 %	25 %	62,5 %	12,5 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
62,5 %	37,5 %	0 %	12,5 %	62,5 %	25 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
25 %	50 %	25 %	25 %	62,5 %	12,5 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
37,5 %	50 %	12,5 %	37,5 %	62,5 %	0 %

The females of group 3 are overwhelmingly neutral to all the questions we focus on apart from Q44 for which they overall disagree. The following table gives you a summary:

Disagree	Q44	Don't disagree	/
Neutral	Q8, Q12, Q15, Q32, Q33, Q40, Q41, Q43, Q45, Q46, Q47, Q48, Q49	Are not neutral	/
Agree	/	Don't agree	/

In summary, the females of group 3 **won't put** many pictures of their private lives if they design their web homepages.

They **don't care** about menus you have to click through. In [WBT02], they are 52 % to say they prefer pull-down menus. This is not confirmed here. They **don't care** about the use of many colours for (hyper)text. Again, we could reckon the opposite at first since in [WBT02], girls emphasised their preference for the use of colours. Spaced-out websites **do not make** any difference to them when 84 % of women said they prefer non-cluttered websites in [WBT02]. They don't care either about the presence of guiding tools and about giving a decription of themselves so that the reader know who they are. In [Khu04], it is stated women have difficulties to navigate in unfamiliar environments, thus guiding tools being a way to help them. But the females of group 3 **do not confirm** this finding. Unlike the males of groups 1 and 2, they **don't show** any interest in describing themselves. This can be considered as a lack of awareness to the reader. Let's remember awareness to the reader was considered as a feminine feature by [AM99a]. Including a guestbook or a counter **doesn't make** any difference to them. Again, we could have thought the contrary since in [MM98a], guestbooks and counters were mainly found on women's pages. Unlike the males of groups 1 and 2, they **don't care** about putting a self-picture on their sites, even jokey self-pictures, unlike the findings of [AM99a] in which the authors have not found any women's pages that use jokey pictures of themselves. Unlike the males of groups 1 and 2, they **don't mind** putting their self-pictures on the main page and inserting pictures of their private lives (and if so, many), and putting the latter on the main page.

- **Group 4's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
30 %	50 %	20 %	70 %	30 %	0 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
10 %	20 %	70 %	0 %	30 %	70 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
0 %	20 %	70 %	10 %	20 %	60 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
60 %	20 %	10 %	20 %	30 %	40 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
30 %	60 %	0 %	30 %	40 %	20 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
60 %	20 %	10 %	50 %	40 %	0 %

The males of group 4 overall have an opinion for most of the questions. They overwhelmingly disagree with Q12 (70 %) but agree with Q32, Q33, Q41 and Q43. Most disagree with Q44 and Q48, but are neutral to Q46. They have a slight tendency to be neutral with Q8 as the males of group 2. Let's note no male of this group agrees with Q49. For the remaining questions, they are more or less equally distributed, thus showing no tendency for a particular answer. Let's summarize this in a table:

Disagree	Q5, Q12, Q17, Q44, Q48	Don't disagree	/
Neutral	Q8, Q46	Are not neutral	/
Agree	Q32, Q33, Q40, Q41, Q43	Don't agree	Q49

In summary, the males of group 4 **do not prefer** pages for which they have to scroll down. They seem to prefer pages with cluttered text. They also **disagree** with the use of many colours for text and hypertext and they **do not like** when a page is not well structured. We could have thought they would be neutral since [Khu04] says they are able to navigate in unfamiliar environments, even when there is no structure. If they put self-pictures when designing their web homepages, they **won't put** many as the males of groups 1 and 2 or the females of group 3. If they insert pictures of their privates lives in their sites, they **won't put** many either as the males of group 1.

Menus you have to click through **do not make** any difference to them as the males of group 2. They **don't mind** putting their self-pictures on the main page unlike the males of group 1.

On the other hand, they **prefer** when there are guiding tools to help them navigate like the males of group 2 and unlike the findings from [Khu04]. If they design their web homepages, they **will describe** themselves to let the reader know who they are. Unlike the males of group 1, they **will include** a guestbook to let the reader sign it. Here again, we could have thought they would give a different answer since guestbooks are considered as a feminine items in [MM98a]. They **will include** a counter as well, what could be considered as feminine feature at first, according to [MM98a]. They **won't forget** to include a self-picture in their site.

These boys **don't agree** with the fact they would put the pictures of their private lives on the main page.

- **Group 5's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
25 %	50 %	25 %	25 %	37,5 %	37,5 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
0 %	0 %	100 %	25 %	75 %	0 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
0 %	12,5 %	87,5 %	37,5 %	37,5 %	25 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
75 %	12,5 %	12,5 %	62,5 %	25 %	12,5 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
37,5 %	50 %	12,5 %	37,5 %	37,5 %	25 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
37,5 %	62,5 %	0 %	87,5 %	0 %	12,5 %

Unlike the females of group 3, the females of group 5 (and of group 6, see next page) have different opinions for the selected questions. As you can see, all the females of group 5 agree with Q32. A great majority agrees with Q41 but disagrees with Q49. They overwhelmingly disagrees with Q44 but are neutral with Q33. They overall disagree with Q45 but are neutral with Q48. The females also have a slight tendency to be neutral with Q8 and Q46. For the remaining questions, they are more or less equally distributed, thus showing no tendency for a particular answer. The following table gives you a summary:

Disagree	Q17, Q44, Q45, Q49	Don't disagree	/
Neutral	Q5, Q8, Q33, Q46, Q48	Are not neutral	/
Agree	Q32, Q40, Q41	Don't agree	/

In summary, the females of group 5 **do not like** when a page is not structured since they then have more difficulties to navigate (see [Khu04]). They **wouldn't put** many pictures of themselves in their own sites like the females of group 3. Let's remember the authors of [Hes] have found men's pages tend to focus more on presenting a self-image to the viewer and that women's pages, in contrast, often exclude their own image. On the other hand, if the females of group 5 put self-pictures on their sites, these **won't be** jokey. This could be guessed since [AM99a] didn't find any jokey pictures on women's pages. Regarding the pictures representing their private lives, they **won't put** these on the main page.

Pages for which you have to scroll down and menus you have to click through **do not make** any difference to them like the females of group 3. Let's remember 84 % of females declared to prefer non-cluttered websites and 52 % prefer pull-down menus instead of menus you have to click through in [WBT02]. They **don't care** about writing a self-description to let the reader know who they are. Again, we could have thought they would agree if we consider this feature as a sign of awareness to the reader, which is then considered as a feminine feature in [MM98a]. Like the females of group 3, they **don't mind** putting their self-pictures on the main page as well as inserting many pictures of their private lives in their own sites.

These girls **prefer** when there guiding tools to help them, confirming the finding of [Khu04]. They **will also include** a guestbook and a counter when designing their web homepages, strengthening the findings of [MM98a].

- **Group 6's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
0 %	33 %	67 %	67 %	22 %	11 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
0 %	22 %	78 %	0 %	33 %	67 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
0 %	22 %	78 %	22 %	44,5 %	33,5 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
44,5 %	22 %	33,5 %	44,5 %	44,5 %	11 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
44,5 %	33,5 %	22 %	33,3 %	33,3 %	33,3 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
44,5 %	33,5 %	22 %	56 %	44 %	0 %

A great majority of the females of group 6 agrees with Q32 and Q41. They also overwhelmingly agree with Q8 and Q33, but disagrees with Q12. The females have a slight tendency to disagree with Q49 (56 %). We can also notice only 11 % agree with Q45. For the remaining questions, they are more or less equally distributed, thus showing no tendency for a particular answer. Let's summarize this with the following table:

Disagree	Q12, Q17, Q49	Don't disagree	/
Neutral		Are not neutral	/
Agree	Q5, Q8, Q32, Q33, Q40, Q41	Don't agree	Q45

In summary, the females of group 6 **do not agree** with the use of many colours for text and hypertext. This is interesting since in [WBT02], girls emphasised their preference for the use of colours. Like the females of group 5, they **do not like** when a page is not structured. If they insert pictures of their private lives in their own sites, they **won't put** these on the main page, like the females of group 5.

The females of this group **prefer** a page for which they have to scroll down than a page in which all text is cluttered, like 84 % of the females in [WBT02]. Unlike the females of groups 3 and 5, they **prefer** menus you have to click through. Like the males of group 4, they will describe themselves to let the reader know who they are. They also **prefer** when there are guiding tools to help them navigate (see [Khu04]) like the females of group 5. If they design their own web homepage, they **will include** a guestbook and a counter, like the females of group 5.

Let's note they **don't agree** with putting jokey self-pictures, showing more or less the same tendency as the females of group 5.

- **Group 7's profile**

Q8 - Disagree	**Q8 - Neutral**	**Q8 -Agree**	**Q12 - Disagree**	**Q12 - Neutral**	**Q12 -Agree**
0 %	67 %	33 %	0 %	67 %	33 %
Q32 - Disagree	**Q32 - Neutral**	**Q32 -Agree**	**Q33 - Disagree**	**Q33 - Neutral**	**Q33 -Agree**
0 %	33 %	67 %	33,5 %	22 %	44,5 %
Q41 - Disagree	**Q41 - Neutral**	**Q41 -Agree**	**Q43 - Disagree**	**Q43 - Neutral**	**Q43 -Agree**
0 %	33 %	67 %	44,5 %	22 %	33,5 %
Q44 - Disagree	**Q44 - Neutral**	**Q44 -Agree**	**Q45 - Disagree**	**Q45 - Neutral**	**Q45 -Agree**
44,5 %	22 %	33,5 %	11 %	44,5 %	44,5 %
Q46 - Disagree	**Q46 - Neutral**	**Q46 -Agree**	**Q47 - Disagree**	**Q47 - Neutral**	**Q47 -Agree**
44,5 %	33,5 %	22 %	33,5 %	44,5 %	22 %
Q48 - Disagree	**Q48 - Neutral**	**Q48 -Agree**	**Q49 - Disagree**	**Q49 - Neutral**	**Q49 -Agree**
44,5 %	33,5 %	22 %	22 %	44,5 %	33,5 %

The males of group 7 are overall neutral with Q8 and Q12, but agree with Q32 and Q41. Let's note only 11 % disagree with Q45. For the remaining questions, they are more or less equally distributed, thus showing no tendency for a particular answer. The following table gives you a summary:

Disagree		Don't disagree	Q45
Neutral	Q8, Q12	Are not neutral	/
Agree	Q17, Q32, Q40, Q41	Don't agree	/

In summary, the males of the last group **don't care** about menus you have to click through like the males of groups 2 and 4. They **don't care** either about an intensive use of colours for text unlike the males of groups 2 and 4 who disagree.

However, it **doesn't matter** to them if a page is not well structured, unlike the males of group 4. They **prefer** when there are guiding tools on a site like the males of groups 2 and 4 and unlike what we could expect after reading [Khu04]. If they design their own web homepages, they **will include** a guestbook and a counter, like the males of group 4 and unlike what we could expect after reading [MM98a].

Let's finally note they **don't disagree** with putting jokey self-pictures on their sites, unlike the males of group 1.

3.4 Conclusions

What we can learn from this chapter is the fact there is **no binary difference** between males and females, that is to say features that men would fit and that women wouldn't or vice versa. From the different analyses, we can bring out **different profiles among men as well as among women**. If we had only conducted a discriminant analysis, we would have stated girls have a **greater** preference for **guiding tools** and **menus you have to click through** than boys. On the other hand, these **tend more** to **describe themselves** than girls, thus being more confident according to [AM99a].

Actually, if we include these features in others, just as we proceeded for the segmentation tree, we can see it is not so obvious as we will explain regarding the way they would design their own sites and their preferences in web design in general.

- **If they had to design their own web homepages...**

The **first profile** among the **males** we could put forward is the one of a **reserved** group of males. They don't seem to show any sign of awareness to the reader (guestbook, counter, self-description). They do not like to include many pictures of themselves, and the few pictures they can show won't be jokey and won't be found on the main page. They won't insert any picture of their private lives and if they do so, there won't be many and these won't be on the main page. Let's remember we had described the main page like a spot representing the private sphere and where private elements were displayed.

However, the **second group of males** is **friendlier**. They are not reluctant to show signs of awareness to the reader (guestbook, counter, self-description). They accept to put self-pictures on their sites, but not many. And these pictures can be jokey. But there is still some restraint regarding the private sphere since like the first group of males, they wouldn't put their self-pictures on the main page nor would they insert pictures of their private lives.

The **third group of males** is a bit **more extrovert** than the males of the second group. Of course, they won't put many self-pictures nor many pictures representing their private lives. But the difference with the other groups of males is they don't mind putting their self-pictures on the main page at all. Let's note they will certainly include at least one self-picture. They show more signs of awareness than the other groups of males since they will absolutely include a guestbook and a counter and will certainly describe themselves so that the reader know who they are. For the private pictures, they are still a bit reluctant to put these on the main page like the other groups.

The **last group of males** don't give **many details** about the way they would design their sites. Like the males of the second and third groups, they will include a guestbook and a counter, and they are not reluctant to put some jokey self-pictures. But this is the only information they provide regarding their web homepages.

The **first group of females** is **atypical** since they don't care about the way they would design their own web homepages. They don't seem to really know what they would like to include or not. They don't care about describing themselves nor putting at least one self-picture. They can't say if they would prefer to include serious pictures or jokey pictures (if they decide to include self-pictures of course). They don't care if their self-pictures are on the main page and if they would include pictures of their privates lives and in which quantity and if they would put these on the main page. They don't have any preference for putting a guestbook and a counter. Let's note the females of this group are sure about one thing: they won't put many pictures of their private lives (tendency of the two following groups of females). Of course, we could wonder why these girls are neutral with most of the questions. It would be interesting to conduct further investigation in order to know whether they do not care at all or if it is because they do not dare to give their opinion etc.

The **second group of females** seems to **focus more on giving much information** about them without putting many self-pictures and especially jokey self-pictures. They don't mind putting their self-pictures on the main page. But regarding the private pictures, they won't put these on the main page. Their private sphere is thus composed of the pictures representing their private lives, and putting these on the main page is like exposing their lives to the whole world. They don't care about writing a self-description, since they seem to prefer to tell who they are by pictures representing their lives (they can even be numerous since they don't care about the quantity). They will include a guestbook and a counter in their sites, thus showing awareness to the reader. So in summary, these girls really want to have an interaction, to share with the reader who they are by pictures representing their lives. They also expect the reader to take part in this exchange.

The **last group of females** do **not say** much about their preferences regarding the way they would design their own websites like the last group of males. However, we know that if they put self-pictures, they don't agree with putting jokey self-pictures, showing more or less the same opinion as the females of the second group. Like the females of the latter, they won't put the pictures representing their private lives on the main page and they will show some signs of awareness to the reader by including a guestbook and a counter.

- **Their preferences in web design...**

Let's first note that **most of the groups do prefer when guiding tools are present** on a site, apart from the first group of males and the first group of females who do not have any opinion. So, as the females, **boys need to get help** to navigate.

The **first group of males** does **not give** any opinion about the topic. Regarding the **males of the second group**, we know they do not like when many colours are used for text as we could expect from boys (the cliché being boys do not like to use many colours like girls). They don't like spaced-out websites either, what we could expect since this feature is the prerogative of women according to [WBT02]. They don't care about menus you have to click through.

As the males of the second group, the **males of the third group** do **not like** when many colours are used for text. They do not like pages for which they have to scroll down, thus going along with the fact the males of the second group do not like spaced-out websites. The males of the third group do not like when a page is not well structured. We could have thought they wouldn't care since [Khu04] says they can navigate in any new environment without difficulties. This preference for structure thus goes along with the presence of guiding tools.

The **males of the last group** are **more tolerant** than the two previous groups of males regarding the use of many colours for text since they don't care about it. They do not show any reluctance if the page is not well structured. We can assume they think they will be able to manage with non-structured pages if there are guiding tools to help them navigate. Regarding the menus you have to click through, they don't care like the second and third group of males.

The **first group of females** do **not care** about the use of many colours for text (like the males of the last group). We could have thought they would have answered they prefer when many colours are used after reading [WBT02]. They don't care about spaced-out websites either. We could have thought they would prefer spaced-out websites after reading the same paper. Like the boys, they don't care about menus you have to click through. We could have thought they wouldn't like this kind of menus since you have to be better at spatial skills (like the boys) since they are not pull-down menus.

Like the males of the third group, the **second group of females** do **not like** when a page is not structured, thus going along with [Khu04]. They don't care if they have to scroll down or if they have to read cluttered text. We could have thought they would prefer pages to scroll down after reading [WBT02]. Like the other groups, they don't care about menus you have to click through.

The **females of the last group** do **not like** when many colours are used for text like the males of the second and third groups (and unlike the females interviewed in [WBT02]). Like the females of the second group, they don't like when a page has no structure. This must be more difficult for them to navigate. They do prefer a page for which they have to scroll down, like the females of [WBT02]. They do also prefer menus you have to click through. They must be able to situated themselves easily in unfamiliar environments then.

