

Dictaat Statistische Methoden

Arnoud Vermeij
arnoudve@sci.ru.nl

December 23, 2005

Contents

1	Introductie	5
1.1	Ter inleiding	5
1.2	Opbouw van deze reader	5
2	Positionering en Voorbereiden van Data	7
2.1	Data Classificeren	7
2.2	Intermezzo: Kwalitatief versus Kwantitatief	8
2.3	Data Coderen	9
2.4	Data Manipuleren	10
2.5	Intermezzo: Soorten meetniveaus	11
2.6	Handmatig of machinaal?	12
2.7	Tot slot	13
3	Analyse met één variabele	15
3.1	Inleiding	15
3.2	Frequentieverdeling	15
3.3	Centrale Tendentie	16
3.3.1	Ratio- en Intervalvariabelen	16
3.3.2	Ordinale Variabelen	16
3.3.3	Alle Variabelen	17
3.4	Spreidingsmaten	17
3.4.1	Variatiebreedte	17
3.4.2	Standaarddeviatie	18
3.4.3	Kwartielverdeling	19
3.4.4	Vijf-Getallen-Resumé	19
3.5	Tenslotte	20
4	Analyse met twee variabelen	21
4.1	Inleiding	21
4.2	Twee Nominale Variabelen	21
4.3	Twee Intervalvariabelen	23
4.4	Eén Nominale en één Intervalvariabele	24
4.4.1	De T-Toets	24
4.4.2	Eenweg Variantie-analyse	25

4.5	Tenslotte	25
5	Analyse met meerdere variabelen	27
5.1	Inleiding	27
5.2	Elaboratietechniek	27
5.2.1	Uitkomsten bij toepassen Elaboratietechniek	28
5.3	Multivariabele Analysetechnieken: een Overzicht	29
5.3.1	Loglineaire Analyse	29
5.3.2	Discriminantanalyse	30
5.3.3	Logistische regressie	30
5.3.4	Variantie-analyse	30
5.3.5	Multipele regressie	30
5.3.6	MANOVA-technieken	30
5.3.7	Pad-Analyse	31
5.3.8	Combinatietechnieken	31
5.4	Dimensietechnieken	31
5.5	Tenslotte	31
6	Cases	33
6.1	Inleiding	33
6.2	Casus 1: de Webstats4U-teller	34
6.2.1	De Teller in het Algemeen	34
6.2.2	Een Specifieke Teller	34
6.2.3	Een Foutje of een Andere Interpretatie?	38
6.3	Semi-Casus 2: Padanalyse en Kunstmatige Intelligentie	38
6.3.1	Neurale Netwerken	39
6.3.2	Agent-based Information Retrieval	40
6.4	Tenslotte	40
7	Conclusie	41
7.1	Ter afsluiting	41
7.2	Ter conclusie	42
A	Data startpagina.nl	43
	Bibliography	59

Chapter 1

Introductie

1.1 Ter inleiding

Statistiek is een zeer uitgebreid onderwerp. Toch vormt het in dit geval slechts een klein onderdeel van een kleine cursus in het Informatiekundecurriculum. Dit in tegenstelling tot bijvoorbeeld een opleiding als Psychologie aan de RU Nijmegen, waar alleen al in de eerste twee jaren 33 EC wordt besteed aan onderzoeksmethoden en -vaardigheden, waarvan 12 puur aan statistiek [4].

Omdat het vrijwel onmogelijk is om in één klein onderdeel van een vak van 3 EC dezelfde hoeveelheid statistische methoden te behandelen als in zo'n opleiding, is hier de keuze gevallen op het geven van een overzicht van beschikbare statistische methoden en technieken. Zo kun je toch inzicht krijgen in de statistische methoden en technieken en waar en wanneer je ze toepast, zonder dat je hiervan meteen alle ins en outs kent. Degene die de diverse onderwerpen toch verder uitgediept wil zien, wordt verwezen naar de cursus 'Statistiek voor Informatiekunde' [15] en de bijbehorende reader [14].

Tijdens de behandeling van de meestal toch vrij droge en abstracte materie, zullen kleine, concrete voorbeeldjes worden gegeven om zo de stof wat tastbaarder en hopelijk begrijpbaarder te maken.

1.2 Opbouw van deze reader

Deze reader is als volgt opgebouwd: allereerst wordt in hoofdstuk 2 ingegaan op het positioneren van statistiek binnen onderzoeksmethoden en op het voorbereiden van data voor statistische analyse. In hoofdstuk 3 zal vervolgens de meest eenvoudige vorm van statistiek aan de orde komen, namelijk die met één variabele, en in hoofdstuk vier worden statistische methoden met twee variabelen onder de loep genomen. Vervolgens wordt in hoofdstuk vijf gekeken naar statistische methoden met meer dan twee variabelen. Hoofdstuk zes bevat een casus, toegespitst op Informatiekunde, ter concretisering van (een onderdeel van) de stof. Tenslotte zullen in hoofdstuk zeven enkele conclusies worden getrokken.

Chapter 2

Positionering en Voorbereiden van Data

Normaliter gaat data in een onderzoek door de volgende vijf fasen heen:

- Data verzamelen
- Data classificeren
- Data coderen
- Data manipuleren
- Data interpreteren

Het verzamelen van data gebeurt middels diverse methoden, zoals bijvoorbeeld surveys of interviews. Die onderwerpen komen aan bod in andere presentaties en dictaten en zullen hier dan ook niet verder worden behandeld.

Op een gegeven moment heb je middels die methoden een bepaalde hoeveelheid data verzameld. Om daar goed mee te kunnen werken zal de data geïnterpreteerd moeten worden. Daarvoor is het noodzakelijk dat de data eerst zodanig bewerkt wordt dat dit ook echt goed mogelijk is. Daarvoor zijn classificeren en coderen. We zullen hieronder kort kijken naar wat classificeren en coderen inhoudt.

2.1 Data Classificeren

Het doel van het classificeren van data is de data kunnen opdelen in ‘hapklare’, verwerkbare brokken. Als je niet goed classificeert, bestaat de kans dat je later, tijdens het interpreteren van de data, tegen problemen aanloopt. Bijvoorbeeld verkeerde interpretaties wegens onvolledigheid van de data, of wegens het toevoegen van verkeerde data. Soms is het vrij eenvoudig om data te classificeren: stel je hebt van mensen hun zwemvaardigheid onderzocht, uitgedrukt in termen van behaalde zwemdiploma’s. Je kunt nu makkelijk categorieën opstellen om

de data in te categoriseren: A, B, Basis, Zwemvaardigheid 1, 2 en 3. Het wordt al lastiger als je een enorme diversiteit hebt in je data, bijvoorbeeld als je alle diersoorten die in China voorkomen wilt indelen. Segers [13] haalt hierbij een artikel van Foucalut aan, die op zijn beurt weer een ‘een zekere Chinese encyclopedie’ aanhaalt. Wat wil het geval: de encyclopedie bevat een dergelijke classificering, te weten: a) eigendom van de Keizer, b) gebalsemd, c) tam, d) speenvarkens, e) sirenes, n) lijkend op een vlieg vanuit de verte, enzovoort. Gevoelsmatig zul je al aangeven dat dit niet de handigste manier is om die dieren te classificeren. Wat moet je bijvoorbeeld met een tam, gebalsemd speenvarken van de Keizer?

Tegenwoordig zijn er dan ook een viertal classificatiecriteria opgesteld waaraan een goede classificatie moet voldoen. Dit zijn [13]:

- Uitputtend. De classificatie moet zoveel mogelijk data kunnen verwerken. Het liefst zelfs alle data, zodat er, na het indelen van de data geen elementen van je verzameling data meer over zijn die eigenlijk in geen enkele categorie passen. Oftewel; je houdt uiteindelijk zo weinig mogelijk (het liefst zelfs helemaal niets) om in de categorie ‘overige’ te stoppen.
- Wederzijds Uitsluitend. De classificatie moet zó zijn ingericht, dat ieder element van je verzamelde data in slechts één categorie past.
- Eenduidig indelingsperspectief. De categorieën dienen consistent gekozen te zijn. Dus bijvoorbeeld op de vraag: ‘hoe vaak eet je per week Chinees’ zijn bijvoorbeeld a) 0, b) 1 of 2 keer, c) 3 of 4 keer d) 5 of meer keer zinvolle antwoorden. Antwoord e) ‘ik hou meer van friet’ zou hier een onzinnig onderdeel van de antwoordmogelijkheden zijn: het heeft totaal niks te maken met de gestelde vraag.
- zinnige indeling vanuit gekozen perspectief. Het heeft weinig zin om bij bovenstaande vraag slechts de antwoordmogelijkheden: a) 6 keer of meer en b) minder dan 6 keer te geven. De kans is erg groot dat de meerderheid van de mensen hier antwoord b zal kiezen.

Meer informatie over de laatste twee criteria zal ongetwijfeld worden aangehaald bij de readers over Surveys, Vragenlijsten en Schaalmethoden.

Op het moment dat je al je data hebt geclassificeerd, heb je een ingevulde datamatrix. Daarmee gaan we aan de slag bij het coderen van data.

2.2 Intermezzo: Kwalitatief versus Kwantitatief

We zijn tot nu toe uitgegaan van ‘data’ in het algemeen. Toch is het, zeker met het oog op de rest van deze reader, nuttig om onderscheid te maken tussen twee verschillende soorten data: kwalitatief en kwantitatief. Kwantitatieve data wordt bijvoorbeeld verkregen middels surveys (gesloten vragen, afgebakende antwoordmogelijkheden, veel respondenten), kwalitatieve data bijvoorbeeld via

interviews (open vragen, geen afgebakende antwoordmogelijkheden, weinig respondenten). Voor het werken met statistieken is vooral kwantitatieve data interessant. Statistische methoden maken immers vooral gebruik van kwantitatieve data: data die je kunt uitdrukken in harde getallen, waarmee gerekend kan worden. En rekenen aan data is immers de hoofdmoot van statistiek (wat immers een method voor kwantitatieve data-analyse is).

Het is echter niet zo dat kwalitatieve data per definitie ongeschikt is om middels statistiek geïnterpreteerd te worden. Er zijn alleen wel wat aanpassingen aan kwalitatieve data nodig om deze geschikt te maken voor kwantitatieve analyse: je moet deze kwantificeerbaar maken. Stel, bijvoorbeeld, dat je vijf mensen geïnterviewd hebt over een bepaald onderwerp en dat je van die vijf mensen een mening hebt, en dat je die mening ook hebt uitgeschreven. Dan heb je vijf lappen tekst voor je liggen. Door wat meta-data aan de teksten toe te voegen wordt de data al snel kwantificeerbaar. Hiervan onderscheiden we twee typen (die overigens ook goed in combinatie met elkaar gebruikt kunnen worden):

- Voeg trefwoorden toe. Door het toevoegen van trefwoorden aan een tekst kunnen gezichtspunten worden ontwikkeld, waarmee naar de teksten gekeken kan worden. Zo is bijvoorbeeld aan de hand van trefwoorden vrij snel duidelijk welke elementen vaak terugkomen in de antwoorden.
- Datareductie. Een stuk tekst in natuurlijke taal is vrij complex om te doorgronden. Om de complexiteit te reduceren kan het mogelijk zijn om de tekst in stukken op te delen en die stukken zelf te classificeren. Op die manier wordt de complexiteit, maar ook vooral de hoeveelheid data gereduceerd.

Eén van de risico's van het kwantificeren van kwalitatieve data is *context-stripping*. Context stripping houdt in dat de tekst in zulke kleine stukken data wordt gedeeld, of op een dusdanig onhandige manier in stukken wordt verdeeld, dat hele stukken tekst eigenlijk uit de context worden gehaald en daarmee een stuk van hun betekenis, en dus hun waarde, verliezen.

2.3 Data Coderen

Nadat de data is geclassificeerd, en de datamatrix is opgesteld, wordt het zaak om de data in de matrix te coderen. Dit houdt in dat je alle variabelen en categorieën in je datamatrix gaat voorzien van een code (een getal), om zo in eerste instantie de data makkelijker te kunnen verwerken met een computer. Een computer is immers vooral een rekenmachine, en geen natuurlijke taalverwerker. Het is uiteraard aan te raden om de gebruikte codes logisch en consistent te houden. Heb je bijvoorbeeld een datamatrix waarbij van vijf ja/nee-vragen de antwoorden zijn opgenomen, kies dan bijvoorbeeld 1 voor ja en 0 voor nee. Heb je eenmaal die keuze gemaakt, houd dit dan ook vol. Dus wijk niet opeens bij vraag drie af door daar 1 voor nee en 0 voor ja te kiezen.

Een voorbeeld van hoe zo'n codering er uit kan zien (merk hier overigens bij

op dat het getal 0 (of soms het getal 9) vaak de standaard is voor antwoorden als ‘overige’, ‘onbekend’, ‘geen mening’, ‘weet niet’, etc.) staat op de volgende pagina.

```
CODEBOOK
var01: Vraag 1: Bent u het eens met dit kabinetsbeleid?
1: ja
2: nee
0: weet niet / geen mening
```

De codes zelf neem je op in een *codeboek*. Een codeboek is feitelijk niks anders dan een verzameling codes en hun betekenis in de context van een onderzoek. Met een codeboek kun je, mits het goed is opgesteld, makkelijk de data ‘terugvertalen’ naar begrijpelijke, zinvolle taal, in plaats van getallen.

Naast verwerkingssnelheid van data middels computerondersteuning kan code ook een rol spelen bij het geheim houden van bepaalde informatie. Voor meer informatie over coderen en decoderen van informatie en de voor- en nadelen van geheimhouding van informatie kun je terecht bij het college ‘Security’ [15].

2.4 Data Manipuleren

Zodra je klaar bent met het classificeren en coderen van data, heb je een ingevulde, gecodeerde datamatrix. Het kan echter goed zijn dat deze matrix nog niet optimaal of nog niet volledig is, of juist overvolledig. Om later met de juiste data te kunnen werken is het uiteraard noodzakelijk dat de beschikbare set data zo optimaal mogelijk is. Daarom zullen de datamatrix en het codeboek nog aangepast moeten worden, oftewel; de data moet nog eenmaal worden gemanipuleerd.

Ook hierbij onderscheiden we een aantal vormen, te weten[13]:

- samenvoegen
- datareductie
- hercoderen

Van samenvoegen bestaan twee vormen, te weten:

- Dezelfde waarnemingen voor nieuwe onderzoekselementen. Het kan gebeuren dat een onderzoek dat is uitgevoerd zoveel data oplevert, dat er meerdere mensen nodig zijn om de data in de computer in te voeren. Het kan lastig zijn om met meerdere mensen tegelijkertijd aan één en hetzelfde databestand te werken. Een mogelijkheid is dan dat, zodra alle databestanden gevuld zijn, deze samen worden gevoegd.
- Nieuwe waarnemingen voor dezelfde onderzoekselementen. Ook kan het voorkomen dat een onderzoek op meerdere momenten wordt uitgevoerd. Tussen het invoeren van de eerder verworven data en het verkrijgen van de

nieuwe data kan dan wat tijd zitten, waardoor een al bestaand en in feite ‘afgerond’ databestand alsnog kan worden aangevuld met nieuwe data.

Ook is het mogelijk dat een databestand niet wordt uitgebreid, maar juist in omvang wordt gereduceerd. Hiervoor is een goede reden te bedenken: overzicht bewaren. Het kan voorkomen dat uit een onderzoek dusdanig veel gegevens komen, en dusdanig veel categorieën gegevens, dat het geheel van data onhanteerbaar wordt. Op zo’n moment is het handig om nog eens naar de onderzoeksvraag te kijken en opnieuw te focussen op wat nou eigenlijk de kern van het onderzoek is. Waarschijnlijk blijkt dan dat een flink aantal variabelen of categorieën (tijdelijk) kunnen worden geschrapt. Op die manier kun je het overzicht bewaren.

Dit probleem, het verlies van overzicht, doet zich al relatief snel voor: bedenk immers dat je, als je tweehonderd variabelen in een kruistabel tegen elkaar uitzet, je al bijna twintigduizend dwarsverbanden hebt! Het zal weinig moeite kosten om te bedenken dat dat al vrij onhandelbaar is. Bedenk ook dat tweehonderd een vrij reëel getal is wat het aantal variabelen in een onderzoek betreft.

Tenslotte is er nog hercoderen als vorm van datamanipulatie. Hercoderen komt vooral voor als blijkt dat de gekozen codering nog net niet optimaal is, of als, bijvoorbeeld door voortschrijdend inzicht in de materie, bepaalde variabelen of categorieën van variabelen geclusterd of gehergroepeerd kunnen worden (zonder dat dit (de resultaten van) het onderzoek aantast, natuurlijk).

Het resultaat van al je inspanningen: een ingevulde, goed gecodeerde datamatrix, klaar om geanalyseerd en geïnterpreteerd te worden.

2.5 Intermezzo: Soorten meetniveaus

We hebben tot nu toe gesproken over variabelen in de algemene zin van het woord. Toch is het noodzakelijk om onderscheid te maken tussen verschillende soorten variabelen. Iedere soort variabele is namelijk verbonden aan een soort meetniveau en heeft derhalve zijn eigen statistische methoden. Het spreekt voor zich dat een verkeerde combinatie van methode en meetniveau tot een verkeerde uitkomst leidt. Dat is uiteraard ongewenst.

We zullen hieronder de vier verschillende soorten meetniveaus kort bekijken:

- **Nominaal meetniveau** Het nominale meetniveau wordt gebruikt om aan te geven of een entiteit gelijk is aan een andere entiteit. Bijvoorbeeld $10 = 10$, en man $\neq$ vrouw, terwijl $\text{lengte} = 181 \text{ centimeter}$, $\text{kleur} = \text{rood}$ en $\text{temperatuur} = 37 \text{ graden Celsius}$ ook voorbeelden zijn van nominale variabelen. Een variabele van nominaal meetniveau zegt dus iets over de identiteit van het onderzochte. We spreken hier dan ook wel, in wiskundige termen, over een identiteitsrelatie.
- **Ordinaal meetniveau** Een variabele van ordinaal meetniveau geeft verhoudingen tussen twee of meer entiteiten weer. Bijvoorbeeld $a < b < c$, of $33 > 25$. We gebruiken dit meetniveau voor het ordenen van elementen. We spreken dan ook wel over een orderelatie.

- **Intervalmeetniveau** Bij het meten op intervalmeetniveau onderzoek je vooral de mate van verschil tussen variabelen, zonder dat je een natuurlijk nulpunt hanteert om je bevindingen aan te spiegelen. Stel bijvoorbeeld: Jan heeft een lengte van 180, Piet heeft een lengte van 170 en Kees heeft een lengte van 160 cm. Het verschil tussen zowel Jan als Piet, als tussen Piet en Kees is 10. Toch is er hier geen ‘natuurlijk nulpunt’ wat lengte betreft: het verschil tussen de heren wordt immers in relatie tot elkaar bekeken. Een wat algemener voorbeeld: $20 - 15 = 10 - 5$. Bij een meting met variabelen op intervalmeetniveau wordt dan ook wel gesproken van een meting met een ‘relatie van vaste afstand’.
- **Ratiomeetniveau** Zodra we een afstand tussen twee variabelen bekijken en we hebben wél een natuurlijk nulpunt, hebben we te maken met variabelen op ratiomeetniveau. Bijvoorbeeld: de afstand van Nijmegen naar Den Bosch is 2x zo groot als de afstand van Nijmegen naar Oss. Je gebruikt hier Nijmegen als natuurlijk nulpunt. Zouden we bijvoorbeeld weer zeggen dat Kees 160 cm lang is en dat Jan 180 cm lang is, en we nemen nu 0 cm als nulpunt, dan kunnen we zeggen dat Jan $9/8$ ste van de lengte van Kees heeft. Deze getallen geven dan ook de verhouding van waarden weer ten opzichte van een bepaald ingesteld nulpunt. We spreken bij variabelen van dit meetniveau dan ook over verhoudingsgetallen.

2.6 Handmatig of machinaal?

Tenslotte, als alle voorbereidingen zijn getroffen en de data klaar is om geïnterpreteerd te worden, dient de laatste keuze zich aan: wordt de data handmatig geanalyseerd of gebruik je een computerprogramma? Beide hebben hun voor- en nadelen.

Een groot voordeel van handmatig is dat je zelf precies weet wat je doet. Je bent zelf getuige van het hele proces van analyse. Bij computeranalyse geef je dit voor een groot deel uit handen. Je moet er dan maar op vertrouwen dat het gebruikte hulpprogramma daadwerkelijk goed in elkaar zit en daadwerkelijk doet wat het moet doen.

Aan de andere kant is werken met computerondersteuning natuurlijk vele malen sneller dan handmatig werken. In sommige gevallen wordt je zelfs bijna gedwongen om de data machinaal te analyseren, domweg omdat je teveel data hebt om handmatig te analyseren.

Welke van de twee mogelijkheden het beste is, dat is geheel afhankelijk van de situatie. Enkele variabelen hier zijn bijvoorbeeld de hoeveelheid data, de competenties van degene die de data moet analyseren, de hoeveelheid beschikbare tijd om de data te analyseren, de vorm waarin de data beschikbaar is (wel of niet in een computerbestand), de beschikbaarheid van een computertool, etcetera.

Mocht je besluiten om per computer je data te analyseren, dan zijn er twee tools die veel gebruikt worden: Microsoft Excel [1] en SPSS [2, 10, 7]. Vooral SPSS is erg populair voor professionele toepassingen en in het statistiekonderwijs. Als

je aan een andere faculteit dan de Beta-faculteit een cursus statistiek volgt, dan is de kans erg groot dat je uiteindelijk een keer met SPSS aan de gang gaat.

2.7 Tot slot

Hiermee zijn alle noodzakelijke voorbereidingen getroffen om op een goede manier de data te kunnen interpreteren middels statistische methoden. Die gaan we dan ook nader bekijken in de volgende hoofdstukken. Om niet al teveel uit te wijden zal de focus vooral liggen op het nominale en intervalmeetniveau. Voor statistische methoden met de overige meetniveaus wordt wederom verwezen naar aanvullende literatuur, zoals bijvoorbeeld het boekje van Rosenthal [6] voor het ordinale niveau en bijvoorbeeld het boek van McClave [12] voor specifieke ratiotechnieken.