- **When we compare some groups of males with some groups of females...**

If we focus on the **second group of females** and the **third group of males**, we will notice they share the same characteristic: they are more extrovert than the other groups of the same gender. Both of them state they don't like non-structured pages. They won't put many pictures of themselves on their sites. Both are neutral regarding menus you have to click through and don't mind putting their self-pictures on the main page. Both will include a guestbook and a counter in their sites. **However**, the **males** do not like pages for which they have to scroll down whereas the **females** are neutral regarding this feature. The males wouldn't insert many pictures of their private lives whereas the females wouldn't mind doing so.

The **last group of females** and the **last group of males** have something in common: they don't give many details about the way they would design their sites. None of the groups is sensitive to the presence of guiding tools on a site. They are neutral regarding the possibility of inserting a guestbook and a counter in their sites. **However**, the **females** don't like when many colours are used for text and hypertext. The **males** don't seem to care about an intensive use of colours. Besides, the females do not like non-structured pages whereas the males do. Regarding the menus you have to click through, the girls prefer these whereas the boys don't have any particular opinion regarding the topic. But for the jokey self-pictures, the girls don't agree to put any of these on their sites whereas the boys tend to be neutral or to agree.

If we compare the **males of the last group** with the **females of the first group**, we will notice both don't care about menus you have to click through. They don't mind either about an intensive use of colours for text. **But** regarding their preference for guiding tools, the **girls** are neutral whereas the **boys** do prefer when the latter can be used. Regarding the inclusion of a guestbook and a counter in their own sites, the girls are neutral but the boys do agree. The girls don't care about putting jokey self-pictures on their sites whereas the boys do not agree.

Now if we focus on the **males of the third group** and the **females of the last group**, we will notice that both do not like an intensive use of colours for text and they also dislike non-structured pages. They are both ready to describe themselves on their own sites. They will surely include a guestbook and a counter. **However**, the **boys** do not like pages for which they have to scroll down whereas the **girls** do. Regarding menus you have to click through, the boys don't care about using these but the girls prefer to use this type of menus.

The last comparison focus on the **second group of males** and the **last group of females**. Both dislike an intensive use of colours for text. Regarding the inclusion of a counter, they will put one on their websites. **However**, the **boys** don't have any particular opinion regarding menus you have to click through when the **girls** prefer these. The males are also neutral concerning the inclusion of a guestbook in their websites whereas the girls will put one. For the jokey self-pictures, the males don't care whereas the females do not agree to put this kind of self-pictures. Finally, the boys tend to be neutral or to agree like the girls when it comes to describing themselves on their own websites.

Conclusion

From this thesis, we have learned gender differences applied to an academic context is not an easy topic. In the first chapter devoted to the literature review, we have realized gender differences in a web design context is an issue preoccupying many sociologists. Many of them conduct investigations in this field. Unfortunately, as one of them says, there is always a study refuting the findings of a previous one.

In the second and third chapters, we tried to figure out whether we could bring out the same tendencies as the ones described by the sociologists in chapter 1. We have also enhanced the characteristics of this first chapter with new features in order not to focus only on the presentation aspect. We have been able to confirm some findings of previous research or to highlight opposite behaviours to those described by the sociologists. It was not always possible to conclude for each feature after having run the statistical tests. Nevertheless, the different analyses we have conducted have allowed us to identify predominantly common profiles and types of distinct behaviours with feminine-higher or masculine-higher tendencies.

However, we have to keep in mind our sample is small: thirty academics divided into fifteen males and fifteen females. That is the reason why we can't generalize and apply our findings to the whole population of male and female academics. Therefore, it would be interesting to go futher with this study and try to conduct it on an international level as well, that is to say in different cultures. Thus we would enhance the scope of this work in order to include the cultural factor.

We could also think of going into the subject in greater depth by applying our study to academics who do not belong to IT departments. Indeed, the basic idea in our study was to choose IT professors in order to have more chance these design their sites on their own since they had the skills to do so. On the face of it, designing a site must be less easy for arts professors. But of course, this should be examined in depth.

Another idea would be that each team who would continue on this task consists of specialists forming an heterogeneous panel: sociologists, statisticians, psychologists, amateur and professional web designers etc. Interviews should also be conducted to examine the results obtained with statistical methods in greater depth and to explore the reasons why the person has designed the website in this way and not in that way.

Lastly, we could also think of analyzing the way professional designers design their websites according to the gender of the final user. So the question would become: "Is there a specific way of designing *for* women and *for* men?"

Bibliography

[Ago04] Denise Agosto. Using gender schema theory to examine gender equity in computing: a preliminary study. *Journal of Women and Minorities in Science and Engineering*, 10:37–53, 2004.

[Ahu02] Manju Ahuja. Women in the information technology profession: a literature review, synthesis and research agenda. *European Journal of Information Systems*, 11:20–34, 2002.

[Akr95] Madeleine Akrich. User representations: Practices, methods and sociology. In *A.Rip, T.Misa, J.Schot (Eds.): Managing Technology in Society: The Approach of Constructive Technology Assessment*, pages 167–184. Pinter Publishers, London (UK) and New York (USA), 1995.

[AM99a] Jill Arnold and Hugh Miller. Gender and web home pages. In *CAL99 Virtuality in Education Conference*, London, UK, March 1999. http://ess.ntu.ac.uk/miller/cyberpsych/cal99.htm (Last visit on December 22nd, 2004).

[AM99b] Jill Arnold and Hugh Miller. The hypertext home: images and metaphors of home on world wide web homepages. In *Design History Society Home and Away Conference*, Nottingham, UK, September 1999. http://ess.ntu.ac.uk/miller/cyberpsych/homeweb.htm (Last visit on December 22nd, 2004).

[AM00] Jill Arnold and Hugh Miller. Same old gender plot? women academics' identities on the web. In *Cultural Diversities in/and Cyberspace Conference*, May 2000. http://ess.ntu.ac.uk/miller/cyberpsych/gendplot.htm (Last visit on December 22nd, 2004).

[AM01a] Jill Arnold and Hugh Miller. Breaking away from home? cyberculture and gendered academic identities on the web. In *Constructing Cyberculture(s): Performance, Pedagogy and Politics in Online Spaces*, Maryland, USA, April 2001. http://ess.ntu.ac.uk/miller/cyberpsych/maryland.htm (Last visit on December 24th, 2004).

[AM01b] Jill Arnold and Hugh Miller. Academic masters, mistresses and apprentices: gender and power in the real world of the web. *Mots pluriels*, 19, October 2001. http://www.arts.uwa.edu.au/MotsPluriels/MP1901jahm.html.

[Bem81] S. Bem. Gender schema theory: a cognitive account of sex typing. *Psychological Review*, 88:354–364, 1981.

[Bro98] Mark Brosnan. The impact of psychological gender, gender-related perception, significant others, and the introducer of technology upon computer anxiety in students. *Journal of Educational Computing Research*, 18(1):63–78, 1998.

[BS00] Ellen Balka and Richard Smith. *Women, Work and Computerization: Charting a Course to the Future*. Kluwer Academic Publishers, Boston, Dordrecht, London, 2000.

[But93] Judith Butler. Gender is burning: Questions of appropriation and subversion. In *N.Mirzoeff (Ed.): The Visual Culture Reader*, pages 448–462. Routledge, London (UK) and New York (USA), 1993.

[Cha99] John Charlton. Biological sex, sex-role identity and the spectrum of computing orientations: A re-appraisal at the end of the 90s. *Journal of Educational Computing Research*, 21(4):393–412, 1999.

[Chi96] Patricia Chisholm. Cyber-sorority: Women begin to feminize the net. *MacLean's*, 109(47), November 1996.

[CJ99] Janet Carter and Tony Jenkins. Gender and programming: What's going on? In *ACM SIGCSE Bulletin, Proceedings of the 4th annual SIGCSE/SIGCUE on Innovation and technology in computer science education 31(3)*, pages 1–4, June 1999.

[CJ02] Janet Carter and Tony Jenkins. Gender differences in programming? In *Proceedings of the 7th annual conference on Innovation and technology in computer science education*, pages 188–192, June 2002.

[CMK98] Cynthia Ching, Sue Marshall, and Yasmin Kafai. Give girls some space: Gender equity in collaborative technology activities. In *Conference Proceedings of the 19th Annual National Educational Computing Conference*, pages 67–78. Int. Soc. Technol. Educ, Eugene, OR, USA, 1998.

[Coc88] Cynthia Cockburn. *Machinery of Dominance: Women, Men, and Technical Know-How*. Northeastern University Press, Boston, USA, 1988.

[Col98] Alix Collard. *Comparaison de deux proportions: une approche bayésienne pratique dans le contexte des tables deux fois deux avec petite taille d'échantillon*. PhD thesis, Facultés universitaires Notre-Dame de la Paix Namur, Namur, Belgique, 1998. Publié par les Presses Universitaires de Namur.

[Cro99] Jennifer Croissant. Engendering technology: Culture, gender, and work. In *1999 International Symposium on Technology and Society Women and Technology: Historical, Societal, and Professional Perspectives*, pages 276–281, 1999. Proceedings. Piscataway: IEEE.

[Dag98] Pierre Dagnelie. *Statistique théorique et appliquée*, volume Tome 2. De Boeck Université, Bruxelles, Belgique, 1998.

[DEH02] Yvonne Dittrich, Sara Eriksen, and Christina Hansson. Pd in the wild: Evolving practices of design in use. In *T.Binder, J.Gregory, I.Wagner (Eds.): PDC 02 Proceedings of the Participatory Design Conference*, pages 124–134, Malm , Sweden, June 2002.

[DHL00] A. Durndell, Zsolt Haag, and Heather Laithwaite. Computer self efficacy and gender: A cross cultural study of scotland and romania. In *Personality and Individual Differences 28(6)*, pages 1037–1044, June 2000.

[dL83] Jean de Lagarde. *Initiation à l'analyse des données*. Dunod, Paris, France, 1983.

[Doe02] Nicola Doering. Personal home pages on the web: A review of research. *JCMC*, 7, April 2002. http://www.ascusc.org/jcmc/vol7/issue3/doering.html (Last visit on November 29th 2004).

[eJP90] Brigitte Escofier et Jérôme Pagès. *Analyses factorielles simples et multiples.* Dunod, 2ème édition edition, Paris, France, 1990.

[eRW91] Tomas Wonnacott et Ronald Wonnacott. *Statistique.* Economica, 4ème édition edition, Paris, France, 1991.

[Eys41] H. Eysenck. A critical and experimental study of color preferences. *The American Journal of Psychology*, 54:385–394, 1941.

[FC00] Julie Fisher and Annemieke Craig. Considering the gender of your web audience. In *Women, Work and Computerization: Charting a Course to the Future*, pages 164–173. Kluwer Academic Publishers, Boston, Dordrecht, London, 2000.

[FMM97] A. Fisher, J. Margolis, and F. Miller. Undergraduate women in computer science: Experience, motivation and culture. In *SIGCSE*, volume 29(1), pages 106–110, 1997.

[Gef00] D. Gefen. Gender differences in the perception and adoption of e-mail and computer-mediated communication media: a sociolinguistics approach. In *A.Kent, L.Lancour (Eds.): The Encyclopedia Of Library And Information Science.* M Dekker, New York, USA, 2000.

[GK99] D. Graves and M. Klawe. Supporting learners in a remote computer-supported collaborative learning environment: The importance of task and communication. In *Proceedings CSCL*, Toronto, Canada, 1999.

[GM00] Cecilia Gorriz and Claudia Medina. Engaging girls with computers through software games. In *Communications of the ACM 43(1)*, pages 42–49, January 2000.

[God04a] Vincent Godard. Fiche mémo n°1. du cours de dea d'analyse de données: Présentation de l'analyse des données multivariées et du logiciel spad. Département de Géographie, Université de Paris 8, Novembre 2004. http://margaux.ipt.univ-paris8.fr/vgodard/enseigne/dea/memodea/mem01dea.htm (Last visit on July 15th 2005).

[God04b] Vincent Godard. Fiche mémo n°2. du cours de dea d'analyse de données: L'analyse en composantes principales (acp). Département de Géographie, Université de Paris 8, Novembre 2004. http://margaux.ipt.univ-paris8.fr/vgodard/enseigne/dea/memodea/mem02dea.htm (Last visit on July 15th 2005).

[God04c] Vincent Godard. Fiche mémo n°4. du cours de dea d'analyse de données: Les classifications hiérarchiques. Département de Géographie, Université de Paris 8, Novembre 2004. http://margaux.ipt.univ-paris8.fr/vgodard/enseigne/dea/memodea/mem04dea.htm (Last visit on July 15th 2005).

[God05] Vincent Godard. Fiche mémo n°3. du cours de dea d'analyse de données: L'analyse factorielle des correspondances (afc). Département de Géographie, Université de Paris 8, Janvier 2005. http://margaux.ipt.univ-paris8.fr/vgodard/enseigne/dea/memodea/mem03dea.htm (Last visit on July 15th 2005).

[Gre95] K. Green. Blue versus periwinkle: Color identification and gender. *Perceptual and Motor Skills*, 80(1):21–32, 1995.

[GS59] J. Guilford and P. Smith. A system of color-preferences. *The American Journal of Psychology*, 73(4):487–502, 1959.

[GS97] D. Gefen and D. Straub. Gender differences in perception and adoption of e-mail: an extension to the technology acceptance model. In *MIS Quarterly*, volume 21, pages 389–400, 1997.

[Gui34] J. Guilford. The affective value of color as a function of hue, tint, and chroma. *Journal of Experimental Psychology*, June 1934.

[GvHNP97] A. Greenhill, L. von Hellens, S. Nielson, and R. Pringle. Australian women in it education: Multiple meanings and multiculturalism. In *Proceedings of the Sixth International IFIP WG 9.1 Conference on Women, Work and Computerisation*, pages 387–397. Springer, Bonn, Germany, May 1997.

[Har86] Sandra Harding. *The Science Question in Feminism.* Cornell University Press, Ithaca (USA) and London (UK), 1986.

[Her94] Susan Herring. Gender differences in computer-mediated communication: bringing familiar baggage to the new frontier. In *American Library Association annual convention*, Miami, USA, June 1994. http://www.cpsr.org/cpsr/gender/herring.txt (Last visit on September 29th 2004).

[Her96] Susan Herring. Posting in a different voice: Gender and ethics in computer-mediated communication. In *C.Ess (Ed.): Philosophical Perspectives on Computer-Mediated Communication*, pages 115–146. State University Press, New York, USA, 1996.

[Hes] Mickey Hess. A nomad faculty: English studies' online representations of work, product and workplace. http://www.louisville.edu/˜mshess01/nomadfaculty.htm (Last visit on December 22nd 2004).

[HMS02] Sue Herring, Anna Martinson, and Rebecca Scheckler. Designing for community: The effects of gender representation in videos on a web site. In *Proceedings of the 35th Hawaii International Conference on System Sciences, IEEE Comput. Soc*, pages 1100–1111, 2002. http://dlib2.computer.org/conferen/hicss/1435/pdf/14350117.pdf (Last visit on August 30th, 2002).

[HR00] Tove Hapnes and Bente Rasmussen. New technology increasing old inequality? In *Women, Work and Computerization: Charting a Course to the Future. IFIP TC9 WG9.1 Seventh International Conference*, pages 241–249, Vancouver, Canada, June 2000.

[HRTT99] A. Huang, A. Ring, S. Toich, and T. Torres. Girls' school to encourage achievement in science, math, and computers, 1999. http://www-cse.stanford.edu/classes/cs201/Projects/gender-gap-in-education/page12.htm (Last visit on August 30th 2002).

[HS04] Geoffrey Hubona and Gregory Shirah. The gender factor performing visualization tasks on computer media. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Hawaii, 2004. http://doi.ieeecomputersociety.org/10.1109/HICSS.2004.1265264 (Last visit on December 3rd 2004).

[iDGCoc] NSF-Project investigators: Denis Gürer and Tracy Camp. Investigating the incredible shrinking pipeline for women in computer science. Final Report NSF Project 9812016, started 1998, http://www.acm.org/women/pipeline-finalreport-ver-2.doc. (Last visit on August 30th 2002).

[Kay92] R. Kay. Understanding gender differences in computer attitudes, aptitude, and use: An invitation to build theory. *Journal of Research on Computing in Education*, 25:159–171, 1992.

[KF94] Nancy Kaplan and Eva Farrell. Weavers of webs: A portrait of young women on the net. *The Arachnet Electronic Journal on Virtual Culture*, 2(3), July 1994. http://raven.ubalt.edu/staff/kaplan/weavers/weavers.html.

[Kho] Natalia Khouw. The meaning of color for gender. http://www.colormatters.com/khouw.html (Last visit on July 31st 2005).

[Khu04] Zeina Khumush. Personal theory of design. University of Technology of Sydney, September 2004.

[Kul76] R. Kuller. The use of space–some physiological and philosophical aspects. In *Third International Architectural Psychology Conference*, Strasbourg, France, 1976.

[LBR02] Andrew Large, Jamshid Beheshti, and Tarjin Rahman. Gender differences in collaborative web searching behavior: an elementary school study. *Information Processing and Management*, 38:424–443, 2002.

[Lie03] Merete Lie. *He, She and IT Revisited: New Perspectives on Gender in the Information Society.* Gyldendal Akademisk, Oslo, Norway, 2003.