Chapter 3

Analyse met één variabele

3.1 Inleiding

In dit hoofdstuk gaan we kijken naar enkele manieren om om te gaan met statistische analyse met één variabele. We zullen hierbij eerst kijken naar frequentieverdelingen. Meer uitleg hierover staat in paragraaf 2. Vervolgens kijken we in paragraaf 3 naar centrale tendentie, vormen om gemiddelden te berekenen. In paragraaf 4 tenslotte kijken we naar spreidingsmaten.

3.2 Frequentieverdeling

De makkelijkste manier van analyse is de frequentieverdeling. Voor een frequentieverdeling heb je één variabele nodig, waarvan je gaat onderzoeken hoe vaak een item in je datamatrix voorkomt die aan je criteria voldoet. Bijvoorbeeld: je hebt 200 mensen geïnterviewd en gevraagd of ze het met het kabinetsbeleid eens zijn. Het kan zijn dat het antwoord ‘ja’ 42x voorkomt, het antwoord ‘nee’ 100 keer, het antwoord ‘neutraal’ 38 keer en ‘weet niet / geen mening’ 20 keer. Als je uitgaat van een nominaal meetniveau, dan is de waarde van positieve reacties dus 42. Je hebt nu de *absolute waarde* te pakken.

Op zich zegt dit getal nog niet zoveel. Het getal 42 zegt meer als je naar de *relatieve frequentieverdeling* gaat kijken. Van de 200 respondenten hebben 20 aangegeven: ‘weet niet / geen mening’. Dan kun je zeggen dat deze mensen voor zogenaamde ‘missing values’ zorgen: waarden die niet onder de geldige antwoorden vallen. We vinden in dit geval slechts ‘ja’, ‘neutraal’ en ‘nee’ geldige antwoorden. In plaats van dat slechts 21 procent van de mensen dus de vraag met ‘ja’ heeft beantwoord, heeft dus 23,3 procent deze met ‘ja’ beantwoord, als je de missing values buiten beschouwing laat. Deze manier van representeren noemt men ook wel het presenteren van de *relatieve frequentieverdeling*.

Stel nu echter dat we niet slechts drie antwoordmogelijkheden hadden, maar vijf: ‘helemaal eens’, ‘gedeeltelijk eens’, ‘neutraal’, ‘gedeeltelijk oneens’ en ‘helemaal oneens’. We kunnen nu voor iedere antwoordcategorie afzonderlijk de frequen-

tieverdeling bekijken, maar we kunnen ook, mits dit is toegestaan in het conceptuele model van je onderzoek, de vier categoriën paarsgewijs samenvoegen tot twee categorieën: ‘ja’ en ‘nee’. Als je dit doet, en daarvan de (relatieve) frequentieverdeling weergeeft, dan spreken we van het weergeven van het *cumulatief resultaat*.

3.3 Centrale Tendentie

Een goede manier om inzicht te krijgen in hoe je onderzochte populatie in elkaar steekt is het kijken naar de centrale tendentie van je populatie. Ofwel: kijken naar gemiddelden. We onderscheiden hierbij drie soorten gemiddelden: voor ratio- en intervalvariabelen, voor ordinale variabelen en voor alle variabelen. Hieronder zal kort worden uitgelegd wat deze inhouden en hoe deze gemiddelden berekend kunnen worden [8, 9, 13].

3.3.1 Ratio- en Intervalvariabelen

Voor ratio- en intervalvariabelen wordt voor het berekenen van het gemiddelde het *rekenkundig gemiddelde* genomen. Dat wil zeggen: tel alle waarden bij elkaar op en deel de som door het aantal waarden.

Bijvoorbeeld: neem de getallenreeks

1,1,1,2,2,3,3,4,5,7,8,10,25.

Deze reeks bestaat uit 13 getallen. De som van deze getallen is 72. Het gemiddelde is dus $\frac{72}{13}$, oftewel afgerond 5,5.

3.3.2 Ordinale Variabelen

Het heeft uiteraard niet zoveel zin om voor een geordende reeks een rekenkundig gemiddelde te nemen, zoals bij de ratio- en intervalvariabele. Daarom gebruiken we voor ordinale variabelen de *mediaan*. Bij de mediaan wordt gekeken naar het middelste element van de geordende populatie. Neem bijvoorbeeld weer de getallenreeks:

1,1,1,2,2,3,3,4,5,7,8,10,25.

Deze reeks bestaat wederom uit 13 elementen en is al geordend (de getallen zijn oplopend geordend). Het middelste element van deze serie is dus het zevende element, in dit geval dus het getal 3. De mediaan van de geordende reeks is dus 3.

Het kan uiteraard ook voorkomen dat je met een reeks met een even aantal elementen te maken krijgt: stel bijvoorbeeld dat je een reeks van 100 elementen hebt. In dat geval ligt de mediaan dus precies tussen element 50 en 51 in. Volgens Ellis [8] is het dan gebruikelijk om het rekenkundig gemiddelde van die twee elementen (hier dus element 50 en 51) te nemen.

3.3.3 Alle Variabelen

Er bestaat echter ook een vorm van gemiddeldeberekening die op alle verschillende vormen van variabelen toepasbaar is. Deze manier heet het berekenen van de *modus*. Deze manier is dus ook geschikt voor het berekenen van het nominale gemiddelde: daarvoor hadden we tot nu toe nog geen manier besproken. De modus is simpelweg het element wat het vaakste voorkomt in een reeks. Stel bijvoorbeeld dat we weer de volgende getallenreeks hebben:

1,1,1,2,2,3,3,4,5,7,8,10,25.

Het element '1' komt drie keer voor, de overige elementen twee keer of minder. Dat betekent dat '1' hier het vaakst voorkomt en dus de modus vormt in deze getallenreeks.

Het kan voorkomen dat meerdere elementen even vaak en het vaakst voorkomen in een reeks. Bijvoorbeeld een groep mannen waarvan de namen Piet en Jan het vaakst voorkomen. In dat geval heb je dus te maken met twee modi. Dit wordt ook wel een bimodale verdeling genoemd. Mocht je zelfs meer dan twee modi tegenkomen, dan heb je te maken met een multimodale verdeling.

3.4 Spreidingsmaten

Uiteraard is het gemiddelde niet de enige manier om wat meer inzicht te krijgen in de populatie van een variabele. Sterker nog: een gemiddelde kan wat beperkend werken. Als je bijvoorbeeld uit een steekproef hebt gehaald dat de gemiddelde leeftijd van je respondenten 30 jaar was, dan kan het zijn dat je daadwerkelijk alleen mensen van 30 jaar hebt gesproken, maar ook dat je jongste respondent 4 jaar jong was en je oudste 110.

Je wilt dus bijvoorbeeld ook wat kunnen zeggen over de omvang van de verschillende elementen in je populatie, of over de grenzen waartussen het merendeel van de elementen uit de populatie zich bevindt. Met andere woorden: je wilt informatie over de spreiding van je populatie. Dat bereken je aan de hand van *spreidingsmaten*. Uiteraard zijn er ook manieren om dat te berekenen. We zullen hieronder ingaan op een drietal veelgebruikte manieren, te weten: de variatiebreedte, de standaarddeviatie en de kwartiilverdeling.

3.4.1 Variatiebreedte

De eenvoudigste spreidingsmaat is de variatiebreedte. Dit is simpelweg het verschil tussen de twee uitersten in je populatie. Deze spreidingsmaat is dus vooral geschikt voor interval- of ratiovariabelen, maar kan soms ook worden gebruikt voor ordinale variabelen.

Een voorbeeld. Stel, we nemen weer de getallenreeks

1,1,1,2,2,3,3,4,5,7,8,10,25.

De variatiebreedte is hier het verschil tussen het kleinste (1) en het grootste (25) element, dus $25 - 1$, oftewel 24.

3.4.2 Standaarddeviatie

Standaarddeviatie betekent standaardafwijking. Je berekent hier de gemiddelde afwijking van een element uit je populatie ten opzichte van het rekenkundig gemiddelde mee. Je kunt deze methode dus vooral toepassen op ratio- en intervalvariabelen. Is je standaardafwijking 0, dan zijn alle elementen uit je populatie hetzelfde. Hoe groter de standaarddeviatie, hoe groter de verschillen in de populatie zijn.

Standaarddeviatie bereken je als volgt:

1. bereken het rekenkundig gemiddelde van je populatie
2. bereken per element uit je populatie de afwijking ten opzichte van dat gemiddelde
3. kwadrateer die afwijkingen
4. tel de uitkomsten bij elkaar op
5. deel die som door het aantal onderzoekselementen
6. trek de wortel uit dat getal

Je hebt nu de standaarddeviatie berekend. Een voorbeeld van hoe dit wordt toegepast: neem wederom de getallenreeks:

1,1,1,2,2,3,3,4,5,7,8,10,25.

Het rekenkundig gemiddelde hebben we al eerder bepaald, dat was 5,5 (afgerond). Nu nemen we eerst per element de afwijking ten opzichte van dit gemiddelde. Dat levert de volgende getallen op:

4.5, 4.5, 4.5, 3.5, 3.5, 2.5, 2.5, 1.5, 0.5, 1.5, 2.5, 4.5, 19.5

Het kwadrateren van deze uitkomsten levert de volgende getallen op:

20.25, 20.25, 20.25, 12.25, 12.25, 6.25, 6.25, 2.25, 0.25, 2.25, 6.25, 20.25, 380.25

De som van deze kwadraten is 509.25. We hebben in totaal 13 elementen in de populatie, dus dienen we dit etal door 13 te delen. Dat levert 39,17 op. Tenslotte nemen we de wortel van dit getal. Dat levert op dat de standaarddeviatie 6.28 is.

Bij een populatie met een zogenaamde *normale verdeling*, een verdeling waar de meeste elementen van de populatie zo rond het gemiddelde liggen, en naarmate de afwijking van het gemiddelde toeneemt, het aantal elementen afneemt (ook wel bekend vanwege de klokvormige grafiek), geldt, dat normaliter 68 procent van alle elementen uit de populatie binnen de standaardafwijking ligt, 95 procent van alle elementen uit de populatie binnen tweemaal de standaarddeviatie ligt en tenminste 99 procent van alle elementen binnen driemaal de standaarddeviatie ligt [8, 9]

3.4.3 Kwartielverdeling

De kwartielverdeling is een manier om de populatie bij een variabele op te delen in kwarten. De kwartielverdeling kan worden gebruikt bij ordinale, interval en ratiovariabelen [8, 13]. Het idee is als volgt: allereerst bepaal je welk element het middelste element van je geordende populatie is, oftewel welke de mediaan is. Daarna kijk je naar het element waar 25 procent van alle elementen uit je populatie beneden valt (dus dat groter is dan tenminste 25 procent van je populatie). Dit noem je het eerste kwartiel. Daarna kijk je naar het element waar 75 procent van al je elementen beneden valt (dus dat groter is dan tenminste 75 procent van alle elementen).

Wederom een voorbeeld aan de hand van de getallenreeks:

1,1,1,2,2,3,3,4,5,7,8,10,25.