[LS03] Merete Lie and Knot Holtan Sorensen. *Strategies of Inclusion: Gender in the Information Society. Vol I: Experiences from public sector initiatives.* NTNU, Norway, 2003.

[MFM00] Jane Margolis, Allan Fisher, and Faye Miller. The anatomy of interest: Women in undergraduate computer science. *Womens Studies Quarterly*, 28(1/2), Spring/Summer 2000.

[Mil95] Hugh Miller. The presentation of self in electronic life: Goffman on the internet. In *Embodied Knowledge and Virtual Space Conference*, London, UK, June 1995. http://ess.ntu.ac.uk/miller/cyberpsych/goffman.htm (Last visit on November 29th 2004).

[Mit97] Tom Mitchell. *Machine learning.* McGraw-Hill International Editions, Singapore, Malaysia, 1997.

[MLM01] Amanda Mitra, Betty LaFrance, and Sandra McCullough. Differences in attitudes between women and men toward computerization. *Journal of Educational Computing Research*, 25(3):227–244, 2001.

[MM98a] Hugh Miller and Russell Mather. The presentation of self in www home pages. In *IRISS '98: Conference Papers*, Bristol, UK, March 1998. http://www.sosig.ac.uk/iriss/papers/paper21.htm (Last visit on December 22nd, 2004).

[MM98b] Janet Morahan-Martin. The gender gap in internet use: Why men tend to use the internet more than women a literature review. *CyperPsychology and Behaviour*, 1(1):3–10, Spring 1998.

[MS64] J. McInnis and J. Shearer. Relationship between color choices and selected preferences for the individual. *Journal of Home Economics*, 56:181–187, 1964.

[NC97] Lori Nelson and Joel Cooper. Gender differences in children's reactions to success and failure with computers. *Computers in Human Behaviour*, 13(2):247–267, May 1997.

[Newtm] BBC News. Education software sex bias 'puts girls off', http://news.bbc.co.uk/1/hi/education/200286.stm. (Last visit on August 30th 2002).

[NF02] M. Noirhomme-Fraiture. Cours de probabilités et statistique, 2002.

[ORvS04] Nelly Oudshoorn, Els Rommes, and Irma van Slooten. *Strategies of Inclusion: Gender in the Information Society. Vol III: Surveys of Women's User Experience.* NTNU, Norway, 2004.

[PH70] E. Pearson and H. Hartley. *Biometrika Tables for Statisticians.* Cambridge University Press, Cambridge, UK, 1970.

[PL00] David Passing and H. Levin. Gender preferences for multimedia interfaces. *Journal of Computer Assisted Learning*, 16(1):64–71, March 2000. http://www.passig.com/pic/PrefInterface.htm (Last visit on August 30th, 2002).

[Pla67] G. Plater. *Adolescent preferences for fabric, color, and design on usual task.* PhD thesis, Indiana State College, Terre Haute, Indiana, 1967.

[Rad90] D. Radeloff. Role of color in perception of attractiveness. *Perceptual and Motor Skills*, 71:151–160, 1990.

[Rod97] Michelle Rodino. Breaking out of binaries: Reconceptualizing gender and its relationship to language. *JCMC*, 3, December 1997. http://www.ascusc.org/jcmc/vol3/issue3/rodino.html (Last visit on October 1st 2004).

[Rom02] Els Rommes. Creating places for women on the internet: The design of a women's square in a digital city. *The European Journal of Women's Studies*, 9(4):400–429, 2002.

[RvOO99] Els Rommes, Ellen van Oost, and Nelly Oudshoorn. Gender in the design of the digital city of amsterdam. *Information Communication Society*, 2(4):476–495, 1999.

[Sap90] Gilbert Saporta. *Probabilités, analyse des données et statistique.* Technip, Paris, France, 1990.

[SE92] Silverman and Eals. Spatial sex differences: Evolutionary theory and data. *The adapted mind: Evolutionary psychology and the generation of culture*, pages 487–503, New York, USA, 1992.

[Sie88] Sidney Siegel. *Nonparametric statistics for the behavioral sciences.* MacGraw-Hill, 2nd edition edition, New York, USA, 1988.

[Sim01] Steven Simon. The impact of gender on web sites: An empirical study. *The DATA BASE for Advances in Information Systems*, 32(1):18–37, Winter 2001.

[SMM01] P. Schumacher and J. Morahan-Martin. Gender, internet and computer attitudes and experiences. *Computers in Human Behaviour*, 17:95–110, 2001.

[Spr89] Peter Sprent. *Applied non-parametric statistical methods.* Chapman and Hall, London, UK, 1989.

[SS80] Thomas Santner and Mark Snell. Small-sample confidence intervals for p1-p2 and p1/p2 in 2x2 contingency tables. *Journal of the American Statistical Association*, 75(370):386–394, June 1980.

[TCB78] L. Thomas, A. Curtis, and R. Bolton. Sex differences in elicited color lexicon size. *Perceptual and Motor Skills*, 47:77–78, 1978.

[TD93] J. Todman and G. Dick. Primary children and teachers attitudes to computers. *Computers in Education*, 20:199–203, 1993.

[TP90] Sherry Turkle and Seymour Papert. Epistemological pluralism: Styles and voices within the computer culture. *Signs*, 16(1):128–157, 1990.

[Tur95] Sherry Turkle. *Life on the Screen: Identity in the Age of the Internet.* Simon and Schuster, New York, USA, 1995.

[WBT02] Ina Wagner, Andrea Birbaumer, and Marianne Tolar. Gender gap in computer science: cultural and psychological factors. In *Widening Women's Work in information and Communication Technology: conceptual framework and state of the art (Deliverable N° 1)*, pages 67–85. European Commission, Brussels, Belgium, 2002.

[Whi84] A. Whitfield. Individual differences in evaluation of architectural colour: Categoriation effects. *Perceptual and Motor Skills*, 59:183–186, 1984.

[Yoe] Chang Yoekyong. Women's strategy in cyberspace: More political imagination. http://lmdedia.nodong.net/1999/archive/e21.htm (Last visit on September 29th 2004).

Appendix A

Site list

Here you will find the list of the selected sites.

A.1 Male academics

1. http://scott.bradcentral.com/
2. http://www.it.jcu.edu.au/~alan/
3. http://www.comp.mq.edu.au/~len/personal/personal.html
4. http://www.sims.monash.edu.au/staff/darnott/
5. http://sky.fit.qut.edu.au/~russells/
6. http://goanna.cs.rmit.edu.au/~jah/
7. http://www.cs.adelaide.edu.au/~esser/
8. http://uob-community.ballarat.edu.au/~cnelson/index.html
9. http://www.cs.mu.oz.au/~lapark/main.html
10. http://www.cs.usyd.edu.au/~deveritt/
11. http://www-staff.it.uts.edu.au/~gerry/
12. http://www.csse.uwa.edu.au/~chris/
13. http://www.uow.edu.au/~phillip/
14. http://www-staff.it.uts.edu.au/~alan/
15. http://www.arch.usyd.edu.au/~john

A.2 Female academics

1. http://members.westnet.com.au/merwood/mer/index.html
2. http://srvcns.it.jcu.edu.au/~marion/
3. http://www.comp.mq.edu.au/~anabel/
4. http://www.sims.monash.edu.au/staff/klynch/
5. http://sky.fit.qut.edu.au/~christir/
6. http://goanna.cs.rmit.edu.au/~liz/index.html
7. http://www.cs.adelaide.edu.au/~cheryl/
8. http://uob-community.ballarat.edu.au/~kkeogh/index.html
9. http://www.cs.mu.oz.au/~linda/
10. http://www.veale.com.au/kylie/index.htm
11. http://www-staff.it.uts.edu.au/~valerie/
12. http://www.csse.uwa.edu.au/~robyn/
13. http://www.itacs.uow.edu.au/school/staff/katina/
14. http://www-staff.it.uts.edu.au/~janeb/
15. http://www.arch.usyd.edu.au/~mary

Appendix B

Site evaluation and computation of the statistics

B.1 Site evaluation

In this section you will find two java programs: the first one assessing the number of images on a page and the second one assessing the number of URLs contained in a site.

In order to retrieve the data to perform the computation in the Interval class, the following class is used to count the number of images that can be found in an HTML page of a website.

```
package countImages;

import java.io.BufferedInputStream;
import java.io.DataInputStream;
import java.io.IOException;
import java.net.MalformedURLException;
import java.net.URL;
import java.net.URLConnection;
import jericho.*;

public class CountImages {

  /**
   * This method returns a String corresponding to the HTML code of the page
   * which URL is given as a parameter.
   * Parameter of the method:
   * - urlString (String) : the URL of the HTML page for which we have to retrieve
   *                        the code.
   */
  public static String getHtmlCodeFromUrl(String urlString)
  {

    URL theURL = null;
    try {

      theURL = new URL(urlString);
    } catch (MalformedURLException e1) {

      e1.printStackTrace();

    }
    //Transformation of the String containing the URL as a URL object that can be reused.

    URLConnection conn = null;
    DataInputStream data = null;
    String line;
    StringBuffer buf = new StringBuffer();
    //Declaration and initialization of the variables which are necessary to open
    //the connection to the page.
```

```
    try {

      conn = theURL.openConnection();
      conn.connect();
      //We open a connection to the URL we have previously obtained and
      //we retrieve the HTML code.

      data = new DataInputStream(
                 new BufferedInputStream(conn.getInputStream())
      );
      while ((line = data.readLine()) != null) {
        buf.append(line + "\n");
      }
      //The HTML code is retrieved through a stream (DataInputStream) that has been opened
      //thanks to the connection.  Each "line" is inserted in a
      //StringBuffer.

      data.close();
      //We close the stream.

    } catch (IOException e) {

      return "IO Error";
      //The method returns an error message if an exception occurs
      //when using the input and output stream, when we open
      //the latter or when we open the
      //connection.

    }

    return buf.toString();
    //We convert the StringBuffer into a String and we send this object to the
    //calling method.

  }
```

```
/**
 * Main method of the class, called via
 * the command line/the shell.
 * The user has to enter the URL to the page of which he/she wants to retrieve
 * the HTML code.
 * Method parameter:
 * - args (String table) : the parameters the user has given
 *                              when calling this
 *                              method
 */
public static void main (String args[])
{

  if (args.length == 0) {

    System.out.println(
        "Hi! Put an URL without \"http://\" as argument"
    );
    //If no parameter is given, the method reminds the user
    //how to use it.

  }
  else {

    String contents = getHtmlCodeFromUrl("http://"+ args[0]);
    //When calling getHtmlCodeFromUrl, the method retrieves the HTML code
    //of the page included in the site corresponding to the specified URL.

    if (contents.equals("IO Error")) {

      System.out.println();
      System.out.println("---> Download of page failed");
      System.out.println();
      //An"IO Error" means the URL is incorrect or that an error
      //occured during the retrieval process
      //of the HTML code of the page.

    }
    else {

      Source source = new Source(contents);
      //Using the functionalities of the "jericho" package, the content
      //of the page is converted into a Source object, that can easily be used
      //to retrieve HTML elements.
```

```
        int nbrImages = source.findAllElements("img").size();
        //The findAllElements method returns a list of elements of
        //the HTML page which are images (HTML img tag).
        //We only focus on their number, that is why
        //only record the size of the list.

        System.out.println();
        System.out.println("---> " + nbrImages + " images that has/have been found for "
                           + args[0]);
        System.out.println();
        //The information we are interested in is displayed in the command line/the
        //user shell.

      }

    }

  }

}
```

The following code is used to retrieve URLs in an HTML page, located on an Internet site. This code is very similar to the one used to count images.

```
package countURL;

import java.io.BufferedInputStream;
import java.io.DataInputStream;
import java.io.IOException;
import java.net.MalformedURLException;
import java.net.URL;
import java.net.URLConnection;
import jericho.*;

public class CountURL {

  /**
   * This method returns a String corresponding to the HTML code of the page
   * which URL is given as a parameter.
   * Parameter of the method:
   * - urlString (String) : the URL of the HTML page for which we have to retrieve
   *                        the code.
   */
  public static String getHtmlCodeFromUrl(String urlString)
  {

    URL theURL = null;
    try {

      theURL = new URL(urlString);

    } catch (MalformedURLException e1) {

      e1.printStackTrace();

    }
    //Transformation of the String containing the URL as a URL object that can be reused.

    URLConnection conn = null;
    DataInputStream data = null;
    String line;
    StringBuffer buf = new StringBuffer();
    //Declaration and initialization of the variables which are necessary to open
    //the connection to the page.
    try {

      conn = theURL.openConnection();
      conn.connect();
      //We open a connection to the URL we have previously obtained and
      //we retrieve the HTML code.

      data = new DataInputStream(
                 new BufferedInputStream(conn.getInputStream())
      );
      while ((line = data.readLine()) != null) {
        buf.append(line + "\n");
      }
      //The HTML code is retrieved through a stream (DataInputStream) that has been opened
      //thanks to the connection.  Each "line" is inserted in a
      //StringBuffer.

      data.close();
      //We close the stream.

    } catch (IOException e) {

      return "IO Error";
      //The method returns an error message if an exception occurs
      //when using the input and output stream, when we open
      //the latter or when we open the
      //connection.

    }

    return buf.toString();
    //We convert the StringBuffer into a String and we send this object to the
    //calling method.

  }
```

```
/**
 * Main method of the class, called in the command line/the
 * shell.
 * The user has to enter the URL to the page of which he/she wants to retrieve
 * the HTML code and obtain the number of URLs
 * Method parameter:
 * - args (String table) : the parameters the user has given
 *                              when calling this
 *                              method
 */
public static void main (String args[])
{

  if (args.length == 0) {

    System.out.println(
        "Hi! Put an URL without \"http://\" as argument"
    );
    //If no parameter is given, the method reminds the user
    //how to use it.

  }
  else {

    String contents = getHtmlCodeFromUrl("http://"+ args[0]);
    //When calling getHtmlCodeFromUrl, the method retrieves the HTML code
    //of the page included in the site corresponding to the specified URL.

    if (contents.equals("IO Error")) {

      System.out.println();
      System.out.println("---> Download of page failed");
      System.out.println();
      //An"IO Error" means the URL is incorrect or that an error
      //occured during the retrieval process
      //of the HTML code of the page.

    }
    else {

      Source source = new Source(contents);
      //Using the functionalities of the "jericho" package, the content
      //of the page is converted into a Source object, that can easily be used
      //to retrieve HTML elements.
```

```
        int nbrUrls = source.findAllElements("url").size();
        //The findAllElements method returns a list of elements of
        //the HTML page which are URls.
        //We only focus on their number, that is why
        //only record the size of the list.

        System.out.println();
        System.out.println("---> " + nbrUrls + " URLs found for "
                           + args[0]);
        System.out.println();
        //L'information trouvee est affichee dans la ligne de commande/le
        //shell de l'utilisateur.

      }

    }

  }

}
```