Deze populatie bevat 13 elementen. De mediaan hebben we al eens berekend, die is 3. Op precies het 3.25ste element bevindt zich het eerste kwartiel. Aangezien we alleen met hele elementen werken en minimaal 25 procent van de elementen onder de grens van het eerste kwartiel moet zitten, kiezen we hier het vierde element om het eerste kwartiel mee aan te geven. Dat is hier dus '2'. Hetzelfde idee wordt gehanteerd voor het berekenen van het derde kwartiel. Dat ligt op het 9.75ste element, oftewel op het tiende element. Dat is hier '7'. De kwartielverdeling is dus als volgt voor deze populatie:

- eerste kwartiel: 2
- mediaan (tweede kwartiel): 3
- derde kwartiel: 7

3.4.4 Vijf-Getallen-Resumé

Om een volledig beeld te geven van een bepaalde populatie wordt vaak gebruik gemaakt van vijf specifieke kengetallen, te weten het vijf-getallen-resumé [8]. Het idee is dat een dergelijk resume een vrij goed en volledig overzicht geeft van hoe de populatie die bij een variabele hoort eruit ziet. Dit resumé is vooral van toepassing op ratio- en intervalvariabelen, en bestaat uit de volgende vijf elementen:

1. het minimum van de populatie, dus het kleinste element uit de verzameling
2. het eerste kwartiel
3. de mediaan
4. het derde kwartiel
5. het maximum van de populatie, dus het grootste element uit de verzameling

Als we weer kijken naar de getallenreeks:

1,1,1,2,2,3,3,4,5,7,8,10,25.

dan is het vijf-getallen-resumé snel te geven. We hebben van deze reeks al zowel de kwartielen als de mediaan berekend, en het minimum en maximum zijn ook snel bekend: 1 en 25. Het vijf-getallen-resumé ziet er dan als geheelals volgt uit:

1. minimum: 1
2. eerste kwartiel: 2
3. de mediaan: 3
4. derde kwartiel: 7
5. maximum: 25

3.5 Tenslotte

We hebben in dit hoofdstuk een aantal veelgebruikte technieken voor de analyse van één variabele gezien. Deze analysetechnieken geven een goed inzicht in hoe de populaite van een variabele is opgebouwd. Hoe dan ook, het kan voorkomen dat je volgens het conceptuele model van je onderzoek niet genoeg hebt aan de analytse van één variabele, maar dat moet worden gekeken naar hoe bepaalde variabelen samenhangen. Dat wordt in de volgende hoofdstukken besproken.

Chapter 4

Analyse met twee variabelen

4.1 Inleiding

Het kan voorkomen dat je vanwege het conceptuele model van het onderzoek wat je doet, twee (of zelfs meer, zie daarvoor hoofdstuk vijf) variabelen met elkaar wilt vergelijken, om te zien of ze enige samenhang vertonen. Een mogelijke vergelijking is bijvoorbeeld geslacht tegenover het al dan niet hebben van een baan, als je in je conceptuele model de hypothese hebt opgenomen dat vrouwen minder vaak een baan hebben dan mannen.

Om dergelijke vergelijkingen tussen variabelen mogelijk te maken zijn een flink aantal technieken ontwikkeld. Om nog een beetje overzicht te bewaren (en omdat dit niet een heel overzichtsboek van statistische methoden hoeft te worden), worden slechts drie soorten relaties aan de orde gesteld: het vergelijken van twee nominale variabelen (paragraaf 2), het vergelijken van twee intervalvariabelen (paragraaf 3) en het vergelijken van een nominale en een intervalvariabele (paragraaf 4).

4.2 Twee Nominale Variabelen

Om twee nominale variabelen met elkaar te vergelijken worden *kruistabellen* gebruikt. Je gebruikt hierbij de kolommen om de populatie van de ene variabele uit te zetten, en de rijen voor die van de andere. Een voorbeeld staat op de volgende pagina.

Hier is wederom voor de verhouding man-vrouw en het hebben van een baan gekozen. De gepresenteerde cijfers zijn overigens geheel fictief en slechts ter illustratie.

De cijfers in de tabel geven op zich al een kleine indruk van hoe de verhoudingen liggen, maar een nog betere indruk zou het geheel uitgedrukt in percentages

Baan Geslacht	Man	Vrouw	Totaal
Ja	124	66	190
Nee	58	98	156
Totaal	182	164	346

geven. De techniek die daar bij hoort is het *percenteren*. Dat houdt in dat je de getallen in de tabel omzet in percentages.

Hierbij onderscheiden we drie soorten: horizontaal, verticaal en totaal. Welke je kiest hangt uiteraard af van je conceptuele model.

- Horizontaal percenteren houdt in dat je per kolom gaat kijken naar de percentages. In dit geval heb je drie kolommen: ja, nee en totaal. Je berekent hiermee dus: zoveel procent van de mensen met een baan is man, en zoveel procent is vrouw. En zoveel procent van de mensen zonder baan is man, en zoveel procent van de mensen zonder baan is vrouw. Een horizontaal gepercenteerde tabel ziet er als volgt uit:

Baan Geslacht	Man	Vrouw	Totaal
Ja	65,26%	34,74%	100%
Nee	37,17%	62,83%	100%
Totaal	52,60%	47,40%	100%

- Verticaal Percenteren werkt hetzelfde als horizontaal percenteren, maar nu bereken je de percentages over de kolommen. Hier bereken je dus: van de mannen heeft zoveel procent een baan en zoveel procent niet. En van de vrouwen ook, uiteraard. Dat levert dus de volgende tabel op in dit voorbeeld:

Baan Geslacht	Man	Vrouw	Totaal
Ja	68,13%	40,24%	54,91%
Nee	31,87%	59,76%	45,09%
Totaal	100%	100%	100%

- Totaal Percenteren heeft betrekking op de gehele populatie. Dus je percentert over de hele tabel. Hier bereken je dus: zoveel procent van de populatie is man en heeft een baan, is vrouw en heeft een baan, is man en heeft geen baan en is vrouw en heeft geen baan. Als ‘bijeffect’ bereken je meteen ook hoeveel procent van je totale populatie man is, hoeveel procent vrouw is, hoeveel procent een baan heeft en hoeveel procent niet. De totaal gepercenteerde tabel staat op de volgende pagina.

Baan Geslacht	Man	Vrouw	Totaal
Ja	35,84%	19,08%	54,91%
Nee	16,76%	28,32%	45,09%
Totaal	52,60%	47,40%	100%

Overigens is percenteren nog ergens goed voor. Het kan namelijk zijn dat je tweemaal een steekproef doet voor een bepaald onderzoek. Als de percentages vrij nauwkeurig overeenkomen (ongeveer 2 procent afwijking is normaal) van die twee steekproeven, dan kun je er van uitgaan dat de steekproeven *statistisch significant* zijn, oftewel representatief voor het gebied wat je hebt onderzocht.

4.3 Twee Intervalvariabelen

De samenhang tussen twee variabelen van intervalniveau wordt aangegeven aan de hand van een *correlatiecoëfficiënt*. Dat wil zeggen: een getal, met de waarde tussen -1.0 en 1.0, dat aangeeft in hoeverre verandering in de ene variabele invloed heeft op de andere variabele: op welke manier (positief of negatief) en in welke mate (sterk of zwak). De correlatiecoëfficiënt gaat uit van lineaire samenhang tussen variabelen. De correlatiecoëfficiënt kan als volgt worden uitgelegd:

- Negatieve samenhang (alle waarden onder 0 tot en met -1.0) houdt in dat als je ene variabele in waarde toeneemt, je andere variabele in die specifieke mate afneemt. Dus groeit bijvoorbeeld je afzet van oliebollen, dan is de kans groot dat de afzet van waterijsjes even af zal nemen, en andersom: er zijn tenslotte niet zoveel mensen die in de zomer oliebollen eten en in de winter waterijs.
- Positieve samenhang (alle waarden boven 0 tot en met 1.0) houdt in dat zodra een bepaalde variabele toe of afneemt, de andere variabele ook in een bepaalde mate afneemt (de sterkte van de toename is afhankelijk van de mate van samenhang). Dus stel bijvoorbeeld dat je de economie van een land ziet inzakken, dan is de kans groot dat het aantal geboekte vakanties door mensen in zo'n land óók af zal nemen. En zodra de economie weer aantrekt, zal het aantal geboekte vakanties ook toenemen.
- Geen samenhang (de waarde is precies 0). In het geval er géén samenhang is tussen twee variabelen, dan is de correlatiecoëfficiënt precies nul. Zo zal bijvoorbeeld het stijgen of dalen van de zeewaarttemperatuur in de Noordzee geen enkele invloed hebben op het aantal treinen van Syntus wat op tijd rijdt: het is absoluut niet zo dat als de zeewatertemperatuur 10% stijgt, er opeens tot 10% meer of minder Syntustreinen op tijd rijden.

De resultaten van deze vergelijking geef je weer in een *correlatiematrix*. Je zet hierbij de relevante intervalvariabelen in je conceptueel model tegenover elkaar uit, waarna je de verbanden (of het ontbreken ervan) paarsgewijs kunt aangeven.

Een voorbeeld van een correlatiematrix, waarin variabele 1 de economische groei, 2 het aantal geboekte vakanties en 3 het aantal werklozen vertegenwoordigt:

	1.	2.	3.
1. Economische Groei	1.0		
2. Geboekte Vakanties	0.5	1.0	
3. Aantal Werklozen	-0.6	-0.3	1.0

Ook hier geldt: de waarden zijn geheel fictief en slechts ter illustratie. Je ziet wel in de tabel dat een variabele met zichzelf altijd een 1.0 relatie heeft (dat is vrij logisch). Het is ook vrij logisch dat het aantal werklozen afneemt als de economie groeit, en dat het aantal werklozen toeneemt als de economie krimpt. Die twee variabelen vertonen dus een negatieve samenhang. Het is ook vrij duidelijk dat als de economie groeit, het aantal geboekte vakanties toeneemt: daar is dan immers meestal weer financiële ruimte voor. Tenslotte is de relatie tussen het aantal geboekte vakanties en het aantal werklozen ook negatief: als er meer werklozen zijn, zijn er minder huishoudens die financiële ruimte zullen hebben om uitgebreid of meerdere keren per jaar op vakantie te gaan dan wanneer er minder werklozen zijn.

4.4 Eén Nominale en één Intervalvariabele

Het berekenen van de relatie tussen een intervalvariabele en een nominale variabele geschiedt aan de hand van de zogenoemde *ANOVA* (*ANalysis Of VAriance*)-*technieken*. Stel bijvoorbeeld dat je onderzoekt of er verschil is tussen mannen en vrouwen wat betreft de tijd die nodig is om een puzzel op te lossen. Je hebt hier een nominale variabele (man versus vrouw) en een intervalvariabele (de tijd om de puzzel op te lossen).