B.2 Computation of the confidence interval for the Wilcoxon-Mann-Whitney test

We run this program when the Wilcoxon-Mann-Whitney test highlights a difference between males and females. The goal of the program is thus to compute the interval representing the extent to which both populations differ regarding a specific attribute. For example, if the Wilcoxon-Mann-Whitney test highlights the fact there is a difference regarding the number of photos, the result of the program can be (0,11) with a Hodges-Lehmann estimator of 1. This means the populations can have a difference between 0 and 11 photos. The mean difference is 1 photo.

```
package Interval;

import java.lang.*;
import java.util.*;

public class Interval {

  public static void main(String[] args){

    String Ma1, Ma2, Ma3, Ma4, Ma5, Ma6, Ma7, Ma8,
           Ma9, Ma10, Ma11, Ma12, Ma13, Ma14, Ma15,
           Fe1, Fe2, Fe3, Fe4, Fe5, Fe6, Fe7, Fe8,
           Fe9, Fe10, Fe11, Fe12, Fe13, Fe14, Fe15;
    //Declaration of 15 String for the males and 15 String for the
    //females in order to retrieve the values entered thanks to the keyboard
    //for each male and each female.

    int i = 0;
    int j = 0;
    int ourthreshold;
    //Declaration of the variable which value will be the critical value
    //below which we can reject the null hypothesis
    //for the Wilcoxon-Mann-Whitney test.

    Float ourconverter;
    //Declaration of float converter that will give us the possibility of converting the values of
    //the males and the females from String to Float.

    HashMap thevectorM = new HashMap();
    //Declaration and initialization of the vector in which the values for the males
    //will be recorded.

    HashMap thevectorF = new HashMap();
    //Declaration and initialization of the vector in which the values for the females
    //will be recorded.

    Vector ourmatrix = new Vector();
    //Declaration and initialization of the vector that will contain the result of
    //the difference between the values of the males' vector, thevectorM, and the values of
    //the females' vector, the vectorF (see body of the program).

    Vector thesortedmatrix= new Vector();
    //Declaration and initialization of the vector thesortedmatrix.
    //This vector will contain the values of ourmatrix sorted by ascending order.

    Float[] thetable = new Float[225];
    //Declaration and initialization of the table in which we will transfer
    //the values from the vector ourmatrix.

    float temp, lowerlimit, upperlimit, average;
    //Declaration and initialization of temp that will allow us to compute the difference
    //between the values of the males' vector, thevectorM and the values of the
    //females' vector, the vectorF.

    while (i < 15) {
      i = i + 1;
      converter = new Float(args[j]);
      thevectorM.put("Ma" + i, converter);
      System.out.println(thevectorM.get("Ma"+i)) ;
      j = j + 1;
    }
    //This first loop allows to fill the males' vector, thevectorM,
    //with the values entered through the keyboard.

    i = 0;
    while (i < 15) {
      i = i + 1;
      converter = new Float(args[j]);
      thevectorF.put("Fe"+i, converter);
      System.out.println(thevectorF.get("Fe"+i)) ;
      j = j + 1;
    }
```

```
//This second loop allows to fill the females' vector, thevectorF,
//with the values entered through the keyboard.

threshold = Integer.parseInt(args[j]);
//Here we get the critical value
//and we transform this one into an integer.

i = 0;
while (i < 15) {
i = i + 1;
j = 0;
  while (j < 15) {
    j = j + 1;
    temp = ((Float) thevectorM.get("Ma"+j)).floatValue()
           - ((Float)thevectorF.get("Fe"+i)).floatValue();
    System.out.println(temp);
    ourmatrix.addElement(new Float(temp));
  }
  //This loop allows to fill the vector ourmatrix with the result of
  //the difference between the value of male j and the value
  //of each female.
}
//This loop allows to carry out the computation of the difference (see internal loop)
//for each male.

i = 0;
while (i< 225){
  thetable[i] = (Float) ourmatrix.get(i);
  System.out.println("value of the table " + thetable[i]);
  i = i + 1;
}
//This loop allows to transfer the values of the vector ourmatrix
//in the float table thetable.
//So, for i belonging to [1,225]: thetable[i] = ourmatrix[i].

ArrayList theList = new ArrayList(Arrays.asList(thetable));
//Transformation of the float table thetable into an ArrayList to apply
//the Collections.sort() method.

Collections.sort(theList);
//Allows to sort theList by ascending order.

i = 0;
while (i < 225) {
  thesortedmatrix.addElement(thesortedmatrix.get(i)) ;
  System.out.println(theList.get(i));
  i = i + 1;
}
//This loop allows to transfer the elements of theList into
//the vector thesortedmatrix.
//So, for i belonging to [1,225], thesortedmatrix[i] = theList[i].

lowerlimit = ((Float) thesortedmatrix.elementAt(threshold)).floatValue();
```

```
    //Gives the lower value of the confidence interval representing
    //the range of the difference between the males and the females.

    upperlimit = ((Float) thesortedmatrix.elementAt(224 - threshold)).floatValue();
    //Gives the upper value of the confidence interval representing
    //the range of the difference between the males and the females.

    average = ((Float) thesortedmatrix.elementAt(112)).floatValue();
    //Gives the location of the difference for the considered attribute (Hodges-
    //Lehmann Estimator).
    //For example: the lower limit for the difference between the males and the females
    //for the attribute Photos is 0 photo whereas the upper limit is 11 photos.
    //The difference is on average 1 photo.

  }

}
```

Appendix C

Results of the numerical analysis

In this appendix, you will find the complete results of the numerical analysis.

C.1 The number of words per page

The problem here is to test if there is a location difference between the males and the females regarding the number of words per page. In table C.1, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 228 and for the females 237. U_m (the males) is 108 and U_n (the females) is 117. From Neave's table of critical values (see appendix D), we see that using a two-tail test at the 5 % level, the critical value is 64. So the conclusion is easily drawn. Since the lowest value (108) is situated above 64, we **cannot reject** the hypothesis that both distributions have the same location.

Value	25,7	62	65,5	78,5	89	105,7	133,4	142,9	151,2	154,3
Rank	1	2	3	4	5	6	7	8	9	10
Value	160,8	164,1	168,7	178,8	197	207	223,4	227	241,1	271
Rank	11	12	13	14	15	16	17	18	19	20
Value	276,9	280,6	283,6	299,33	303,2	329	481,4	524	568	614
Rank	21	22	23	24	25	26	27	28	29	30

Table C.1: WMW's table for the number of words per page

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 232,21 words per page whereas it is 234,95 words per page for the females. In table C.2, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the males is T = 4852. The mean of the squared ranks for all thirty observations is 315,17. Here S equals 770,26. Thus Z = (4852 - 15*315,17)/770,26 equals 1,87. Since Z is below 1,96, we can conclude that the variances are **equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that both samples come from identically distributed populations. In table C.3, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,1333 (final column). With a two-tail test at a nominal 5 % level, the critical value is 0,5333. So we **cannot reject** the null hypothesis saying that both samples come from

Deviation	7,95	8,89	11,55	25,21	36,05	37,95	44,69	45,65	48,85	53,41
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	63,51	67,12	70,85	70,99	71,41	80,65	83,75	89,31	94,05	98,81
Rank	11	12	13	14	15	16	17	18	19	20
Square	121	144	169	196	225	256	289	324	361	400
Deviation	129,25	143,21	153,71	166,71	172,95	209,25	246,45	291,79	333,05	381,79
Rank	21	22	23	24	25	26	27	28	29	30
Square	441	484	529	576	625	676	729	784	841	900

Table C.2: SR table for the number of words per page

Females	Males	S(f)	S(m)	S(f) - S(m)
25,7		0,0667	0	0,0667
62		0,1333	0	0,1333
	65,5	0,1333	0,0667	0,0667
	78,5	0,1333	0,1333	0
	89	0,1333	0,2	0,0667
105,7		0,2	0,2	0
	133,4	0,2	0,26667	0,0667
	142,9	0,2	0,3333	0,1333
151,2		0,2667	0,3333	0,0667
154,3		0,3333	0,3333	0
	160,8	0,3333	0,4	0,0667
164,1		0,4	0,4	0
	168,7	0,4	0,4667	0,0667
	178,8	0,4	0,53333	0,1333
197		0,46667	0,5333	0,0667
	207	0,4667	0,6	0,1333
223,4		0,5333	0,6	0,1333
227		0,6	0,6	0
	241,1	0,6	0,6667	0,0667
271		0,6667	0,6667	0
	276,9	0,6667	0,7333	0,0667
280,6		0,7333	0,7333	0
283,8		0,8	0,7333	0,0667
	299,33	0,8	0,8	0
	303,2	0,8	0,8667	0,0667
329		0,8667	0,8667	0
481,4	49123	0,9333	0,8667	0,0667
	524	0,9333	0,9333	0
568		1	0,9333	0,0667
	614	1	1	0

Table C.3: SMIR's table for the number of words per page

identically distributed populations.

C.2 The number of characters per word

The problem here is to test if there is a location difference between the males and the females regarding the length of the words. In table C.4, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 199 and for the females 266. U_m (the males) is 79 and U_n (the females) is 146. Since the lowest value (79) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Value	$\underline{4,5}$	$\underline{4,8}$	4,8	$\underline{4,9}$	$\underline{4,9}$	$\underline{4,9}$	$\underline{5,1}$	$\underline{5,2}$	5,2	5,3
Rank	1	2,5	2,5	4,33	4,33	4,33	7	8,5	8,5	10
Value	$\underline{5,4}$	5,4	5,5	5,5	5,5	$\underline{5,6}$	$\underline{5,6}$	$\underline{5,6}$	5,7	5,7
Rank	11,5	11,5	13,33	13,33	13,33	16,33	16,33	16,33	19,5	19,5
Value	$\underline{5,8}$	$\underline{5,9}$	5,9	6,3	6,3	6,5	6,6	6,6	$\underline{6,9}$	$\underline{8,4}$
Rank	21	22,5	22,5	24,5	24,5	26	27,5	27,5	29	30

Table C.4: WMW's table for the number of characters per word

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 5,57 characters per word whereas it is 5,79 characters per word for the females. In table C.5, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the males is T = 4997,67. The mean of the squared ranks for all thirty observations is 310,65. Here S equals 769,83. Thus Z = (4997,67- 15*310,65)/769,83 equals 0,44. Since Z is below 1,96, we can conclude that the variances are **equal**.

Deviation	$\underline{0,03}$	$\underline{0,03}$	$\underline{0,03}$	0,09	0,09	0,11	$\underline{0,17}$	$\underline{0,23}$	0,29	0,29
Rank	1,33	1,33	1,33	4,5	4,5	6	7	8	9,33	9,33
Square	1,78	1,78	1,78	20,25	20,25	36	49	64	87,11	87,11
Deviation	0,29	$\underline{0,33}$	$\underline{0,37}$	0,39	$\underline{0,47}$	0,49	0,51	0,51	0,59	$\underline{0,67}$
Rank	9,33	12	13	14	15	16	17,5	17,5	19	20,33
Square	87,11	144	169	196	225	256	306,25	306,25	361	413,44
Deviation	$\underline{0,67}$	$\underline{0,67}$	0,71	$\underline{0,77}$	0,81	0,81	0,99	$\underline{1,07}$	$\underline{1,33}$	$\underline{2,83}$
Rank	20,33	20,33	23	24	25,5	25,5	27	28	29	30
Square	413,44	413,44	529	576	650,25	650,25	729	784	841	900

Table C.5: SR table for the number of characters per word

Let's now use the Smirnov test to determine if it is reasonable to assume that both samples come from identically distributed populations. In table C.6, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,4 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.3 The proportion of white spaces

The problem here is to test if there is a location difference between the males and the females regarding the proportion of white spaces. In table C.7, you will find all sample values and the associated

Females	Males	S(f)	S(m)	S(f) - S(m)
	4,5	0	0,0667	0,0667
	4,8	0	0,1333	0,1333
4,8		0,0667	0,1333	0,0667
	4,9	0,0667	0,2	0,1333
	4,9	0,0667	0,2667	0,2
	4,9	0,0667	0,3333	0,2667
	5,1	0,0667	0,4	0,3333
	5,2	0,0667	0,4667	0,4
5,2		0,1333	0,4667	0,3333
5,3		0,2	0,4667	0,2667
	5,4	0,2	0,5333	0,3333
5,4		0,2667	0,5333	0,2667
5,5		0,3333	0,53333	0,2
5,5		0,4	0,5333	0,1333
5,5		0,4667	0,5333	0,0667
	5,6	0,4667	0,6	0,1333
	5,6	0,4667	0,6667	0,2
	5,6	0,4667	0,7333	0,2667
5,7		0,5333	0,7333	0,2
5,7		0,6	0,7333	0,1333
	5,8	0,6	0,8	0,2
	5,9	0,6	0,8667	0,2667
5,9		0,6667	0,8667	0,2
6,3		0,7333	0,8667	0,1333
6,3		0,8	0,8667	0,0667
6,5		0,8667	0,8667	0
6,6		0,9333	0,8667	0,0667
6,6		1	0,8667	0,1333
	6,9	1	0,9333	0,0667
	8,4	1	1	0

Table C.6: SMIR's table for the number of characters per word

ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 270 and for the females 195. U_m (the males) is 150 and U_n (the females) is 75. Since the lowest value (75) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Value	$\underline{0{,}1145}$	0,1228	0,1297	0,1384	0,1398	0,1406	0,1408	$\underline{0{,}1411}$	$\underline{0{,}1432}$	$\underline{0{,}1458}$
Rank	1	2	3	4	5	6	7	8	9	10
Value	$\underline{0{,}1473}$	0,1476	$\underline{0{,}1481}$	0,1501	0,1511	0,1538	$\underline{0{,}1549}$	0,1565	$\underline{0{,}169}$	0,1709
Rank	11	12	13	14	15	16	17	18	19	20,5
Value	0,1709	$\underline{0{,}171}$	$\underline{0{,}1713}$	$\underline{0{,}1763}$	0,1808	$\underline{0{,}185}$	0,1855	$\underline{0{,}1874}$	$\underline{0{,}204}$	$\underline{0{,}2072}$
Rank	20,5	22	23	24	25	26	27	28	29	30

Table C.7: WMW's table for the proportion of white spaces

Now we are going to test the equality of variances by the squared rank test for variance. The

estimation of the mean for the males is 0,1644 whereas it is 0,1520 for the females. In table C.8, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the males is T = 5443. The mean of the squared ranks for all thirty observations is 312,74. Here S equals 769,86. Thus Z = (5443- 15*312,74)/769,86 equals 0,98. Since Z is below 1,96, we can conclude that the variances are **equal**.

Deviation	0,0009	0,0018	0,0019	0,0044	0,0046	0,0066	0,0069	0,0095	0,0112	0,0114
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	0,0119	0,0122	0,0136	0,0163	0,0171	0,0186	0,0189	0,0189	0,0189	0,0206
Rank	11	12	13	14	15	16	17,33	17,33	17,33	20
Square	121	144	169	196	225	300,44	300,44	300,44	400	
Deviation	0,0212	0,0223	0,0230	0,0233	0,0288	0,0292	0,0335	0,0396	0,0428	0,0499
Rank	21	22	23	24	25	26	27	28	29	30
Square	441	484	529	576	625	676	729	784	841	900

Table C.8: SR table for the proportion of white spaces

Let's now use the Smirnov test to determine if it is reasonable to assume that both samples come from identically distributed populations. In table C.9, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,3333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.4 The number of paragraphs per page

The problem here is to test if there is a location difference between the males and the females regarding the number of paragraphs per page. In table C.10, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 214 and for the females 251. U_m (the males) is 94 and U_n (the females) is 131. Since the lowest value (94) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 16,4 paragraphs per page whereas it is 21,3 paragraphs per page for the females. In table C.11, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the males is T = 6362. The mean of the squared ranks for all thirty observations is 315,17. Here S equals 770,26. Thus Z = (6362- 15*315,17)/770,26 equals 2,12. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine it is reasonable to assume that both samples come from identically distributed populations. In table C.12, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,2667 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

Females	Males	S(f)	S(m)	S(f) - S(m)
	0,1145	0	0,0667	0,0667
0,1228		0,0667	0,0667	0
0,1297		0,1333	0,0667	0,0667
0,1384		0,2	0,0667	0,1333
0,1398		0,2667	0,0667	0,2
0,1406		0,3333	0,0667	0,2667
0,1408		0,4	0,0667	0,3333
	0,1411	0,4	0,1333	0,2667
	0,1432	0,4	0,2	0,2
	0,1458	0,4	0,2667	0,1333
	0,1473	0,4	0,3333	0,0667
0,1476		0,4667	0,3333	0,1333
	0,1481	0,4667	0,4	0,0667
0,1501		0,5333	0,4	0,1333
0,1511		0,6	0,4	0,2
0,1538		0,6667	0,4	0,2667
	0,1549	0,6667	0,4667	0,2
0,1565		0,7333	0,4667	0,2667
	0,169	0,7333	0,5333	0,2
0,1709		0,8	0,5333	0,2667
0,1709		0,8667	0,5333	0,3333
	0,171	0,8667	0,6	0,2667
	0,1713	0,8667	0,6667	0,2
	0,1763	0,8667	0,7333	0,1333
0,1808		0,9333	0,7333	0,2
	0,185	0,9333	0,8	0,1333
0,1855		1	0,8	0,2
	0,1874	1	0,8667	0,1333
	0,204	1	0,9333	0,0667
	0,2072	1	1	0

Table C.9: SMIR's table for the proportion of white spaces

Value	3,2	$\underline{6,1}$	$\underline{6,8}$	6,9	8,0	8,0	8,8	$\underline{11,0}$	$\underline{11,4}$	$\underline{12,0}$
Rank	1	$\underline{2}$	$\underline{3}$	4	5,5	5,5	7	$\underline{8}$	$\underline{9}$	$\underline{10}$
Value	$\underline{13,0}$	$\underline{13,0}$	13,8	$\underline{14,4}$	$\underline{14,5}$	$\underline{16,7}$	17,1	18,8	$\underline{20,7}$	$\underline{20,9}$
Rank	$\underline{11,5}$	$\underline{11,5}$	13	$\underline{14}$	$\underline{15}$	$\underline{16}$	17	18	$\underline{19}$	$\underline{20,5}$
Value	21,0	23,0	$\underline{24,0}$	$\underline{27,0}$	30,1	31,1	34,3	$\underline{34,8}$	44,2	51,0
Rank	20,5	22	$\underline{23}$	$\underline{24}$	25	26	27	$\underline{28}$	29	30

Table C.10: WMW's table for the number of paragraphs per page

C.5 The number of fonts

The problem here is to test if there is a location difference between the males and the females regarding the number of fonts. In table C.13, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 235 and for the females 115 . U_m (the males) is 110 and U_n (the females) is 131. Since the lowest value (110) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the

Deviation	0,3	0,32	1,67	1,9	2,02	2,54	3,38	3,4	4,23	4,3
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	4,45	4,5	4,99	5,36	7,5	7,56	8,76	9,56	9,76	10,34
Rank	11	12	13	14	15	16	17	18	19	20
Square	121	144	169	196	225	289	324	361	400	
Deviation	10,64	12,47	12,95	13,3	13,33	14,37	18,06	18,41	22,85	29,7
Rank	21	22	23	24	25	26	27	28	29	30
Square	441	484	529	576	625	676	729	784	841	900

Table C.11: SR table for the number of paragraphs per page

Females	Males	S(f)	S(m)	S(f) - S(m)
3,2		0,0667	0	0,0667
	6,1	0,0667	0,0667	0
	6,8	0,0667	0,1333	0,0667
6,9		0,1333	0,1333	0
8,0		0,2	0,1333	0,0667
8,0		0,2667	0,1333	0,1333
8,8		0,3333	0,1333	0,2
	11,0	0,3333	0,2	0,1333
	11,4	0,3333	0,2667	0,0667
	12,0	0,3333	0,3333	0
	13,0	0,3333	0,4	0,0667
	13,0	0,3333	0,4667	0,1333
13,8		0,4	0,4667	0,0667
	14,4	0,4	0,5333	0,1333
	14,5	0,4	0,6	0,2
	16,7	0,4	0,6667	0,2667
17,1	0,4667	0,6667	0,2	
18,8		0,5333	0,6667	0,1333
	20,7	0,5333	0,7333	0,2
	20,9	0,5333	0,8	0,2667
21,0		0,6	0,8	0,2
23,0		0,6667	0,8	0,1333
	24,0	0,6667	0,8667	0,2
	27,0	0,6667	0,9333	0,2667
30,1		0,7333	0,9333	0,2
31,1		0,8	0,9333	0,1333
34,3		0,8667	0,9333	0,0667
	34,8	0,8667	1	0,1333
44,2		0,9333	1	0,0667
51,0		1	1	0

Table C.12: SMIR's table for the number of paragraphs per page

same location.

Now we are going to test the equality of variances by the squared rank test for variance. The

Value	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	1	1	1	1
Rank	1,077	1,077	1,077	1,077	1,077	1,077	1,077	1,077	1,077	1,077
Value	1	1	1	$\underline{2}$	$\underline{2}$	$\underline{2}$	$\underline{2}$	$\underline{2}$	$\underline{2}$	$\underline{2}$
Rank	1,077	1,077	1,077	14,071	14,071	14,071	14,071	14,071	14,071	14,071
Value	$\underline{2}$	2	2	2	2	2	2	$\underline{3}$	3	3
Rank	14,071	14,071	14,071	14,071	14,071	14,071	14,071	28,33	28,33	28,33

Table C.13: WMW's table for the number of fonts

estimation of the mean for the males is 1,67 fonts as well as for the females. In table C.14, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the males is

Deviation	$\underline{0,33}$	$\underline{0,33}$	$\underline{0,33}$	$\underline{0,33}$	$\underline{0,33}$	$\underline{0,33}$	$\underline{0,33}$	$\underline{0,33}$	0,33	0,33
Rank	1,071	1,071	1,071	1,071	1,071	1,071	1,071	1,071	1,071	1,071
Square	1,148	1,148	1,148	1,148	1,148	1,148	1,148	1,148	1,148	1,148
Deviation	0,33	0,33	0,33	0,33	$\underline{0,67}$	$\underline{0,67}$	$\underline{0,67}$	$\underline{0,67}$	$\underline{0,67}$	$\underline{0,67}$
Rank	1,071	1,071	1,071	1,071	15,077	15,077	15,077	15,077	15,077	15,077
Square	1,148	1,148	1,148	1,148	227,31	227,31	227,31	227,31	227,31	
Deviation	0,67	0,67	0,67	0,67	0,67	0,67	0,67	$\underline{1,33}$	1,33	1,33
Rank	15,077	15,077	15,077	15,077	15,077	15,077	15,077	28,33	28,33	28,33
Square	227,31	227,31	227,31	227,31	227,31	227,31	227,31	802,59	802,59	802,59

Table C.14: SR table for the number of fonts

T = 3203,26. The mean of the squared ranks for all thirty observations is 179,3. Here S equals 651,21. Thus Z = (3203,26- 15*179,3)/651,21 equals 0,79. Since Z is below 1,96, we can conclude that the variances are **equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that both samples come from identically distributed populations. In table C.15, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,4667 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.6 The number of colours for text and hypertext

The problem here is to test if there is a location difference between the males and the females regarding the number of colours for text and hypertext. In table C.16, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 278,5 and for the females 186,5. U_m (the males) is 158,5 and U_n (the females) is 66,5. Since the lowest value (66,5) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 4,93 colours whereas it is 3,87 colours for the females. In table C.17, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the males is T = 4977,39. The mean of the squared ranks for all thirty observations is 279,07. Here S equals 794,21. Thus Z = (4977,39- 15*279,07)/794,21 equals 1. Since Z is below 1,96, we can conclude that the variances are **equal**.

Females	Males	S(f)	S(m)	S(f) - S(m)
	1	0	0,0667	0,0667
	1	0	0,1333	0,1333
	1	0	0,2	0,2
	1	0	0,2667	0,2667
	1	0	0,3333	0,3333
	1	0	0,4	0,4
1		0,0667	0,4	0,3333
1		0,1333	0,4	0,2667
1		0,2	0,4	0,2
1		0,2667	0,4	0,1333
1		0,3333	0,4	0,0667
1		0,4	0,4	0
1		0,4667	0,4	0,0667
	2	0,4667	0,4667	0
	2	0,4667	0,5333	0,0667
	2	0,4667	0,6	0,1333
	2	0,4667	0,6667	0,2
	2	0,4667	0,7333	0,2667
	2	0,4667	0,8	0,3333
	2	0,4667	0,8667	0,4
	2	0,4667	0,9333	0,4667
2		0,5333	0,9333	0,4
2		0,6	0,9333	0,3333
2		0,6667	0,9333	0,2667
2		0,7333	0,9333	0,2
2		0,8	0,9333	0,1333
2		0,8667	0,9333	0,0667
	3	0,8667	1	0,1333
3		0,9333	1	0,0667
3		1	1	0

Table C.15: SMIR's table for the number of fonts

Value	2	2	2	3	3	3	3	3	3	3
Rank	1,33	1,33	1,33	4,09	4,09 4,09	4,09	4,09	4,09	4,09	
Value	3	3	3	3	4	4	4	4	4	4
Rank	4,09	4,09	4,09	4,09	15,11	15,11	15,11	15,11	15,11	15,11
Value	4	4	4	5	6	7	8	10	10	11
Rank	15,11	15,11	15,11	24	25	26	27	28,5	28,5	30

Table C.16: WMW's table for the number of colours for text and hypertext

Let's now use the Smirnov test to determine if it is reasonable to assume that both samples come from identically distributed populations. In table C.18, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,4 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

Deviation	0,07	0,13	0,13	0,87	0,87	0,87	0,87	0,87	0,87	0,87
Rank	1	2,5	2,5	4,14	4,14	4,14	4,14	4,14	4,14	4,14
Square	1	6,25	6,25	17,16	17,16	17,16	17,16	17,16	17,16	17,16
Deviation	0,93	0,93	0,93	0,93	0,93	0,93	0,93	1,87	1,87	1,87
Rank	11,14	11,14	11,14	11,14	11,14	11,14	11,14	18,33	18,33	18,33
Square	124,16	124,16	124,16	124,16	124,16	124,16	124,16	335,99	335,99	335,99
Deviation	1,93	1,93	1,93	1,93	2,13	3,07	3,13	5,07	6,07	6,13
Rank	21,25	21,25	21,25	21,25	25	26	27	28	29	30
Square	451,56	451,56	451,56	451,56	625	676	729	784	841	900

Table C.17: SR table for the number of colours for text and hypertext

Females	Males	S(f)	S(m)	S(f) - S(m)
2		0,0667	0	0,0667
2		0,1333	0	0,1333
2		0,2	0	0,2
	3	0,2	0,0667	0,1333
	3	0,2	0,1333	0,0667
	3	0,2	0,2	0
	3	0,2	0,2667	0,0667
3		0,2667	0,2667	0
3		0,3333	0,2667	0,0667
3		0,4	0,2667	0,1333
3		0,4667	0,2667	0,2
3		0,5333	0,2667	0,2667
3		0,6	0,2667	0,3333
3		0,6667	0,2667	0,4
	4	0,6667	0,3333	0,3333
	4	0,6667	0,4	0,2667
	4	0,6667	0,4667	0,2
	4	0,6667	0,5333	0,1333
	4	0,6667	0,6	0,0667
	4	0,6667	0,6667	0
	4	0,6667	0,7333	0,0667
4		0,7333	0,7333	0
4		0,8	0,7333	0,0667
	5	0,8	0,8	0
6		0,8667	0,8	0,0667
7		0,9333	0,8	0,1333
	8	0,9333	0,8667	0,0667
	10	0,9333	0,9333	0
10		1	0,9333	0,0667
	11	1	1	0

Table C.18: SMIR's table for the number of colours for text and hypertext

C.7 The number of words for the main page

The problem here is to test if there is a location difference between the males and the females regarding the number of words for the main page. In table C.19, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 229,5 and for the females 235,5. U_m (the males) is 109,5 and U_n (the females) is 115,5. Since the lowest value (109,5) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Value	2	$\underline{13}$	17	$\underline{43}$	57	74	$\underline{76}$	$\underline{113}$	$\underline{126}$	178
Rank	1	2	3	4	5	6	7	8	9	10
Value	185	$\underline{191}$	$\underline{196}$	$\underline{214}$	271	$\underline{278}$	322	355	$\underline{394}$	394
Rank	11	12	13	14	15	16	17	18	19,5	19,5
Value	405	$\underline{414}$	$\underline{420}$	$\underline{459}$	499	705	$\underline{860}$	1715	$\underline{1860}$	2272
Rank	21	22	23	24	25	26	27	28	29	30

Table C.19: WMW's table for the number of words for the main page

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 37,13 words whereas it is 496,73 words for the females. In table C.20, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T = 5404. The mean of the squared ranks for all thirty observations is 315,17. Here S equals 770,26. Thus Z = (5404 - 15*315,17)/770,26 equals 0,88. Since Z is below 1,96, we can conclude that the variances are **equal**.

Deviation	2,27	$\underline{5,87}$	$\underline{24,13}$	$\underline{38,87}$	$\underline{75,87}$	$\underline{88,87}$	91,73	102,73	141,73	$\underline{153,87}$
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	$\underline{158,87}$	174,73	$\underline{176,87}$	208,87	225,73	$\underline{240,87}$	311,73	318,73	$\underline{356,87}$	$\underline{376,87}$
Rank	11	12	13	14	15	16	17	18	19	20
Square	121	144	169	196	225	256	289	324	361	400
Deviation	$\underline{382,87}$	$\underline{421,87}$	422,73	439,73	479,73	494,73	$\underline{822,87}$	1218,27	1775,27	$\underline{1822,87}$
Rank	21	22	23	24	25	26	27	28	29	30
Square	441	484	529	576	625	676	729	784	841	900

Table C.20: SR table for the number of words for the main page

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.21, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,1333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.8 The number of colours for the background

The problem here is to test if there is a location difference between the males and the females regarding the number of colours for the background. In table C.22, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 260 and for the females 205. U_m (the males) is 140 and U_n (the females) is 85. Since the lowest value (85) is situated above the critical value, we **cannot reject** the hypothesis that both

Females	Males	S(f)	S(m)	S(f) - S(m)
2		0,0667	0	0,0667
	13	0,0667	0,0667	0
17		0,1333	0,0667	0,0667
	43	0,1333	0,1333	0
57		0,2	0,1333	0,0667
74		0,2667	0,1333	0,1333
	76	0,2667	0,2	0,0667
	113	0,2667	0,2667	0
	126	0,2667	0,3333	0,0667
178		0,3333	0,3333	0
185		0,4	0,3333	0,0667
	191	0,4	0,4	0
	196	0,4	0,4667	0,0667
	214	0,4	0,5333	0,1333
271		0,4667	0,5333	0,0667
	278	0,4667	0,6	0,1333
322		0,5333	0,6	0,0667
355		0,6	0,6	0
	394	0,6	0,6667	0,0667
394		0,6667	0,6667	0
405		0,7333	0,6667	0,0667
	414	0,7333	0,7333	0
	420	0,7333	0,8	0,0667
	459	0,7333	0,8667	0,1333
499		0,8	0,8667	0,0667
705		0,8667	0,8667	0
	860	0,8667	0,9333	0,0667
1715		0,9333	0,9333	0
	1860	0,9333	1	0,0667
2272		1	1	0

Table C.21: SMIR's table for the number of words for the main page

Value	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	1	1	1
Rank	1,06	1,06	1,06	1,06	1,06	1,06	1,06	1,06	1,06	1,06
Value	1	1	1	1	1	1	1	$\underline{2}$	$\underline{2}$	2
Rank	1,06	1,06	1,06	1,06	1,06	1,06	1,06	18,33	18,33	18,33
Value	$\underline{3}$	$\underline{3}$	3	3	3	$\underline{4}$	5	$\underline{6}$	$\underline{7}$	$\underline{13}$
Rank	21,2	21,2	21,2	21,2	21,2	26	27	28	29	30

Table C.22: WMW's table for the number of colours for the background

distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 3,13 colours whereas it is 1,73 colours for the females. In table C.23, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T = 5991,14. The mean of the squared ranks for all thirty observations is 261,77. Here S equals 752,12. Thus Z = (5991,14 - 15*261,77)/752,12 giving 2,74 as a result. Since Z is above 1,96,

Deviation	$\underline{0{,}13}$	$\underline{0{,}13}$	0,27	0,71	0,71	0,71	0,71	0,71	0,71	0,71
Rank	$\underline{1{,}5}$	$\underline{1{,}5}$	3	4,1	4,1	4,1	4,1	4,1	4,1	4,1
Square	2,25	2,25	9	16,81	16,81	16,81	16,81	16,81	16,81	16,81
Deviation	0,71	0,71	0,71	$\underline{0{,}87}$	$\underline{1{,}13}$	$\underline{1{,}13}$	1,27	1,27	1,27	$\underline{2{,}13}$
Rank	4,1	4,1	4,1	$\underline{14}$	$\underline{15{,}5}$	$\underline{15{,}5}$	17,33	17,33	17,33	$\underline{20{,}14}$
Square	16,81	16,81	16,81	196	240,25	240,25	300,33	300,33	300,33	405,73
Deviation	$\underline{2{,}13}$	$\underline{2{,}13}$	$\underline{2{,}13}$	$\underline{2{,}13}$	$\underline{2{,}13}$	$\underline{2{,}13}$	$\underline{2{,}87}$	3,27	$\underline{3{,}87}$	$\underline{9{,}87}$
Rank	$\underline{20{,}14}$	$\underline{20{,}14}$	$\underline{20{,}14}$	$\underline{20{,}14}$	$\underline{20{,}14}$	$\underline{20{,}14}$	$\underline{27}$	28	$\underline{29}$	$\underline{30}$
Square	405,73	405,73	405,73	405,73	405,73	405,73	729	784	841	900

Table C.23: SR table for the number of colours for the background

we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.24, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,2667 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.9 The number of words for the self-description

The problem here is to test if there is a location difference between the males and the females regarding the number of words for the self-description. In table C.25, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 221 and for the females 244. U_m (the males) is 101 and U_n (the females) is 124. Since the lowest value (101) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 147 words whereas it is 200 words for the females. In table C.26, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T = 4859,2. The mean of the squared ranks for all thirty observations is 293,90. Here S equals 718,16. Thus Z = (4859,2 - 15*293,90)/718,16 equals 0,63. Since Z is below 1,96, we can conclude that the variances are **equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.27, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,3333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.10 The ratio of personal pages

The problem here is to test if there is a location difference between the males and the females regarding the ratio of personal pages. In table C.28, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 253,5 and for the females 211,5. U_m (the males) is 133,5 and U_n (the females) is 91,5. Since the

Females	Males	S(f)	S(m)	S(f) - S(m)
	1	0	0,0667	0,0667
	1	0	0,1333	0,1333
	1	0,0667	0,1333	0,0667
	1	0,1333	0,1333	0
	1	0,2	0,1333	0,0667
	1	0,2667	0,1333	0,1333
	1	0,3333	0,1333	0,2
1		0,3333	0,2	0,1333
1		0,4	0,2	0,2
1		0,4667	0,2	0,2667
1		0,4667	0,2667	0,2
1		0,4667	0,3333	0,1333
1		0,5333	0,3333	0,2
1		0,5333	0,4	0,1333
1		0,5333	0,4667	0,0667
1		0,6	0,4667	0,1333
1		0,6	0,5333	0,0667
	2	0,6667	0,5333	0,1333
	2	0,7333	0,5333	0,2
2		0,7333	0,6	0,1333
	3	0,8	0,6	0,2
	3	0,8	0,6667	0,1333
3		0,8	0,7333	0,0667
3		0,8	0,8	0
3		0,8	0,8667	0,0667
	4	0,8	0,9333	0,1333
5		0,8	1	0,2
	6	0,8667	0,1	0,1333
	7	0,9333	1	0,0667
	13	1	1	0

Table C.24: SMIR's table for the number of colours for the background

Value	0	0	0	0	0	0	0	0	0	9
Rank	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11	10
Value	31	57	76	90	97	101	109	120	136	137
Rank	11	12	13	14	15	16	17	18	19,5	19,5
Value	147	148	153	161	204	242	262	424	869	1632
Rank	21	22	23	24	25	26	27	28	29	30

Table C.25: WMW's table for the number of words for the self-description

lowest value (91,5) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 0,1602 (16,02%) whereas it is 0,0844 (8,44%) for the females. In table C.29, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T=6428,125. The mean of the squared ranks for all thirty observations is