Helaas zijn de meeste in deze paragraaf genoemde technieken niet uit te leggen zonder minstens enige flinke berekeningen aan te wenden of de achterliggende conceptuele modellen aan te halen. Dat gaat voor deze reader te ver, vandaar dat voor de specifieke *hoe* van de technieken wordt verwezen naar de statistiekboeken en -cursussen[12, 8, 9, 14, 15]

4.4.1 De T-Toets

De eenvoudigste variant van de ANOVA-technieken is de *t-toets*. Om echt rendement te hebben van je t-toets is het noodzakelijk om meerdere steekproeven te nemen. De t-toets is namelijk gebaseerd op het idee dat vooral gemiddelden met elkaar vergeleken kunnen worden, en om meerdere gemiddelden te hebben, zijn meerdere steekproeven nodig. Het idee is vooral dat je aan de hand van een t-toets kunt zien of de gemiddelden al dan niet normaal verdeeld zijn. In dit geval heet dat echter t-verdeeld (of Student-verdeeld)[12, 9]. Helaas is de t-toets alleen bruikbaar in het geval dat in de nominale variabele slechts twee

categorieën onderscheiden kunnen worden. Voor de meer complex opgebouwde variabelen zijn andere ANOVA-technieken beschikbaar.

4.4.2 Eenweg Variantie-analyse

Eén van die technieken is de eenweg variantie-analyse, ofde *one way ANOVA*. Bij de one way ANOVA wordt er, net als bij de t-toets, vanuit gegaan dat er meerdere steekproeven worden genomen en worden vergeleken met elkaar, om op die manier samenhang tussen variabelen aan te tonen (of juist van de hand te wijzen). De one way ANOVA is dus een generalisatie van de t-toets. Stel bijvoorbeeld dat we niet twee categorieën hadden bij het puzzelvoorbeeld, maar dat we willen kijken of Friezen, Limburgers of Utrechters beter zijn in puzzelen. Dan hebben we dus met meer categorieën en zijn we dus aangewezen op one way ANOVA.

4.5 Tenslotte

We hebben in dit hoofdstuk een aantal manieren gezien om twee (soorten) variabelen met elkaar te vergelijken, om zo de samenhang (of het ontbreken ervan) aan te tonen. Er zijn nog véél meer verschillende technieken om met variabelen om te gaan, zoals bijvoorbeeld de chi-kwadraat-toets, de F-toets, de Wilcoxon, Kendalls tau, etcetera. het gaat te ver om die hier allemaal zelfs maar te noemen. Wel is er nog één hoofdcategorie over: analyse van samenhang tussen meer dan twee variabelen. Dat is, op hoofdlijnen, dan ook onderwerp van het volgende hoofdstuk.

Chapter 5

Analyse met meerdere variabelen

5.1 Inleiding

In dit hoofdstuk kijken we tenslotte naar een overzicht van verschillende technieken om samenhang tussen meer dan twee variabelen aan te tonen. Voor dat dat aan de orde komt kijken we eerst naar een manier om te testen of er überhaupt wel samenhang is tussen die hoeveelheid variabelen, en naar mogelijke redenen waarom het toch kan lijken alsof er samenhang is, of, als er toch samenhang is, hoe die samenhang dan precies in elkaar steekt. Dit heet de *elaboratietechniek*, en zal verder worden behandeld in paragraaf 2. Het overzicht van de verschillende technieken staat verder in paragraaf drie. Tenslotte volgt nog een terugblik op dit hoofdstuk in de vierde paragraaf.

5.2 Elaboratietechniek

Waarom zou je eigenlijk met meer dan twee variabelen willen werken bij het analyseren van je data? Daar zijn twee mogelijke redenen voor: ten eerste kan het zijn dat het conceptueel model van je onderzoek hierom vraagt, en ten tweede kan het zijn dat je de samenhang tussen twee variabelen wilt toetsen door te kijken wat de invloed van een derde variabele is. Het kan namelijk gebeuren dat er een causaal verband lijkt te bestaan tussen twee variabelen, maar dat dit verband eigenlijk alleen bestaat bij de gratie van een derde variabele die zijn invloed op die relatie uitoefent.

Er is een manier om te testen of dit het geval is. Die toets heet de *elaboratietechniek* [13]. Elaboratietechniek maakt gebruik van multidimensionale kruistabellen voor het in kaart brengen van de variabelen. De manier om te testen of er überhaupt sprake is van een causale relatie tussen twee variabelen is hier door de derde variabele constant te maken. Hiervoor dien je een steekproef te nemen

waarbij je die derde factor constant houdt. Stel bijvoorbeeld dat je denkt dat er een causaal verband bestaat tussen iemands leeftijd en diens oordeel over niet-Westerse culturen. Het kan zijn dat er een causaal verband lijkt te bestaan (bijvoorbeeld: naarmate mensen ouder worden, worden ze intoleranter). Het kan echter ook zijn dat, omdat er relatief meer jonge mannen zijn dan jonge vrouwen en relatief meer oude vrouwen dan jonge mannen, dit verschil wordt veroorzaakt door de onderzochte populatie, mocht de populatie een afspiegeling van de bevolking zijn. In dat geval zou geslacht dus een derde variabele zijn die onbewust invloed heeft op het verband. De manier om dit te testen is door een onderzoek uit te voeren alleen onder mannen en een alleen onder vrouwen. Op die manier heb je de derde variabele, geslacht, constant gemaakt. Het kan ook zijn dat de bevolking op verschillende leeftijden anders is samengesteld. Bijvoorbeeld omdat er meer jongere Nederlanders van niet-Westerse afkomst zijn dan oudere. In dat geval kun je onderzoeken of er daadwerkelijk wel een causaal verband is tussen de genoemde variabelen. Tenslotte kan ook de leefomgeving waarin iemand verkeert, dus bijvoorbeeld of iemand in een Randstedelijke stad woont of in een Drents dorp, van invloed zijn op diens beeld in deze kwestie. Hieronder kijken we naar de vier mogelijke uitkomsten die de elaboratietechniek hierbij kan geven.

5.2.1 Uitkomsten bij toepassen Elaboratietechniek

Het gebruik van de elaboratietechniek kan leiden tot vier mogelijke soorten uitkomsten. Hieronder zullen we ze kort bespreken.

1. Interne Replicatie

Interne Replicatie houdt in dat, nadat de derde variabele constant is gehouden, de causale relatie tussen de overige twee variabelen nog steeds aanwezig blijkt. Stel dus dat bij zowel mannen als vrouwen ongeveer dezelfde uitkomsten naar boven komen als in het onderzoek waarin geslacht niet specifiek als derde variabele constant is gehouden, dan is er sprake van interne replicatie. In dat geval heeft de derde variabele dus geen enkele invloed op het resultaat, en kan worden geconcludeerd dat er daadwerkelijk een causaal verband is tussen die twee variabelen.

2. Interactie

Interactie houdt in dat er, slechts afhankelijk van de waarde van de derde variabele, een mogelijke causale relatie tussen twee variabelen bestaat. In dat geval is er bij het wijzigen van de waarde van die derde variabele totaal geen samenhang meer te vinden, en blijkt dus dat die derde variabele zelf een onderdeel is van de relatie. Stel bijvoorbeeld dat er voor mannen en voor vrouwen totaal andere uitsagen naar boven komen, dan is er sprake van interactie.

3. Schijnsamenhang

Het kan ook voorkomen dat je nog andere vormen van toevalligheden wilt uitsluiten. In dit geval heb je het vermoedelijke causale verband: hoe

jonger hoe intoleranter. Maar stel nou dat we leeftijd even als constante nemen en bijvoorbeeld bedenken dat jonge mensen vaak zelf uit een andere cultuur afkomstig zijn (bijvoorbeeld omdat mensen van niet-Westerse afkomst er vaker grote gezinnen op nahouden). De kans is zeker groot dat die mensen al bijvoorbaat positief tegenover niet-westerse culturen staan. Dat kan het beeld ook flink vertekenen. Stel dat uit het onderzoek naar voren komt dat jongeren inderdaad toleranter zijn dan ouderen, maar dat er ook veel meer niet-Westerse jongeren in de onderzochte populatie zaten dan niet-Westerse ouderen. In dat geval is er dus eigenlijk helemaal geen sprake van een causaal verband tussen leeftijd en tolerantie: we spreken hier dan van een schijnsamenhang.

4. Interpretatie

Uiteraard speelt het ook mee in hoeverre iemand zelf te maken heeft met Nederlanders van niet-Westerse afkomst. Zo zijn er legio dorpen te vinden waar geen enkele niet-Westerse Nederlander woont, terwijl steden als Rotterdam en Den Haag al bijna meer niet-Westerse Nederlanders herbergen dan autochtone Nederlanders. Door deze factor, leefomgeving, mee te nemen in het model kan naar voren komen dat bijvoorbeeld jongeren in de steden relatief zeer tolerant zijn, en de ouderen al iets minder tolerant, maar nog steeds tolerant, en dat de jongeren in dorpen een stuk minder tolerant zijn dan de jongeren in de steden, maar nog altijd toleranter dan de ouderen in de dorpen. In dat geval is wel een deel van de causaliteit van het verband onderuit gehaald, maar de strekking van het verhaal blijft overeind staan. In feite komt het er op neer dat er geen causaal verband bestaat tussen twee, maar tussen drie variabelen. In dat geval spreekt men van interpretatie.

5.3 Multivariabele Analysetechnieken: een Overzicht

Tenslotte zullen we kijken naar een overzicht van enkele technieken die voor multivariabele statistische analyse. Uiteraard zijn er nog veel meer, maar een poging om alle technieken op te noemen en kort te bespreken zou al bijvoorbaat falen. Bovendien ligt ook hier de focus op twee soorten variabelen: intervalvariabelen en nominale variabelen. Merk overigens op dat een groot aantal technieken in de basis hetzelfde zijn; het is vaak vooral het achterliggende conceptuele model van die techniek dat verschilt van de andere technieken [3].

5.3.1 Loglineaire Analyse

Loglineaire analyse is een statistische methode die kan worden gebruikt bij het analyseren van één afhankelijke variabele van nominaal niveau, en minimaal één onafhankelijke variabele van nominaal niveau. Loglineaire analyse kan dus vooral gebruikt worden voor het analyseren van multidimensionale kruistabellen. Het conceptuele model achter deze techniek is gebaseerd op het idee dat

de logaritme van de afhankelijke variabele, in het geval de overige variabelen inderdaad geheel onafhankelijk zijn (en dus geen samenhang vertonen), de relatie tussen de afhankelijke en de onafhankelijke variabelen lineair is.

5.3.2 Discriminantanalyse

Discriminantanalyse gaat uit van één onafhankelijke variabele van nominaal niveau en één of meer onafhankelijke variabelen van intervalniveau. Het idee achter discriminantanalyse is vooral dat het de onderscheidende variabelen tussen twee of meer groepen in een populatie onderscheidt. Als je bijvoorbeeld naar lengte kijkt van mensen, zal blijken dat mannen gemiddeld langer zijn dan vrouwen. Hier is geslacht dus een onderscheidende variabele.

5.3.3 Logistische regressie

Net als bij discriminantanalyse wordt voor logistische regressie gebruik gemaakt van één onafhankelijke variabele van nominaal niveau en één of meer onafhankelijke variabelen van intervalniveau. Het doel van Logistische regressie is echter om op basis van data voorspellingen te kunnen doen over hoe vaak een verwachte uitkomst zal voorkomen.

5.3.4 Variantie-analyse

Variantie-analyse is vooral een verzamelbak van diverse technieken die gebruikt kunnen worden voor de analyse van één onafhankelijke variabele van intervalniveau en één of meer onafhankelijke variabelen van nominaal niveau. Deze verzameling staat ook wel bekend als ANOVA-technieken. Eén ervan hebben we al gezien in het vorige hoofdstuk.

5.3.5 Multipele regressie

Multipele regressie is, net als Variantie-analyse, een ANOVA-techniek. Bij multipele regressie wordt uitgegaan van één onafhankelijke variabele van intervalniveau en één of meer onafhankelijke variabelen van intervalniveau. Multipele regressie onderscheidt zich in het feit dat bij multipele regressie wél enige vorm van onderlinge samenhang mag voorkomen tussen de andere variabelen.

5.3.6 MANOVA-technieken

Indien je te maken hebt met méér dan één afhankelijke variabele van nominaal niveau, dan zijn MANOVA-technieken de meest waarschijnlijke technieken om te gebruiken. MANOVA staat voor Multiple ANalysis of Variance.

5.3.7 Pad-Analyse

Padanalyse gaat uit van samenhang tussen twee of meer afhankelijke intervalvariabelen, waarbij vooral wordt gekeken naar de paden waarop die variabelen samenhangen. Bijvoorbeeld: neem een docent, en kijk naar in hoeverre er verbanden bestaan tussen de manier waarop deze doceert, zijn leefsituatie, zijn hobby's, zijn slaapgewoonten, zijn eetgewoonten, zijn geloofsovertuiging, etcetera. Tussen al deze variabelen zijn verbanden aan te wijzen en ook is aan te wijzen in hoeverre en hoe sterk die verbanden zijn, oftewel: hoe sterk de relatie tussen twee van die variabelen is.

5.3.8 Combinatietechnieken

In de loop van de tijd zijn er ook technieken ontwikkeld die verschillende technieken combineren. Eén ervan is bijvoorbeeld LISREL. Deze techniek combineert factoranalyse en regressieanalyse en wordt vooral gebruikt voor bijvoorbeeld het aangeven van verbanden tussen de afhankelijke variabelen en latente variabelen (variabelen die er wel zijn maar niet zichtbaar zijn gemaakt), of het onderzoeken van meetfouten en voorspellingsfouten.

5.4 Dimensietechnieken

Naast technieken waarmee de variabelen aan sich te analyseren zijn, zijn er ook technieken om de veelheid van variabelen te reduceren (oftewel: op hoofdlijnen terug te brengen). Stel bijvoorbeeld dat je enkele honderden variabelen in je onderzoek hebt. Dan is het niet erg prettig om steeds alle variabelen mee te moeten nemen, voor zover het al haalbaar is om nog te rekenen met een veelheid van variabelen. Dimensietechnieken maken het mogelijk om op een abstracter niveau naar die variabelen te kijken, door bijvoorbeeld enkele variabelen te combineren tot een algemenere variabele. Voor verschillende meetniveaus zijn technieken beschikbaar:

- clustertechnieken voor het nominaal niveau
- schaaltechnieken voor het ordinaal niveau
- factoranalyse voor het intervalniveau

5.5 Tenslotte

In dit hoofdstuk hebben we gekeken naar hoe uit te zoeken is of er daadwerkelijk sprake is van een relatie tussen twee of meer variabelen, middels elaboratietechniek. Ook hebben we gekeken naar een overzicht van diverse manieren om de samenhang van meer dan twee variabelen te kunnen analyseren, alsook naar manieren om de hoeveelheid gebruikte variabelen te reduceren.

Chapter 6

Cases

6.1 Inleiding

Allereerst een kleine waarschuwing:

‘There are three kinds of lies: lies, damned lies, and statistics’ –
Benjamin Disraeli

Benjamin Disraeli is een voormalige negentiende-eeuwse Britse premier. Hij had, getuige ook deze uitspraak, niet al teveel op met hoe er met statistieken werd omgegaan. Laten we hierbij twee zaken onderscheiden: de manier waarop statistieken tot stand komen en de manier waarop met de totstandgekomen statistieken wordt omgegaan. Indien zorgvuldig de juiste soort variabele wordt onderkend en de juiste analysetechnieken juist worden toegepast, zullen de statistieken niet liegen: ze zijn dan volgens de geldende regels geformuleerd. Het is echter vaak de manier waarop met deze statistieken wordt omgegaan die de waarde van statistieken in diskrediet brengt. Het komt vaak genoeg voor dat iemand ergens een setje statistieken vandaan tovert om zijn mening kracht bij te zetten, zonder dat de statistieken in hun context worden gelaten of geheel van toepassing zijn op die specifieke situatie. Let daar dus ook op bij het interpreteren en gebruiken van statistieken.

In dit hoofdstuk worden een aantal van de in deze reader gepresenteerde technieken uitgelegd aan de hand van een voorbeeld uit de ICT-wereld. We kijken hiervoor naar een voorbeeld met de webstats4u-teller[5]. Dit is de opvolger van de vroeger zeer bekende en veelvuldig gebruikte NedStat-teller (die overigens nog wel bestaat, maar nu een commercieel product is geworden waar flink voor moet worden betaald in plaats van het freeware-product wat het was. Bovendien richt NedStat zich tegenwoordig meer op professionele toepassingen van websurf-statistieken). Deze case is op een zeer concreet toepassingsniveau beschreven: hierin worden veel concrete voorbeelden gegeven van hoe je statistiek toegepast ziet in het ICT-domein. Echter geldt hierbij de beperking dat niet alle gegeven voorbeelden tot in detail uitgewerkt zullen worden. De focus van deze reader ligt immers op het kunnen herkennen van de verschillende soorten variabelen en

het kunnen vinden en benoemen van de geschikte statistische analysetechniek voor ieder specifiek geval. Daar ligt dan dus ook de focus van de hier genoemde voorbeelden.

Tevens zullen nog kort twee toepassingen van Padanalyse in de ICT-sector kort bediscussieerd worden: neurale netwerken en agent-based information retrieval. Aangegeven zal worden hoe deze statistische techniek bijdraagt aan de ontwikkeling van kunstmatig intelligente systemen.

6.2 Casus 1: de Webstats4U-teller

Veel websites hebben er een: een teller die netjes bijhoudt hoeveel bezoekers er op een site komen, waar die bezoekers vandaan komen, via welke links de bezoekers op de website zijn gekomen, etcetera. In deze casus kijken we naar een aantal van de gegevens die via deze site beschikbaar zijn. We zullen ons daarna specifiek richten op de cijfers van één specifieke website, te weten startpagina.nl. Tenslotte zullen we nog kijken naar een mogelijke fout in de berekening van een van de statistieken van Webstats4U. De data die is gebruikt voor deze casus is terug te vinden in appendix A.

6.2.1 De Teller in het Algemeen

De hoofdpagina van de webstats4U-site geeft een overzicht van drie soorten statistieken: de pagina die de dag ervoor het meest bezocht is, de top tien van relatief snelste stijgers qua bezoekersaantallen en de top tien van nieuwkomers. De eerste en derde categorie zijn hier ordinale variabelen: de getallen worden gebruikt om een reeks te ordenen van het aantal bezoekers van websites per dag. De tweede categorie beschrijft een ratiovariabele: vanuit een bepaald nulpunt wordt gemeten in hoeverre er een bepaalde toe- of afname van het bezoek van die site is ten opzichte van de dag ervoor.

6.2.2 Een Specifieke Teller

Laten we nu eens kijken naar hoe de index van één specifieke teller er uit ziet. We nemen hiervoor de statistieken van de vaakst bekeken website: startpagina.nl. Allereerst zijn er wat algemene karakteristieken van de website:

- de datum waarop de teller is aangemaakt,
- het totale aantal bezoeken tot nu toe (gewoon de som van alle bezoeken),
- de drukste dag tot nu toe en het aantal bezoekers die dag,
- het aantal bezoekers van de huidige dag tot op het moment van opvragen,
- en tenslotte het aantal bezoekers van de dag ervoor.

Aangezien aantal bezoekers kan worden gedefinieerd als een intervalvariabele, kan, op basis van het aantal dagen dat de site bestaat (het aantal items in de populatie), worden berekend wat het rekenkundig gemiddelde aantal bezoekers is voor die site per dag. De teller is op 30 september 1998 aangemaakt, zodat we, tot op het moment van schrijven, precies 2638 dagen verder zijn. startpagina.nl heeft sinds die tijd 4.318.907.754 pageviews gehad, oftewel gemiddeld 1.637.190 pageviews per dag.

Een Eerste Analyse: het Rekenkundig Gemiddelde

Als we dit vergelijken met een steekproef uit die populatie, te weten de bezoekers van de afgelopen 4 weken, dan vinden we een gemiddelde van 3.586.534 bezoekers per dag. We kunnen in ieder geval constateren dat het de laatste tijd erg druk was in vergelijking met het gemiddelde. Stel dat we de hypothese hadden dat er de laatste maand veel meer internetters uit Nederland actief zijn dan gemiddeld in de afgelopen zeven jaar, en we willen dat meten aan de hand van de beschikbare data op de site die al jarenlang één van de meest bezochte sites van Nederland is. We kunnen dan allereerst constateren dat verreweg de meeste bezoekers van die site uit Nederland afkomstig zijn, en dat het gemiddelde bezoekersaantal in de afgelopen vier weken gemiddeld 2,19 keer zo hoog lag als gedurende het gemiddelde over de afgelopen zeven jaar. Die hypothese lijkt dus enige grond van waarheid te hebben.

Een Tweede Analyse: de T-Toets

Stel nu dat we het idee hebben dat er de afgelopen vier weken in het weekend significant minder mensen zijn die startpagina.nl bezoeken dan op werkdagen in de week. We hebben dan te maken met een nominale variabele met niet meer dan twee classificaties (weekend en werkdag), en met een intervalvariabele (het aantal bezoekers per dag). In dit geval zou je ervoor kunnen kiezen om het geheel aan de hand van een t-toets te analyseren. Laten we nu, aangezien we de gegevens van de afgelopen vier weken paraat hebben, het geheel eens in kaart brengen. In de onderstaande tabel staan allereerst de soort dagen, het gemiddelde aantal bezoekers per dag en de absolute afwijking hiervan van het rekenkundig totale gemiddelde vermeld, alsook de absolute frequentie (dus hoeveel elementen voor deze steekproef verantwoordelijk zijn):

Dag	Gem. aantal bezoekers	Abs. Afw.	Abs. Freq
Werkdag	3.770.703	184.169	20
Weekend	3.126.111	-460.432	8
Totaal	3.586.534	0	28

De hypothese lijkt te kloppen: in het weekend bezoeken veel minder mensen startpagina.nl. Voor een precieze uitwerking van hoe deze met de t-toets te berekenen valt wordt verwezen naar de statistiekboeken. de inhoudelijke behandeling van de methoden valt buiten de scope van deze reader: het aangeven

Dag	Gem. aantal bezoekers	Absolute afw.
Maandag	3.944.540	358.006
Dinsdag	3.816.991	230.457
Woensdag	3.815.096	228.562
Donderdag	3.694.315	107.781
Vrijdag	3.582.575	-3.959
Zaterdag	2.948.128	-638.406
Zondag	3.304.093	-282.441
Totaal	3.586.534	0

van hoe je (soorten) variabelen kunt herkennen en alvast voorbereidend werk kunt treffen voor het berekenen van de precieze uitkomsten valt wél hierbinnen.