Deviation	0	38	39	42	46	47	50	52	57	57
Rank	1	2	3	4	5	6	7	8	9,5	9,5
Square	1	4	9	16	25	36	49	64	90,25	90,25
Deviation	62	63	64	71	80	90	116	147	147	147
Rank	11	12	13	14	15	16	17	18,25	18,25	18,25
Square	121	144	169	196	225	256	289	333,06	333,06	333,06
Deviation	147	191	200	200	200	200	200	277	722	1432
Rank	18,25	22	23,2	23,2	23,2	23,2	23,2	28	29	30
Square	333,06	484	538,24	538,24	538,24	538,24	538,24	784	841	900

Table C.26: SR table for the number of words for the self-description

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
	0	0	0,1333	0,1333
	0	0	0,2	0,2
	0	0	0,2667	0,2667
0		0,0667	0,2667	0,2
0		0,1333	0,2667	0,1333
0		0,2	0,2667	0,0667
0		0,2667	0,2667	0
0		0,3333	0,2667	0,0667
9		0,4	0,2667	0,1333
	31	0,4	0,3333	0,0667
	57	0,4	0,4	0
	76	0,4	0,4667	0,0667
	90	0,4	0,5333	0,1333
	97	0,4	0,6	0,2
	101	0,4	0,6667	0,2667
	109	0,4	0,7333	0,3333
120		0,4667	0,7333	0,2667
136		0,5333	0,7333	0,2
137		0,6	0,7333	0,1333
	147	0,6	0,8	0,2
148		0,6667	0,8	0,1333
153		0,7333	0,8	0,0667
161		0,8	0,8	0
	204	0,8	0,8667	0,0667
242		0,8667	0,8667	0
262		0,9333	0,8667	0,0667
	424	0,9333	0,9333	0
	869	0,9333	1	0,0667
1632		1	1	0

Table C.27: SMIR's table for the number of words for the self-description

256,87. Here S equals 762,01. Thus Z = (6428,125 - 15*256,87)/762,01 giving 3,38 as a result. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples

Value	0	0	0	0	0	0	0	0	0	0
Rank	1,01	1,01	1,01	1,01	1,01	1,01	1,01	1,01	1,01	1,01
Value	0	0	0	0	0	0	0	0	0,006	0,03
Rank	1,01	1,01	1,01	1,01	1,01	1,01	1,01	1,01	19	20,5
Value	0,03	0,06	0,07	0,17	0,25	0,33	0,4	0,65	0,69	0,99
Rank	20,5	22	23	24	25	26	27	28	29	30

Table C.28: WMW's table for the ratio of personal pages

Deviation	0,0144	0,0544	0,0784	0,0844	0,0844	0,0844	0,0844	0,0844	0,0844	0,0844
Rank	1	2	3	4,1	4,1	4,1	4,1	4,1	4,1	4,1
Square	1	4	9	16,81	16,81	16,81	16,81	16,81	16,81	16,81
Deviation	0,0844	0,0844	0,0844	0,0856	0,0893	0,1007	0,1307	0,1607	0,1607	0,1607
Rank	4,1	4,1	4,1	15 16	17	18,125	18,125	18,125		
Square	16,81	16,81	16,81	225	256	289	328,52	328,52	328,52	
Deviation	0,1607	0,1607	0,1607	0,1607	0,1607	0,1693	0,2393	0,4893	0,5293	0,9056
Rank	18,125	18,125	18,125	18,125	18,125	26	27	28	29	30
Square	328,52	328,52	328,52	328,52	328,52	676	729	784	841	900

Table C.29: SR table for the ratio of personal pages

come from identically distributed populations. In table C.30, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,5333 (final column). Since it equals the critical value, we **cannot really reject** the null hypothesis saying that both samples come from identically distributed populations.

C.11 The number of links

The problem here is to test if there is a location difference between the males and the females regarding the number of links. In table C.31, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 258 and for the females 207. U_m (the males) is 138 and U_n (the females) is 87. Since the lowest value (87) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 370,47 links whereas it is 932,27 links for the females. In table C.32, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T = 7446,5. The mean of the squared ranks for all thirty observations is 315,15. Here S equals 770,14. Thus Z = (7446,5 - 15*315,15)/770,14 equals 3,53. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.33, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,3333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
	0	0	0,1333	0,1333
	0	0	0,2	0,2
	0	0	0,2667	0,2667
	0	0	0,3333	0,3333
	0	0	0,4	0,4
	0	0	0,4667	0,4667
	0	0	0,5333	0,5333
0		0,0667	0,5333	0,4667
0		0,1333	0,5333	0,4
0		0,2	0,5333	0,3333
0		0,2667	0,5333	0,2667
0		0,3333	0,5333	0,2
0		0,4	0,5333	0,1333
0		0,4667	0,5333	0,0667
0		0,5333	0,5333	0
0		0,6	0,5333	0,0667
0		0,6667	0,5333	0,1333
0,006		0,7333	0,5333	0,2
	0,03	0,7333	0,6	0,1333
0,03		0,8	0,6	0,2
	0,06	0,8	0,6667	0,1333
0,07		0,8667	0,6667	0,2
0,17		0,9333	0,6667	0,2667
	0,25	0,9333	0,7333	0,2
	0,33	0,9333	0,8	0,1333
	0,4	0,8667	0,9333	0,0667
	0,65	0,9333	0,9333	0
	0,69	0,9333	1	0,0667
0,99		1	1	0

Table C.30: SMIR's table for the ratio of personal pages

Value	2	12	14	17	20	21	39	41	44	63
Rank	1	2	3	4	5	6	7	8	9	10
Value	70	88	118	126	133	133	151	158	176	237
Rank	11	12	13	14	15,5	15,5	17	18	19	20
Value	281	293	7326	486	609	785	1057	2114	5764	6163
Rank	21	22	23	24	25	26	27	28	29	30

Table C.31: WMW's table for the number of links

C.12 The number of links to non-personal pages

The problem here is to test if there is a location difference between the males and the females regarding the number of links to non-personal pages. In table C.34, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 237,5 and for the females 227,5. U_m (the males) is 117,5 and U_n (the females) is 107,5. Since the lowest value (107,5) is situated above the critical value, we **cannot reject** the hypothesis

Deviation	44,47	89,47	115,53	124,73	133,47	212,47	219,47	238,53	244,47	252,47
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	282,47	326,47	350,47	356,47	414,53	639,27	756,27	799,27	799,27	862,27
Rank	11	12	13	14	15	16	17	18,5	18,5	20
Square	121	144	169	196	225	256	289	342,25	342,25	400
Deviation	869,27	891,27	893,27	911,27	915,27	920,27	930,27	1743,53	4831,73	5230,73
Rank	21	22	23	24	25	26	27	28	29	30
Square	441	484	529	576	625	676	729	784	841	900

Table C.32: SR table for the number of links

Females	Males	S(f)	S(m)	S(f) - S(m)
2		0,0667	0	0,0667
12		0,1333	0	0,1333
	14	0,1333	0,0667	0,0667
17		0,2	0,0667	0,1333
	20	0,2	0,1333	0,0667
21		0,2667	0,1333	0,1333
39		0,3333	0,1333	0,2
41		0,4	0,1333	0,2667
	44	0,4	0,2	0,2
63		0,46667	0,2	0,2667
70		0,5333	0,2	0,3333
	88	0,5333	0,2667	0,2667
	118	0,5333	0,3333	0,2
	126	0,5333	0,4	0,1333
133		0,6	0,4	0,2
133		0,6667	0,4	0,2667
	151	0,6667	0,4667	0,2
	158	0,6667	0,5333	0,1333
176		0,7333	0,5333	0,2
	237	0,7333	0,6	0,1333
	281	0,7333	0,6667	0,0667
293		0,8	0,6667	0,1333
	326	0,8	0,7333	0,0667
	486	0,8	0,8	0
	609	0,8	0,8667	0,0667
	785	0,8	0,9333	0,1333
1057		0,8667	0,9333	0,0667
	2114	0,8667	1	0,1333
5764		0,9333	1	0,0667
5764		0,9333	1	0,0667
6163		1	1	0

Table C.33: SMIR's table for the number of links

that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The

Value	0	1	4	5	6	6	8	11	14	14
Rank	1	2	3	4	5,5	5,5	7	8	9,5	9,5
Value	17	17	18	28	31	41	42	48	58	68
Rank	11,5	11,5	13	14	15,5	15,5	17	18	19	20
Value	73	98	101	120	126	164	379	429	802	3531
Rank	21	22	23	24	25	26	27	28	29	30

Table C.34: WMW's table for the number of links to non-personal pages

estimation of the mean for the males is 93,73 links whereas it is 323,6 links for the females. In table C.35, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T = 7490,5. The mean of the squared ranks for all thirty observations is 315,13. Here S equals 770,07. Thus Z = (7490,5 - 15*315,13)/770,07 equals 3,59. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Deviation	20,73	25,73	26,27	45,73	51,73	62,73	70,27	75,73	79,73	85,73
Rank	1	2	3	4	5	6	7	8	9	10
Square	1	4	9	16	25	36	49	64	81	100
Deviation	87,73	87,73	93,73	197,6	222,6	225,6	265,6	282,6	285,27	295,6
Rank	11,5	11,5	13	14	15	16	17	18,5	18,5	20
Square	132,25	132,25	169	196	225	256	289	342,25	342,25	400
Deviation	306,6	306,6	309,6	312,6	318,6	319,6	322,6	335,27	478,4	3207,4
Rank	21,5	21,5	23	24	25	26	27	28	29	30
Square	462,25	462,25	529	576	625	676	729	784	841	900

Table C.35: SR table for the number of links to non-personal pages

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.36, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,1333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.13 The number of personal links

The problem here is to test if there is a location difference between the males and the females regarding the number of links to ohter people's pages. In table C.37, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 226 and for the females 239. U_m (the males) is 106 and U_n (the females) is 119. Since the lowest value (106) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 16,2 links whereas it is 44,27 links for the females. In table C.38, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T=6947,83. The mean of the squared ranks for all thirty observations is 302,90. Here S equals 770,84. Thus Z = (6947,83 - 15*302,90)/770,84 equals 3,12. Since Z is above 1,96, we can

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
1		0,0667	0,0667	0
4		0,1333	0,0667	0,0667
5		0,2	0,0667	0,1333
	6	0,2	0,1333	0,0667
	6	0,2	0,2	0
	8	0,2	0,2667	0,0667
11		0,2667	0,2667	0
	14	0,2667	0,3333	0,0667
14		0,3333	0,3333	0
17		0,4	0,3333	0,0667
17		0,4667	0,3333	0,1333
	18	0,4667	0,4	0,0667
28		0,5333	0,4	0,1333
	31	0,5333	0,4667	0,0667
41		0,6	0,4667	0,1333
	42	0,6	0,5333	0,0667
	48	0,6	0,6	0
58		0,6667	0,6	0,0667
	68	0,6667	0,6667	0
	73	0,6667	0,7333	0,0667
98		0,7333	0,7333	0
101		0,8	0,7333	0,0667
	120	0,8	0,8	0
126		0,8	0,8667	0,0667
	164	0,8667	0,8667	0
	379	0,8667	0,9333	0,0667
	429	0,8667	1	0,1333
802		0,9333	1	0,0667
3531		1	1	0,0667

Table C.36: SMIR's table for the number of links to non-personal pages

Value	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	0	0	0	$\underline{1}$	1
Rank	1,125	1,125	1,125	1,125	1,125	1,125	1,125	1,125	2,33	2,33
Value	1	$\underline{3}$	3	3	$\underline{4}$	$\underline{4}$	4	4	4	$\underline{5}$
Rank	2,33	12,33	12,33	12,33	15,2	15,2	15,2	15,2	15,2	20,5
Value	5	$\underline{7}$	11	$\underline{21}$	$\underline{21}$	23	$\underline{83}$	$\underline{94}$	166	439
Rank	20,5	22	23	24,5	24,5	26	27	28	29	30

Table C.37: WMW's table for the number of personal links

conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.39, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,3333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

Deviation	4,8	4,8	9,2	11,2	12,2	12,2	13,2	15,2	16,2	16,2
Rank	1,5	1,5	3	4	5,5	5,5	7	8	9,2	9,2
Square	2,25	2,25	9	16	30,25	30,25	49	64	84,64	84,64
Deviation	16,2	16,2								
underline16,2	21,27	33,27	39,27	40,27	40,27	40,27	41,27			
Rank	9,2	9,2	9,2	14	15	16	17,33	17,33	17,33	20,5
Square	84,64	84,64	84,64	196	225	256	300,33	300,33	300,33	420,25
Deviation	41,27	43,27	43,27	44,27	44,27	44,27	66,8	77,8	121,73	394,73
Rank	20,5	22,5	22,5	24,33	24,33	24,33	27	28	29	30
Square	420,25	506,25	506,25	591,95	591,95	591,95	729	784	841	900

Table C.38: SR table for the number of personal links

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
	0	0	0,1333	0,1333
	0	0	0,2	0,2
	0	0	0,2667	0,2667
	0	0	0,3333	0,3333
0		0,0667	0,3333	0,2667
0		0,1333	0,3333	0,2
0		0,2	0,3333	0,1333
	1	0,2	0,4	0,2
1		0,2667	0,4	0,1333
1		0,3333	0,4	0,0667
	3	0,3333	0,4667	0,1333
3		0,4	0,4667	0,0667
3		0,4667	0,4667	0
	4	0,4667	0,5333	0,0667
	4	0,4667	0,6	0,1333
4		0,5333	0,6	0,0667
4		0,6	0,6	0
4		0,6667	0,6	0,0667
	5	0,6667	0,6667	0
5		0,7333	0,6667	0,0667
	7	0,7333	0,7333	0
11		0,8	0,7333	0,0667
	21	0,8	0,8	0
	21	0,8	0,8667	0,0667
23		0,8667	0,8667	0
	83	0,8667	0,9333	0,0667
	94	0,8667	1	0,1333
166		0,9333	1	0,0667
439		1	1	0

Table C.39: SMIR's table for the number of personal links

C.14 The number of self-photos

The problem here is to test if there is a location difference between the males and the females regarding the number of self-photos. In table C.40, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 231,5 and for the females 233,5. U_m (the males) is 111,5 and U_n (the females) is 113,5. Since the lowest value (111,5) is situated above the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Value	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	0	0	0	0	0
Rank	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1
Value	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	$\underline{1}$	1	1	1	1	$\underline{2}$
Rank	11,11	11,11	11,11	11,11	11,11	11,11	11,11	11,11	11,11	20,25
Value	2	2	2	$\underline{3}$	3	$\underline{5}$	6	$\underline{8}$	$\underline{21}$	89
Rank	20,25	20,25	20,25	24,5	24,5	26	27	28	29	30

Table C.40: WMW's table for the number of self-photos

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 2,93 self-photos whereas it is 7,2 self-photos for the females. In table C.41, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T=6760,95. The mean of the squared ranks for all thirty observations is 283,18. Here S equals 735,21. Thus Z = (6760,95 - 15*283,18)/735,21 equals 3,42. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Deviation	$\underline{0,07}$	$\underline{0,93}$	1,2	$\underline{1,93}$	$\underline{1,93}$	$\underline{1,93}$	$\underline{1,93}$	$\underline{1,93}$	$\underline{2,07}$	$\underline{2,93}$
Rank	1	2	3	4,2	4,2	4,2	4,2	4,2	9	10,2
Square	1	4	9	17,64	17,64	17,64	17,64	17,64	81	104,04
Deviation	$\underline{2,93}$	$\underline{2,93}$	$\underline{2,93}$	$\underline{2,93}$	4,2	$\underline{5,07}$	5,2	5,2	6,2	6,2
Rank	10,2	10,2	10,2	10,2	15	16	17,5	17,5	19,25	19,25
Square	104,04	104,04	104,04	104,04	225	256	306,25	306,25	370,56	370,56
Deviation	6,2	6,2	7,2	7,2	7,2	7,2	7,2	$\underline{18,07}$	52	81,8
Rank	19,25	19,25	23,2	23,2	23,2	23,2	23,2	28	29	30
Square	370,56	370,56	538,24	538,24	538,24	538,24	538,24	784	841	900

Table C.41: SR table for the number of self-photos

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.42, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,3333 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.15 The number of photos

The problem here is to test if there is a location difference between the males and the females regarding the number of photos. In table C.43, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 283,5 and for the females 163,5. U_m (the males) is 61,5 and U_n (the females) is 113,5. Since the lowest value (61,5) is situated below the critical value, we **can reject** the hypothesis that both distributions have the same location. There is indeed a location difference situated at 1 (photo) and the confidence

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
	0	0	0,1333	0,1333
	0	0	0,2	0,2
	0	0	0,2667	0,2667
	0	0	0,3333	0,3333
0		0,0667	0,3333	0,2667
0		0,1333	0,3333	0,2
0		0,2	0,3333	0,1333
0		0,2667	0,3333	0,0667
0		0,3333	0,3333	0
	1	0,3333	0,4	0,0667
	1	0,3333	0,4667	0,1333
	1	0,3333	0,5333	0,2
	1	0,3333	0,6	0,2667
	1	0,3333	0,6667	0,3333
1		0,4	0,6667	0,2667
1		0,4667	0,6667	0,2
1		0,5333	0,6667	0,1333
1		0,6	0,6667	0,0667
	2	0,6	0,7333	0,1333
2		0,6667	0,7333	0,0667
2		0,7333	0,7333	0
2		0,8	0,7333	0,0667
	3	0,8	0,8	0
3		0,8667	0,8	0,0667
	5	0,8667	0,8667	0
6		0,9333	0,8667	0,0667
	8	0,9333	0,9333	0
	21	0,9333	1	0,0667
89		1	1	0

Table C.42: SMIR's table for the number of self-photos

Value	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	0	0	0	0
Rank	1,05	1,05	1,05	1,05	1,05	1,05	1,05	1,05	1,05	1,05
Value	0	0	0	0	0	0	0	0	0	$\underline{1}$
Rank	1,05	1,05	1,05	1,05	1,05	1,05	1,05	1,05	1,05	20
Value	$\underline{2}$	2	$\underline{4}$	$\underline{10}$	$\underline{11}$	$\underline{14}$	$\underline{28}$	$\underline{90}$	$\underline{116}$	1083
Rank	21,5	21,5	23	24	25	26	27	28	29	30

Table C.43: WMW's table for the number of photos

interval for that difference is (0,11)[1].

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 18,4 photos whereas it is 72,33 photos for the females. In table C.44, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for

[1] This result has been obtained thanks to a Java program you can find in appendix B

Deviation	$\underline{4,4}$	$\underline{7,4}$	$\underline{8,4}$	$\underline{9,6}$	$\underline{14,4}$	$\underline{16,4}$	$\underline{17,4}$	$\underline{18,4}$	$\underline{18,4}$	$\underline{18,4}$
Rank	1	2	3	4	5	6	7	8,17	8,17	8,17
Square	1	4	9	16	25	36	49	66,69	66,69	66,69
Deviation	$\underline{18,4}$	$\underline{18,4}$	$\underline{18,4}$	70,33	$\underline{71,6}$	72,33	72,33	72,33	72,33	72,33
Rank	8,17	8,17	8,17	14	15	16,08	16,08	16,08	16,08	16,08
Square	66,69	66,69	66,69	196	225	258,47	258,47	258,47	258,47	258,47
Deviation	72,33	72,33	72,33	72,33	72,33	72,33	72,33	72,33	$\underline{97,6}$	1010,67
Rank	16,08	16,08	16,08	16,08	16,08	16,08	16,08	16,08	29	30
Square	258,47	258,47	258,47	258,47	258,47	258,47	258,47	258,47	841	900

Table C.44: SR table for the number of photos

the females is T=4456,08. The mean of the squared ranks for all thirty observations is 202,07. Here S equals 575,61. Thus Z = (4456,08 - 15*202,07)/575,61 equals 2,48. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.45, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,47 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

C.16 The number of graphics

The problem here is to test if there is a location difference between the males and the females regarding the number of graphics. In table C.46, you will find all sample values and the associated ranks. The underlined values belong to the male sample. Here the sum of the ranks for the males is 252,5 and for the females 212,5. U_m (the males) is 132,5 and U_n (the females) is 92,5. Since the lowest value (92,5) is situated below the critical value, we **cannot reject** the hypothesis that both distributions have the same location.

Now we are going to test the equality of variances by the squared rank test for variance. The estimation of the mean for the males is 76,93 graphics whereas it is 305,27 graphics for the females. In table C.47, you can find the deviations, ranks and squares of these ranks. The sum of the squared ranks for the females is T = 7841,31. The mean of the squared ranks for all thirty observations is 311,48. Here S equals 754,30. Thus Z = (7841,31- 15*311,48)/754,30 equals 4,20. Since Z is above 1,96, we can conclude that the variances are **not equal**.

Let's now use the Smirnov test to determine if it is reasonable to assume that the two samples come from identically distributed populations. In table C.48, you will find the computation of the sample cumulative distribution functions S(f) and S(m) and the differences S(f) - S(m). The difference of greatest magnitude is 0,27 (final column). Since it is below the critical value, we **cannot reject** the null hypothesis saying that both samples come from identically distributed populations.

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
	0	0	0,1333	0,1333
	0	0	0,2	0,2
	0	0	0,2667	0,2667
	0	0	0,3333	0,3333
	0	0	0,4	0,4
0		0,0667	0,4	0,3333
0		0,1333	0,4	0,2667
0		0,2	0,4	0,2
0		0,2667	0,4	0,1333
0		0,3333	0,4	0,0667
0		0,4	0,4	0
0		0,4667	0,4	0,0667
0		0,5333	0,5	0,1333
0		0,6	0,4	0,2
0		0,6667	0,4	0,2667
0		0,7333	0,4	0,3333
0		0,8	0,4	0,4
0		0,8667	0,4	0,4667
	1	0,8667	0,4667	0,4
	2	0,8667	0,5333	0,3333
2		0,9333	0,5333	0,4
	4	0,9333	0,6	0,3333
	10	0,9333	0,6667	0,2667
	11	0,9333	0,7333	0,2
	14	0,9333	0,8	0,1333
	28	0,9333	0,8667	0,0667
	90	0,9333	0,9333	0
	116	0,9333	1	0,0667
1083		1	1	0

Table C.45: SMIR's table for the number of photos

Value	0	0	0	0	0	2	3	4	4	4
Rank	1,2	1,2	1,2	1,2	1,2	6	7	8,33	8,33	8,33
Value	6	6	6	12	14	14	17	25	30	42
Rank	11,33	11,33	11,33	14	15,5	15,5	17	18	19	20
Value	64	74	83	97	172	286	341	355	1217	2855
Rank	21	22	23	24	25	26	27	28	29	30

Table C.46: WMW's table for the number of graphics

Deviation	2,93	6,07	20,07	34,93	49,73	59,93	62,93	64,93	70,93	70,93
Rank	1	2	3	4	5	6	7	8	9,5	9,5
Square	1	4	9	16	25	36	49	64	90,25	90,25
Deviation	72,93	76,93	76,93	95,07	209,07	241,27	264,07	275,27	280,27	291,27
Rank	11	12,5	12,5	14	15	16	17	18	19	20
Square	121	156,25	156,25	196	225	256	289	324	361	400
Deviation	299,27	301,27	301,27	302,27	303,27	305,27	305,27	305,27	911,73	2549,73
Rank	21	22,5	22,5	24	25	26,33	26,33	26,33	29	30
Square	441	506,25	506,25	576	625	693,27	693,27	693,27	841	900

Table C.47: SR table for the number of graphics

Females	Males	S(f)	S(m)	S(f) - S(m)
	0	0	0,0667	0,0667
	0	0	0,1333	0,1333
0		0,0667	0,1333	0,0667
0		0,1333	0,1333	0
0		0,2	0,1333	0,0667
2		0,2667	0,1333	0,1333
3		0,3333	0,1333	0,2
	4	0,3333	0,2	0,1333
4		0,4	0,2	0,2
4		0,4667	0,2	0,2667
	6	0,4667	0,2667	0,2
	6	0,4667	0,3333	0,1333
6		0,5333	0,3333	0,2
	12	0,5333	0,4	0,1333
	14	0,5333	0,4667	0,0667
14		0,6	0,4667	0,1333
	17	0,6	0,5333	0,0667
25		0,6667	0,5333	0,1333
30		0,7333	0,5333	0,2
	42	0,7333	0,6	0,1333
64		0,8	0,6	0,2
	74	0,8	0,0667	0,1333
	83	0,8	0,7333	0,0667
	97	0,8	0,8	0
	172	0,8	0,8667	0,0667
	286	0,8	0,9333	0,1333
	341	0,8	1	0,2
355		0,8667	1	0,1333
1217		0,9333	1	0,0667
2855		1	1	0

Table C.48: SMIR's table for the number of graphics

Appendix D

Fisher's and binomial tables

D.1 Tables for the analysis of the numerical variables

D.1.1 Neave's table for the Wilcoxon-Mann-Whitney's test

D.1.2 Table for the Smirnov's test

D.2 Tables for the analysis of the binary variables

D.2.1 Extract of the Fisher's tables from Biometrika

D.2.2 Extract of the Fisher's tables from Siegel's Nonparametric statistics for the behavioral sciences

Part I

Part II

D.2.3 Extract of the confidence limits tables for the binomial distribution

Appendix E

Results for the binary variables

In this appendix, you will find all the results of the Fisher's test, the binomial test and the multiple correspondence analysis for the binary variables.

E.1 The Fisher's test

Professional description	yes	no	total
M	10	5	15
F	9	6	15
Total	19	11	30

Since C is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Private description	yes	no	total
M	2	13	15
F	3	12	15
Total	5	25	30

Since D is above 7, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Technological website	yes	no	total
M	2	13	15
F	4	11	15
total	6	24	30

Since D is above 7, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Classic fonts	yes	no	total
M	15	0	15
F	14	1	15
total	29	1	30

Since C is above 11, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Girlish fonts	yes	no	total
M	0	15	15
F	1	14	15
total	1	29	30

Since D is above 11, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Black colour (txt)	yes	no	total
M	13	2	15
F	13	2	15
total	26	4	30

Since C is above 7, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

White colour (txt)	yes	no	total
M	2	13	15
F	3	12	15
total	5	25	30

Since D is above 7, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Grey colour (txt)	yes	no	total
M	1	14	15
F	3	12	15
total	4	26	30

Since D is above 9, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Links to other people's pages	yes	no	total
M	10	5	15
F	12	3	15
total	22	8	30

Since C is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Non-personal links	yes	no	total
M	14	1	15
F	15	0	15
total	29	1	30

Since C is above 9, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Self-photos	yes	no	total
M	10	5	15
F	10	5	15
total	20	10	30

Since C is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Official self-photos	yes	no	total
M	8	2	10
F	9	1	10
total	17	3	20

Since C is above 2, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Non-official self-photos	yes	no	total
M	2	8	10
F	1	9	10
total	3	17	20

Since D is above 2, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Family self-photos	yes	no	total
M	3	7	10
F	3	7	10
total	6	14	20

Since D is above 1, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Friends self-photos	yes	no	total
M	2	8	10
F	1	9	10
total	3	17	20

Since D is above 2, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Colleagues self-photos	yes	no	total
M	0	10	10
F	1	9	10
total	1	19	20

Since D is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Pets self-photos	yes	no	total
M	0	10	10
F	0	10	10
total	0	20	20

Since D is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Leisure time self-photos	yes	no	total
M	4	6	10
F	2	8	10
total	6	14	20

Since C is above 0, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Computer-related self-photos	yes	no	total
M	1	9	10
F	0	10	10
total	1	19	20

Since D is above 3, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Quality of self-photos	yes	no	total
M	10	0	10
F	10	0	10
total	20	0	20

Since C is above 6, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Family photos	yes	no	total
M	2	7	9
F	1	1	2
total	3	8	11

For your information, from this variable until the computer-related photos, we will use the Fisher's tables appearing in Siegel-Tukey. To know how to read these tables, please see appendix D where they are displayed. Regarding the result, since the total probability is 1,00 and thus above the 0,05 significance level, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Friends photos	yes	no	total
M	2	7	9
F	1	1	2
total	3	8	11

Since the total probability is 1,00 and thus above the 0,05 significance level, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Colleagues photos	yes	no	total
M	2	7	9
F	1	1	2
total	3	8	11

Since the total probability is 1,00 and thus above the 0,05 significance level, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Pets photos	yes	no	total
M	2	7	9
F	0	2	2
total	2	9	11

Since the total probability is 1,00 and thus above the 0,05 significance level, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Leisure time photos	yes	no	total
M	6	3	9
F	1	1	2
total	7	4	11

Since the total probability is 1,00 and thus above the 0,05 significance level, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Computer-related photos	yes	no	total
M	3	6	9
F	0	2	2
total	3	8	11

Since the total probability is 0,564 and thus above the 0,05 significance level, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Quality of photos	yes	no	total
M	9	0	9
F	1	1	2
total	10	1	11

Since C is above 0, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Basic graphics	yes	no	total
M	11	2	13
F	11	1	12
total	22	3	25

Since C is above 5, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Modern graphics	yes	no	total
M	1	12	13
F	2	10	12
total	3	22	25

Since D is above 6, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Trendy graphics	yes	no	total
M	2	11	13
F	2	10	12
total	4	21	25

Since D is above 5, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Artistic graphics	yes	no	total
M	2	11	13
F	3	9	12
total	5	20	25

Since D is above 5, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Comics	yes	no	total
M	3	10	13
F	2	10	12
total	5	20	25

Since D is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Computer-related graphics	yes	no	total
M	6	7	13
F	6	6	12
Total	12	13	25

Since D is above 1, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Soft colours for background	yes	no	total
M	6	9	15
F	6	9	15
Total	12	18	30

Since D is above 3, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Dark colours for background	yes	no	total
M	4	11	15
F	1	14	15
total	5	25	30

Since D is above 5, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Reddish colours (bck)	yes	no	total
M	7	8	15
F	5	10	15
total	12	18	30

Since D is above 2, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Blueish colours (bck)	yes	no	total
M	5	10	15
F	2	13	15
total	7	23	30

Since D is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Black colour (bck)	yes	no	total
M	3	12	15
F	1	14	15
total	4	26	30

Since D is above 6, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

White colour (bck)	yes	no	total
M	13	2	15
F	10	5	15
Total	23	7	30

Since C is above 7, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Grey colour (bck)	yes	no	total
M	5	10	15
F	5	10	15
Total	10	20	30

Since D is above 4, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Classic background	yes	no	total
M	14	1	15
F	15	0	15
total	29	1	30

Since C is above 9, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Original background	yes	no	total
M	3	12	15
F	2	13	15
total	5	25	30

Since D is above 6, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

Graphics	yes	no	total
M	13	2	15
F	12	3	15
total	25	5	30

Since C is above 7, we **cannot** reject the homogeneity hypothesis with an error risk of 5 %.

E.2 The binomial test

1. **The self-description**

   Pr[Having a self-description] = [0,51 ; 0,7]

2. **The professional description**

   Pr[Having a professional description] = [0,43 ; 0,63]

3. **The private description**

   Pr[Not having a private description] = [0,65 ; 0,83]

4. **The focus on credentials**

   Pr[Not focussing on credentials] = [0,47 ; 0,66]

5. **The personal pages (ratio)**

   Pr[Not having personal pages] = [0,41 ; 0,6]

6. **The technological website**

   Pr[Having a technological website] = [0,62 ; 0,8]

7. **The classic fonts**

   Pr[Having classic fonts] = [0,84 ; 0,96]

8. **The girlish fonts**

   Pr[Not having girlish fonts] = [0,84 ; 0,96]

9. **The reddish colours for text and hypertext**

   Pr[Using reddish colours] = [0,32 ; 0,5]

10. **The blueish colours for text and hypertext**

    Pr[Using blueish colours] = [0,78 ; 0,93]

11. **The black colour for text and hypertext**

    Pr[Using black] = [0,69 ; 0,86]

12. **The white colour for text and hypertext**

    Pr[Not using white] = [0,65 ; 0,83]

13. **The grey colour for text and hypertext**

    Pr[Not using grey] = [0,69 ; 0,86]

14. **The graphic accents**

    Pr[Not showing graphic accents] = [0,65 ; 0,83]

15. **The links to other people's pages**

    Pr[Including links to other people's pages] = [0,57 ; 0,73]

16. **The non-personal links**

    Pr[Including non-personal links] = [0,84 ; 0,96]

17. **The self-photos**

    Pr[Showing self-photos] = [0,47 ; 0,66]

18. **The photos**

    Pr[Not showing photos] = [0,43 ; 0,63]

19. **The graphics**

    Pr[Including graphics] = [0,65 ; 0,83]

20. **The soft colours for the background**

    Pr[Not having soft colours] = [0,41 ; 0,6]

21. **The dark colours for the background**

    Pr[Not having dark colours] = [0,65 ; 0,83]

22. **The reddish colours for the background**

    Pr[Not using reddish colours] = [0,41 ; 0,6]

23. **The blueish colours for the background**

    Pr[Not using blueish colours] = [0,58 ; 0,76]

24. **The black colour for the background**

    Pr[Not using black] = [0,69 ; 0,86]

25. **The white colour for the background**

    Pr[Having white] = [0,58 ; 0,76]

26. **The grey colour for the background**

    Pr[Not having grey] = [0,47 ; 0,66]

27. **The classic background**

    Pr[Having a classic background] = [0,84 ; 0,96]

28. **The original background**

    Pr[Not having an original background] = [0,65 ; 0,83]

E.3 The multiple correspondence analysis

Here is the full description of the first ten factors

Axis 1	
Negative	**Positive**
Photos: no	Photos: yes
Self-photos:no	Self-photos: yes
Leisure time self-photos: no	Leisure time self-photos: yes
Personal content (denoted by ratio): no	Personal content (denoted by ratio): yes

Axis 2	
Negative	**Positive**
Artistig graphics: no	Artistic graphics: yes
Blueish background: no	Blueish background: yes
Computer-related background: no	Computer-related graphics: yes
Family self-photos: yes	Family self-photos: no

Axis 3	
Negative	**Positive**
Girlish fonts: yes	Girlish fonts: no
Classic fonts: no	Classic fonts: yes
Not official self-photos: yes	Not official self-photos: no
Blueish colours (txt): no	Blueish colours (txt): yes

Axis 4	
Negative	**Positive**
Original background: no	Original background: yes
Trendy graphics: no	Trendy graphics: yes
White background: yes	White background: no
Leisure time photos: yes	Leisure time photos: no

Axis 5	
Negative	**Positive**
Comics graphics: no	Comics graphics: yes
Colleagues self-photos: no	Colleagues self-photos: yes
Grey (txt): no	Grey (txt): yes
Original background: yes	Original background: no

Axis 6	
Negative	**Positive**
Dark colours for background: no	Dark colours for background: yes
Private self-description: no	Private self-description: yes
Reddish colours (txt): no	Reddish colours (txt): yes
Grey for background: yes	Grey for background: no

Axis 7	
Negative	**Positive**
Highly-technological website: no	Highly-technological website: yes
White (txt): yes	White (txt): no
Modern graphics: yes	Modern graphics: no
Colleagues photos: no	Colleagues photos: yes

Axis 8	
Negative	**Positive**
Pets photos: yes	Pets photos: no
Dark colours for background: yes	Dark colours for background: no
Soft colours for background: no	Soft colours for background: yes
Black for background: yes	Black for background: no

Axis 9	
Negative	**Positive**
Family photos: no	Family photos: yes
Black (txt): yes	Black (txt): no

Axis 10	
Negative	**Positive**
Classic background: no	Classic background: yes
Pets photos: no	Pets photos: yes

Appendix F

Survey results

F.1 Results of the principal components analysis

Here is the full description of the 14 axes for the questions.