Een Derde Analyse: One-Way ANOVA

Stel nu dat we het probleem uit de vorige analyse iets dieper willen uitwerken, en niet specifiek iets willen weten over de bezoekersaantallen in het weekend of op een werkdag, maar over een specifieke dag, bijvoorbeeld zaterdag. Stel dat de hypothese is dat op zaterdag de minste bezoekers naar startpagina.nl komen. In dat geval heb je méér dan twee nominale variabelen (weekend of werkdag), namelijk zeven stuks (de dagen van de week). Hier kan een analyse aan de hand van de t-toets dus niet meer worden uitgevoerd, maar kan one-way ANOVA uitkomst bieden. Ook hier is vast weer wat voorbereidend werk getroffen. In de tabel bovenaan de volgende pagina staan alvast de nominale variabelen (de dagen van de week) uitgestald, met daarachter de gemiddelde bezoekersaantallen per dag, gedurende vier weken. Ook hiër geldt weer: het concrete uitwerken aan de hand van de specifieke techniek valt buiten de scope van dit paper.

Een vijf-getallen-resumé

We gaan nu een vijf-getallen-resumé opstellen voor het volgende geval: aan de hand van de gegevens van twaalf steekproeven (zie de appendix voor de gegevens), kijken we naar hoeveel procent van de bezoekers van startpagina.nl gemiddeld uit Nederland komt op een bepaald moment. Het gemiddelde wat op Webstats4U staat is weliswaar berekend over de afgelopen zeven jaar, maar we kijken nu ten eerste naar in hoeverre dit gemiddelde representatief is voor een willekeurig moment en ten tweede hoe het de afgelopen paar dagen gemiddeld zat. De tijdstippen van de metingen van de steekproeven zijn geheel willekeurig gekozen.

Allereerst berekenen we nu de relatieve frequenties uitgedruk in percentages Nederlanders voor de twaalf steekproeven. Mocht de waarde ‘Onbekend’ voorkomen, dan wordt die uiteraard gezien als een missing value en uit de populatie geschrapt. Deze frequenties zijn:

- proef 1: 98 %
- proef 2: 94 %
- proef 3: 98 %
- proef 4: 96 %
- proef 5: 94 %
- proef 6: 98 %
- proef 7: 98 %
- proef 8: 100 %
- proef 9: 92 %
- proef 10: 94 %
- proef 11: 100%
- proef 12: 96 %

Als we deze percentages ordenen dan krijgen we de volgende rij:

92%, 94%, 94%, 94%, 96%, 96%, 98%, 98%, 98%, 98%, 100%, 100%

Het minimum en het maximum zijn inmiddels bekend: 92% en 100%.

De mediaan ligt precies tussen het zesde en zevende element in. Hier is dat dus precies tussen 96% en 98%, dus 97 %. Het eerste kwartiel ligt na precies een kwart van de elementen, dus in dit geval na het derde van de twaalf elementen, oftewel bij het vierde element. Dat is 94%. Het derde kwartiel ligt na het negende, dus op het tiende element. Dat is hier 98%. We hebben nu voor deze steekproeven het volgende vijf-getallen-resumé op kunnen stellen:

- minimum: 92%
- eerste kwartiel: 94%
- mediaan: 97%
- derde kwartiel: 98%
- maximum: 100%

We kunnen overigens ook het rekenkundig gemiddelde berekenen van deze percentages: dat is 96.5% Oftewel: gemiddeld komt 96.5% van alle bezoekers van startpagina.nl uit Nederland volgens deze steekproeven. Of 97% als je de mediaan hanteert.

6.2.3 Een Foutje of een Andere Interpretatie?

De teller van Webstats4U lijkt echter ook op één specifiek punt niet helemaal de juiste gegevens weer te geven. Het gaat hier specifiek over de gegevens die in de rechter kolom van tabel A.2 van appendix A staan. Hier staat namelijk een relatieve frequentieverdeling van de landen van herkomst. Nou is het vrij gebruikelijk, zie hiervoor ook sectie 3.1, dat je zogenaamde ‘missing values’, oftewel antwoorden die je niet kunt gebruiken in je datamatrix, uit de uitslag schrapt. In dit geval is de categorie ‘Onbekend’ een typisch voorbeeld van ‘missing values’. Er is immers geen enkel land of domein op de wereld dat ‘Onbekend’ heet. Dergelijke resultaten opnemen in de relatieve frequentieverdeling (de laatste kolom van die tabel) verstoort slechts de werkelijke relatieve verdeling. De kans is dan ook erg groot dat, als het wél mogelijk was om ‘Onbekend’ nader te analyseren, hieruit zou blijken dat dat ongeveer een afspiegeling zou zijn van de relatieve frequentieverdeling zoals die zou zijn zonder dat je ‘onbekend’ zou opnemen hierin.

In dit geval is ‘Onbekend’ dus **niet** uit de relatieve frequentieverdeling geschrapt. Hierdoor geeft de relatieve frequentieverdeling een vertekend beeld. De tabel met daarin beschreven hoe de werkelijke relatieve frequentieverdeling er uit zou kunnen zien als Onbekend daadwerkelijk als Missing Value werd behandeld staat op de volgende pagina.

Uiteraard is er al een mail verstuurd naar Webstats4U, met het verzoek om uitleg over de motivatie om dit niet te doen, maar tot op heden is er nog geen reactie op gekomen.

6.3 Semi-Casus 2: Padanalyse en Kunstmatige Intelligentie

Naast tellers van webstatistieken zijn er natuurlijk nog veel meer toepassingen van statistiek in het ICT-domein. Een voorbeeld van een techniek die relatief

nr.	Land van Herkomst	Aantal	percentage
1.	Nederland	4090527263	97.3 %
2.	België	20470670	0.5 %
3.	Verenigde Staten	16811548	0.4 %
4.	Duitsland	11258392	0.3 %
5.	Netwerk	9129221	0.2 %
6.	Verenigd Koninkrijk	8600845	0.2 %
7.	VS Commercieel	8297498	0.2 %
8.	Zweden	6623837	0.2 %
9.	Frankrijk	5416227	0.1 %
10.	Spanje	3425709	0.1 %
	De rest	21826265	0.5 %
	Onbekend	116520279	%
	Totaal	4318907754	100.0 %

Table 6.1: Gegevens land van herkomst bezoekers (aangepast)

vaak wordt gebruikt in toepassingen van Kunstmatige Intelligentie is padanalyse. Hier worden twee concrete aspecten uitgelicht en wordt beschreven hoe padanalyse als statistische techniek bijdraagt aan die aspecten.

6.3.1 Neurale Netwerken

De eerste toepassing is die in neurale netwerken [11]. Een gecomputeriseerd neuraal netwerk is in eerste instantie een nabootsing van het menselijk associatief geheugen. Bijvoorbeeld: zeg Den Haag, en veel mensen denken meteen aan het Binnenhof, Scheveningen, de Lange Poten, en 't Plein. Die laatste twee termen zullen waarschijnlijk, mits in combinatie genoemd, een associatie met Harrie Jekkers oproepen. Op deze manier valt een heel conceptueel netwerk aan te leggen van kennis in de vorm van associaties. Een volgend aspect hierbij is het toekennen van waarden aan de associaties. Dit gebeurt ook weer op basis van statistieken (meestal relatieve frequenties): dus als bijvoorbeeld bij het woord Ajax in 60 van de 100 gevallen de voetbalclub wordt bedoeld, dan zal de waarde die aan die verbinding wordt toegekend hoger zijn dan de associatie met bijvoorbeeld het schoonmaakmiddel of de Griekse held. Uiteraard veranderen deze waarde naarmate er meer statistische waarden bekend raken over diverse associaties.

Tot op de dag van vandaag worden neurale netwerken gebruikt om te proberen delen van het menselijke brein juist te simuleren. Vooralsnog functioneren alle neurale netwerken vooral dankzij padanalyse.

Padanalyse is dus uiterst geschikt voor deze vorm van statistische analyse. Bij padanalyse draait het juist ook om de verschillende paden die gekozen kunnen worden vanuit een bepaald uitgangspunt. Ook hier gebeurt dat vaak op basis van waarden die aan de mogelijke verbindingen zijn toegekend.

6.3.2 Agent-based Information Retrieval

Een mogelijke vorm van information retrieval is agent-based information retrieval. In dat geval wordt het internetgedrag van een individu vastgelegd in de vorm van statistieken. Surft een individu bijvoorbeeld vaak naar nieuwssites, en dan vaak naar het sportgedeelte, en dan vaak naar autosport, dan zal die persoon dus wel een flinke interesse hebben in autosport. Surft die persoon daarna bijvoorbeeld ook vaak naar sites over fotografie, dan kan een webagent bijvoorbeeld tot de conclusie komen dat er een sterke link bestaat tussen autosport en fotografie, en voor zo'n persoon bijvoorbeeld bij het zoeken naar informatie over een van die twee aspecten sites of nieuwsgroepen aanbieden die over allebei gaan, dus bijvoorbeeld over autosport fotograferen. Enige vraag is nog in hoeverre een dergelijke manier van informatieanalyse als legaal kan worden bestempeld: misbruik van dergelijke gegevens is erg makkelijk en aantrekkelijk. Ook de voor dit soort doeleinden benodigde statistieken worden echter vooral via padanalyse verkregen: een dergelijke agent volgt gewoon de stappen van punt A naar punt B naar punt C enzovoort. Indien bepaalde paden of punten vaker worden aangedaan, dan stijgt zo'n punt of pad in de waardering van de bot. Op die manier ontstaan via padanalyse bepaalde surfprofielen.

6.4 Tenslotte

In dit hoofdstuk werd een casus uit het ICT-gebied gekozen, te weten websites-tatistiekanalyse, specifiek die van Webstats4u, en helemaal specifiek die van startpagina.nl. Aan de hand van de in dat domein verzamelde data zijn enkele voorbeelden van het herkennen van de juiste variabelen en het kiezen van de juiste analysetechniek gekozen.

Chapter 7

Conclusie

7.1 Ter afsluiting

We hebben in deze reader gekeken naar het prepareren van data voor kwantitatieve analyse en naar diverse vormen van kwantitatieve analyse.

Hierbij valt op dat het goed voorbereiden van de data voor kwantitatieve analyse nog een behoorlijk ingrijpende klus is, die zeker niet te lichtvaardig mag worden opgevat. Als het voorbereiden van de data al niet zorgvuldig gebeurt, dan is de kans erg groot dat de resultaten van de kwantitatieve analyse een verkeerd beeld zullen geven van de resultaten van het verrichte onderzoek.

Wat de analyse zelf betreft hebben we eerst gekeken naar analyse van één variabele. We hebben hierbij onderscheid gemaakt tussen vier soorten variabelen en mogelijke technieken voor analyse erbij genoemd. Het is uiteraard van het grootste belang de juiste soort variabele te benoemen: ook het verkeerd benoemen van een variabele heeft negatieve gevolgen voor de analyse van je data.