Axis 1				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q50	0,55	90,00	2,933	1,073
Q38	0,56	90,00	2,811	0,965
Q35	0,61	90,00	2,411	0,930
Q45	0,64	90,00	2,700	1,048
Q39	0,65	90,00	2,767	0,907
Q43	0,66	90,00	2,933	1,062
Q49	0,69	90,00	2,422	1,054
Q46	0,73	90,00	2,622	1,060
Q44	0,75	90,00	2,378	1,111
Q48	0,76	90,00	2,489	1,035
Q47	0,76	90,00	2,644	1,119

Axis 2				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q34	-0,63	90,00	3,344	0,871
Q10	-0,63	90,00	3,322	0,917
Q32	-0,58	90,00	3,500	0,980
Q40	-0,52	90,00	3,333	1,065
Q41	-0,49	90,00	3,511	0,969
Q19	-0,47	90,00	3,589	0,918
Q27	-0,45	90,00	3,267	0,786
Q25	-0,43	90,00	3,478	0,792
Q26	-0,42	90,00	3,300	0,888
Q13	-0,40	90,00	3,311	0,755
Q33	-0,36	90,00	3,344	0,871
Centre	*Centre*	*Centre*	*Centre*	*Centre*

Next part of table on next page...

Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q47	0,18	90,00	2,644	1,119
Q11	0,23	90,00	2,622	0,824
Q44	0,24	90,00	2,378	1,111
Q12	0,24	90,00	2,544	0,979
Q45	0,27	90,00	2,700	1,048
Q30	0,34	90,00	2,778	1,009
Q49	0,34	90,00	2,422	1,054
Q48	0,35	90,00	2,489	1,035
Q35	0,35	90,00	2,411	0,930
Q17	0,41	90,00	2,178	1,060
Q18	0,52	90,00	2,078	0,980

Axis 3				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q23	-0,37	90,00	3,211	0,782
Q25	-0,34	90,00	3,478	0,792
Q22	-0,34	90,00	3,433	0,700
Q43	-0,32	90,00	2,933	1,062
Q50	-0,31	90,00	2,933	1,073
Q33	-0,31	90,00	3,344	0,871
Q38	-0,30	90,00	2,811	0,965
Q45	-0,27	90,00	2,700	1,048
Q48	-0,21	90,00	2,489	1,035
Q28	-0,17	90,00	3,244	0,793
Q47	-0,16	90,00	2,644	1,119
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q27	0,32	90,00	3,267	0,786
Q18	0,34	90,00	2,078	0,980
Q4	0,34	90,00	2,378	0,902
Q15	0,34	90,00	3,033	0,983
Q17	0,38	90,00	2,178	1,060
Q20	0,39	90,00	3,233	0,943
Q16	0,40	90,00	2,911	0,890
Q3	0,42	90,00	2,900	0,831
Q5	0,42	90,00	2,944	1,026
Q9	0,43	90,00	2,689	0,985
Q14	0,44	90,00	3,056	0,794

Axis 4				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q44	-0,37	90,00	2,378	1,111
Q43	-0,36	90,00	2,933	1,062
Q4	-0,32	90,00	2,378	0,902
Q3	-0,32	90,00	2,900	0,831
Q16	-0,28	90,00	2,911	0,890

Next part of table on next page...

Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q5	-0,27	90,00	2,944	1,026
Q15	-0,23	90,00	3,033	0,983
Q21	-0,22	90,00	3,078	0,885
Q39	-0,22	90,00	2,767	0,907
Q48	-0,20	90,00	2,489	1,035
Q46	-0,19	90,00	2,622	1,060
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q18	0,23	90,00	2,078	0,980
Q24	0,24	90,00	3,289	1,003
Q17	0,25	90,00	2,178	1,060
Q29	0,30	90,00	3,011	0,901
Q30	0,36	90,00	2,778	1,009
Q50	0,36	90,00	2,933	1,073
Q9	0,36	90,00	2,689	0,985
Q26	0,37	90,00	3,300	0,888
Q12	0,40	90,00	2,544	0,979
Q27	0,41	90,00	3,267	0,786
Q28	0,58	90,00	3,244	0,793

Axis 5				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q29	-0,45	90,00	3,011	0,901
Q24	-0,44	90,00	3,289	1,003
Q41	-0,43	90,00	3,511	0,969
Q40	-0,36	90,00	3,333	1,065
Q5	-0,28	90,00	2,944	1,026
Q30	-0,18	90,00	2,778	1,009
Q39	-0,14	90,00	2,767	0,907
Q19	-0,12	90,00	3,589	0,918
Q23	-0,12	90,00	3,211	0,782
Q12	-0,10	90,00	2,544	0,979
Q44	-0,09	90,00	2,378	1,111
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q50	0,14	90,00	2,933	1,073
Q34	0,19	90,00	3,344	0,871
Q45	0,21	90,00	2,700	1,048
Q16	0,23	90,00	2,911	0,890
Q37	0,26	90,00	2,800	0,991
Q32	0,31	90,00	3,500	0,980
Q15	0,32	90,00	3,033	0,983
Q22	0,34	90,00	3,433	0,700
Q13	0,34	90,00	3,311	0,755
Q33	0,41	90,00	3,344	0,871
Q36	0,48	90,00	2,856	0,961

Axis 6				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q30	-0,35	90,00	2,778	1,009
Q33	-0,33	90,00	3,344	0,871
Q21	-0,32	90,00	3,078	0,885
Q11	-0,28	90,00	2,622	0,824
Q5	-0,26	90,00	2,944	1,026
Q12	-0,26	90,00	2,544	0,979
Q8	-0,23	90,00	3,256	0,811
Q38	-0,22	90,00	2,811	0,965
Q36	-0,21	90,00	2,856	0,961
Q9	-0,21	90,00	2,689	0,985
Q19	-0,19	90,00	3,589	0,918
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q16	0,10	90,00	2,911	0,890
Q3	0,17	90,00	2,900	0,831
Q43	0,18	90,00	2,933	1,062
Q50	0,21	90,00	2,933	1,073
Q28	0,22	90,00	3,244	0,793
Q10	0,22	90,00	3,322	0,917
Q49	0,23	90,00	2,422	1,054
Q35	0,26	90,00	2,411	0,930
Q37	0,40	90,00	2,800	0,991
Q31	0,43	90,00	3,033	0,875
Q14	0,51	90,00	3,056	0,794

Axis 7				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q23	-0,46	90,00	3,211	0,782
Q4	-0,37	90,00	2,378	0,902
Q45	-0,33	90,00	2,700	1,048
Q15	-0,32	90,00	3,033	0,983
Q3	-0,30	90,00	2,900	0,831
Q16	-0,28	90,00	2,911	0,890
Q36	-0,27	90,00	2,856	0,961
Q30	-0,22	90,00	2,778	1,009
Q40	-0,19	90,00	3,333	1,065
Q19	-0,19	90,00	3,589	0,918
Q50	-0,17	90,00	2,933	1,073
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q37	0,16	90,00	2,800	0,991
Q7	0,17	90,00	3,244	0,923
Q29	0,17	90,00	3,011	0,901
Q46	0,17	90,00	2,622	1,060
Q22	0,17	90,00	3,433	0,700
Q33	0,18	90,00	3,344	0,871
Q43	0,19	90,00	2,933	1,062
Q20	0,22	90,00	3,233	0,943

Next part of table on next page...

Variable	Coordinate	Weight	Mean	Std.Deviation
Q27	0,25	90,00	3,267	0,786
Q21	0,32	90,00	3,078	0,885
Q8	0,35	90,00	3,256	0,811

Axis 8				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q11	-0,51	90,00	2,622	0,824
Q22	-0,38	90,00	3,433	0,700
Q25	-0,29	90,00	3,478	0,792
Q4	-0,28	90,00	2,378	0,902
Q35	-0,28	90,00	2,411	0,930
Q13	-0,26	90,00	3,311	0,755
Q7	-0,19	90,00	3,244	0,923
Q31	-0,19	90,00	3,033	0,875
Q3	-0,18	90,00	2,900	0,831
Q40	-0,16	90,00	3,333	1,065
Q49	-0,15	90,00	2,422	1,054
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q34	0,12	90,00	3,344	0,871
Q21	0,13	90,00	3,078	0,885
Q26	0,13	90,00	3,300	0,888
Q36	0,17	90,00	2,856	0,961
Q46	0,18	90,00	2,622	1,060
Q50	0,20	90,00	2,933	1,073
Q20	0,25	90,00	3,233	0,943
Q15	0,26	90,00	3,033	0,983
Q19	0,27	90,00	3,589	0,918
Q47	0,30	90,00	2,644	1,119
Q42	0,31	90,00	3,100	1,012

Axis 9				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q20	-0,45	90,00	3,233	0,943
Q7	-0,36	90,00	3,244	0,923
Q14	-0,33	90,00	3,056	0,794
Q23	-0,30	90,00	3,211	0,782
Q39	-0,24	90,00	2,767	0,907
Q37	-0,23	90,00	2,800	0,991
Q9	-0,22	90,00	2,689	0,985
Q30	-0,19	90,00	2,778	1,009
Q19	-0,19	90,00	3,589	0,918
Q28	-0,16	90,00	3,244	0,793
Q22	-0,16	90,00	3,433	0,700
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q26	0,13	90,00	3,300	0,888

Next part of table on next page...

Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q46	0,14	90,00	2,622	1,060
Q8	0,17	90,00	3,256	0,811
Q11	0,20	90,00	2,622	0,824
Q13	0,21	90,00	3,311	0,755
Q32	0,21	90,00	3,500	0,980
Q45	0,22	90,00	2,700	1,048
Q18	0,22	90,00	2,078	0,980
Q31	0,24	90,00	3,033	0,875
Q5	0,27	90,00	2,944	1,026
Q10	0,32	90,00	3,322	0,917

Axis 10				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q37	-0,36	90,00	2,800	0,991
Q30	-0,29	90,00	2,778	1,009
Q32	-0,25	90,00	3,500	0,980
Q38	-0,23	90,00	2,811	0,965
Q26	-0,23	90,00	3,300	0,888
Q18	-0,17	90,00	2,078	0,980
Q27	-0,17	90,00	3,267	0,786
Q41	-0,16	90,00	3,511	0,969
Q5	-0,16	90,00	2,944	1,026
Q33	-0,15	90,00	3,344	0,871
Q21	-0,15	90,00	3,078	0,885
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q49	0,13	90,00	2,422	1,054
Q3	0,13	90,00	2,900	0,831
Q31	0,16	90,00	3,033	0,875
Q25	0,17	90,00	3,478	0,792
Q12	0,19	90,00	2,544	0,979
Q50	0,20	90,00	2,933	1,073
Q11	0,24	90,00	2,622	0,824
Q28	0,26	90,00	3,244	0,793
Q7	0,28	90,00	3,244	0,923
Q19	0,37	90,00	3,589	0,918
Q8	0,40	90,00	3,256	0,811

Axis 11				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q42	-0,37	90,00	3,100	1,012
Q13	-0,33	90,00	3,311	0,755
Q23	-0,23	90,00	3,211	0,782
Q19	-0,23	90,00	3,589	0,918
Q5	-0,21	90,00	2,944	1,026
Q14	-0,21	90,00	3,056	0,794

Next part of table on next page...

Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q41	-0,18	90,00	3,511	0,969
Q39	-0,18	90,00	2,767	0,907
Q35	-0,17	90,00	2,411	0,930
Q45	-0,16	90,00	2,700	1,048
Q27	-0,16	90,00	3,267	0,786
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q9	0,12	90,00	2,689	0,985
Q25	0,15	90,00	3,478	0,792
Q47	0,15	90,00	2,644	1,119
Q46	0,16	90,00	2,622	1,060
Q31	0,18	90,00	3,033	0,875
Q38	0,21	90,00	2,811	0,965
Q29	0,23	90,00	3,011	0,901
Q34	0,26	90,00	3,344	0,871
Q12	0,27	90,00	2,544	0,979
Q15	0,27	90,00	3,033	0,983
Q16	0,32	90,00	2,911	0,890

Axis 12				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q16	-0,27	90,00	2,911	0,890
Q24	-0,23	90,00	3,289	1,003
Q19	-0,23	90,00	3,589	0,918
Q20	-0,22	90,00	3,233	0,943
Q23	-0,21	90,00	3,211	0,782
Q4	-0,18	90,00	2,378	0,902
Q21	-0,18	90,00	3,078	0,885
Q29	-0,18	90,00	3,011	0,901
Q26	-0,15	90,00	3,300	0,888
Q49	-0,14	90,00	2,422	1,054
Q22	-0,10	90,00	3,433	0,700
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q38	0,10	90,00	2,811	0,965
Q15	0,12	90,00	3,033	0,983
Q50	0,12	90,00	2,933	1,073
Q30	0,15	90,00	2,778	1,009
Q14	0,16	90,00	3,056	0,794
Q36	0,19	90,00	2,856	0,961
Q39	0,20	90,00	2,767	0,907
Q18	0,22	90,00	2,078	0,980
Q41	0,30	90,00	3,511	0,969
Q40	0,34	90,00	3,333	1,065
Q7	0,56	90,00	3,244	0,923

Axis 13				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q28	-0,32	90,00	3,244	0,793
Q32	-0,30	90,00	3,500	0,980
Q12	-0,26	90,00	2,544	0,979
Q16	-0,22	90,00	2,911	0,890
Q29	-0,19	90,00	3,011	0,901
Q5	-0,17	90,00	2,944	1,026
Q22	-0,16	90,00	3,433	0,700
Q13	-0,16	90,00	3,311	0,755
Q11	-0,13	90,00	2,622	0,824
Q10	-0,13	90,00	3,322	0,917
Q43	-0,12	90,00	2,933	1,062
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q20	0,12	90,00	3,233	0,943
Q15	0,14	90,00	3,033	0,983
Q17	0,18	90,00	2,178	1,060
Q26	0,19	90,00	3,300	0,888
Q18	0,20	90,00	2,078	0,980
Q31	0,22	90,00	3,033	0,875
Q24	0,26	90,00	3,289	1,003
Q4	0,26	90,00	2,378	0,902
Q25	0,27	90,00	3,478	0,792
Q33	0,30	90,00	3,344	0,871
Q8	0,30	90,00	3,256	0,811

Axis 14				
Variable	**Coordinate**	**Weight**	**Mean**	**Std.Deviation**
Q15	-0,29	90,00	3,033	0,983
Q41	-0,25	90,00	3,511	0,969
Q11	-0,23	90,00	2,622	0,824
Q37	-0,23	90,00	2,800	0,991
Q3	-0,22	90,00	2,900	0,831
Q9	-0,22	90,00	2,689	0,985
Q25	-0,18	90,00	3,478	0,792
Q13	-0,16	90,00	3,311	0,755
Q21	-0,16	90,00	3,078	0,885
Q47	-0,12	90,00	2,644	1,119
Q40	-0,11	90,00	3,333	1,065
Centre	*Centre*	*Centre*	*Centre*	*Centre*
Q26	0,10	90,00	3,300	0,888
Q46	0,11	90,00	2,622	1,060
Q44	0,11	90,00	2,378	1,111
Q16	0,13	90,00	2,911	0,890
Q7	0,17	90,00	3,244	0,923
Q23	0,17	90,00	3,211	0,782
Q5	0,18	90,00	2,944	1,026
Q34	0,20	90,00	3,344	0,871

Next part of table on next page...

Variable	Coordinate	Weight	Mean	Std.Deviation
Q32	0,26	90,00	3,500	0,980
Q39	0,35	90,00	2,767	0,907
Q17	0,37	90,00	2,178	1,060