Vervolgens hebben we gekeken naar hoe geanalyseerd kan worden als we met twee gerelateerde variabelen te maken hebben. Hierbij hebben we ons beperkt tot twee soorten variabelen: nominale en intervalvariabelen, aangezien we anders zeker zouden verdwalen in en oerwoud van verschillende technieken.

Tenslotte hebben we gekeken naar analyse van meerdere variabelen. Allereerst hebben we gekeken naar een manier om te achterhalen of we daadwerkelijk te maken hebben met een situatie waarin meerdere variabelen een rol spelen: de elaboratietechniek. Voor de verdere analyse van meerdere variabelen is domweg een greep uit de beschikbare technieken of soorten technieken gekozen die ook goed kunnen worden gebruikt voor analyse van nominale en intervalvariabelen. Ook hier geldt dat het oerwoud van beschikbare technieken niet te overzien is. Bovendien zijn er dan ook nog technieken die meerdere technieken combineren en technieken om de variabelen te clusteren zodat er nog enig overzicht kan worden gecreëerd.

In de casus bleek dat het vrij goed gaat om van een gegeven set data toch wat analyses te maken met een of twee variabelen. Hier blijkt ook dat het vooral

belangrijk is om de juiste analysemethode te kiezen. Als dat eenmaal is gelukt dan kan een bijpassende methode makkelijk worden gevonden en toegepast.

7.2 Ter conclusie

Te concluderen valt dat er een enorme hoeveelheid bestaat van beschikbare analysetechnieken voor kwantitatieve analyse. Het is dan ook niet lastig om al snel de verkeerde variabele te benoemen, de verkeerde techniek te kiezen of domweg verkeerde conclusies te trekken, bijvoorbeeld wegens schijnsamenhang. En dan hebben we het nog niet eens over het interpreteren of toepassen van de gegenereerde statistieken: ook daar zijn de nodige invalshoeken te benoemen. De doelstelling van deze reader was vooral om een overzicht te geven van beschikbare analysetechnieken die gebruikt kunnen worden bij het analyseren van kwantitatieve data en de weg naar die kwantitatieve analyse toe: het voorbereiden en voorbewerken van data. Dat doel lijkt vrij goed geslaagd.

Appendix A

Data startpagina.nl

In deze appendix zijn de tabellen met de statistische gegevens zoals gebruikt op webstats4u opgenomen die het uitgangspunt vormen voor de voorbeelden zoals gegeven in de eerste casus.

Meet sinds	30 september 1998
Totaal aantal pageviews tot nu toe	4318907754
Drukste dag tot nu toe	24 oktober 2005
Pageviews	4089331
Pageviews vandaag	3751586
Pageviews gisteren	3984021

Table A.1: Algemene gegevens van de teller op startpagina.nl

nr.	Land van Herkomst	Aantal	percentage
1.	Nederland	4090527263	94.7 %
2.	België	20470670	0.5 %
3.	Verenigde Staten	16811548	0.4 %
4.	Duitsland	11258392	0.3 %
5.	Netwerk	9129221	0.2 %
6.	Verenigd Koninkrijk	8600845	0.2 %
7.	VS Commercieel	8297498	0.2 %
8.	Zweden	6623837	0.2 %
9.	Frankrijk	5416227	0.1 %
10.	Spanje	3425709	0.1 %
	Onbekend	116520279	2.7 %
	De rest	21826265	0.5 %
	Totaal	4318907754	100.0 %

Table A.2: gegevens land van herkomst bezoekers

Datum	Pageviews
23 november 2005, woensdag	3843954
24 november 2005, donderdag	3804283
25 november 2005, vrijdag	3844663
26 november 2005, zaterdag	3023317
27 november 2005, zondag	3356678
28 november 2005, maandag	4030275
29 november 2005, dinsdag	3908645
30 november 2005, woensdag	3831276
1 december 2005, donderdag	3722745
2 december 2005, vrijdag	3528004
3 december 2005, zaterdag	2951228
4 december 2005, zondag	3254588
5 december 2005, maandag	3818416
6 december 2005, dinsdag	3815455
7 december 2005, woensdag	3807861
8 december 2005, donderdag	3620314
9 december 2005, vrijdag	3420074
10 december 2005, zaterdag	2800239
11 december 2005, zondag	3245148
12 december 2005, maandag	3945447
13 december 2005, dinsdag	3792278
14 december 2005, woensdag	3777292
15 december 2005, donderdag	3629919
16 december 2005, vrijdag	3537560
17 december 2005, zaterdag	3017728
18 december 2005, zondag	3359956
19 december 2005, maandag	3984021
20 december 2005, dinsdag	3751586
Totaal	100422950

Table A.3: Aantal bezoekers startpagina per dag van de afgelopen vier weken

nr.	datum	tijd	provider
1.	21 december	02:16	Chello, Nederland
2.	21 december	02:16	Planet Internet, Nederland
3.	21 december	02:16	Casema NV, Nederland
4.	21 december	02:16	Essent Kabelcom, Nederland
5.	21 december	02:16	Het Net, Nederland
6.	21 december	02:16	Chello, Nederland
7.	21 december	02:16	Planet Internet, Nederland
8.	21 december	02:16	Chello, Helmond, Nederland
9.	21 december	02:16	XS4all Internet B.V., Nederland
10.	21 december	02:16	Vertis BV, Nederland
11.	21 december	02:16	Chello, Amsterdam, Nederland
12.	21 december	02:16	Infopact, Nederland
13.	21 december	02:16	Wanadoo Nederland BV, Nederland
14.	21 december	02:16	Demon Internet, Nederland
15.	21 december	02:16	Planet Internet, Nederland
16.	21 december	02:16	@Home, Den Bosch, Nederland
17.	21 december	02:16	@Home, Roosendaal, Nederland
18.	21 december	02:16	Wanadoo Nederland BV, Nederland
19.	21 december	02:16	Universiteit Twente, Enschede, Nederland
20.	21 december	02:17	Wanadoo Nederland BV, Nederland
21.	21 december	02:17	BetuweNet, Nederland
22.	21 december	02:17	Wanadoo Nederland BV, Nederland
23.	21 december	02:17	Cambrium BV, Nederland
24.	21 december	02:17	Planet Internet, Nederland
25.	21 december	02:17	Chello, Nederland
26.	21 december	02:17	Planet Internet, Nederland
27.	21 december	02:17	Chello, Nederland
28.	21 december	02:17	Chello, Apeldoorn, Nederland
29.	21 december	02:17	Quicknet BV, Nederland
30.	21 december	02:17	@Home, Roosendaal, Nederland
31.	21 december	02:17	Versatel, Nederland
32.	21 december	02:17	Wanadoo Nederland BV, Nederland
33.	21 december	02:17	KLM, Schiphol, Nederland
34.	21 december	02:17	Royal Wessanen N.V., Amstelveen, Nederland
35.	21 december	02:17	Planet Internet, Nederland
36.	21 december	02:17	Wanadoo Nederland BV, Nederland
37.	21 december	02:17	Wanadoo Nederland BV, Nederland
38.	21 december	02:17	EuroNet Internet, Nederland
39.	21 december	02:17	XS4all Internet B.V., Nederland
40.	21 december	02:17	Chello Broadband GmbH, Oostenrijk
41.	21 december	02:17	@Home, Roosendaal, Nederland
42.	21 december	02:17	Chello, Nederland
43.	21 december	02:17	@Home, Den Bosch, Nederland
44.	21 december	02:17	Chello, Nijmegen, Nederland
45.	21 december	02:17	@Home, Tilburg, Nederland
46.	21 december	02:17	MCI, Nederland
47.	21 december	02:17	Chello, Nederland
48.	21 december	02:17	XS4all Internet B.V., Nederland
49.	21 december	02:17	Chello, Nederland
50.	21 december	02:17	@Home, Nederweert, Nederland

Table A.4: De laatste vijftig bezoekers, steekproef 1

nr.	datum	tijd	provider
1.	21 december	12:32	Versatel, Nederland
2.	21 december	12:32	France Telecom, Frankrijk
3.	21 december	12:32	@Home, Roosendaal, Nederland
4.	21 december	12:32	Planet Internet, Nederland
5.	21 december	12:32	Planet Internet, Nederland
6.	21 december	12:32	Planet Internet, Nederland
7.	21 december	12:32	Tiscali, Nederland
8.	21 december	12:32	Wanadoo Nederland BV, Nederland
9.	21 december	12:32	SpeedXS, Nederland
10.	21 december	12:32	Chello, Purmerend, Nederland
11.	21 december	12:33	Chello, Nederland
12.	21 december	12:33	Planet Internet, Nederland
13.	21 december	12:33	Telefonia Bonairiano N.V., Nederlandse Antillen
14.	21 december	12:33	Planet Internet, Nederland
15.	21 december	12:33	Chello, Zutphen, Nederland
16.	21 december	12:33	Kumamoto University, Kumamoto, Japan
17.	21 december	12:33	CAIW, Aalsmeer, Nederland
18.	21 december	12:33	Quicknet BV, Nederland
19.	21 december	12:33	Chello, Zaandam, Nederland
20.	21 december	12:33	Planet Internet, Nederland
21.	21 december	12:33	Quicknet BV, Nederland
22.	21 december	12:33	Wanadoo Nederland BV, Nederland
23.	21 december	12:33	Chello, Nederland
24.	21 december	12:33	Versatel, Nederland
25.	21 december	12:33	Chello, Nederland
26.	21 december	12:33	Quicknet BV, Nederland
27.	21 december	12:33	Tiscali, Nederland
28.	21 december	12:33	Wanadoo Nederland BV, Nederland
29.	21 december	12:33	Chello, Helmond, Nederland
30.	21 december	12:33	@Home, Tilburg, Nederland
31.	21 december	12:33	Scarlet Network, Rotterdam, Nederland
32.	21 december	12:33	Versatel, Nederland
33.	21 december	12:33	Tiscali, Nederland
34.	21 december	12:33	Chello, Nederland
35.	21 december	12:33	Chello, Nederland
36.	21 december	12:33	@Home, Landgraaf, Nederland
37.	21 december	12:33	Royal Wessanen N.V., Amstelveen, Nederland
38.	21 december	12:33	Chello, Lelystad, Nederland
39.	21 december	12:33	@Home, Landgraaf, Nederland
40.	21 december	12:33	Planet Internet, Nederland
41.	21 december	12:33	Demon Internet, Nederland
42.	21 december	12:33	Planet Internet, Nederland
43.	21 december	12:33	Chello, Nederland
44.	21 december	12:33	Wanadoo Nederland BV, Nederland
45.	21 december	12:33	Chello, Almere, Nederland
46.	21 december	02:33	Wanadoo Nederland BV, Nederland
47.	21 december	02:33	Wanadoo Nederland BV, Nederland
48.	21 december	12:33	Tiscali, Nederland
49.	21 december	12:33	Chello, Nederland
50.	21 december	12:33	Wanadoo Nederland BV, Nederland

Table A.5: De laatste vijftig bezoekers, steekproef 2

nr.	datum	tijd	provider
1.	22 december	16:05	XS4all Internet B.V., Nederland
2.	22 december	16:05	XS4all Internet B.V., Nederland
3.	22 december	16:05	Wanadoo Nederland BV, Nederland
4.	22 december	16:05	Chello, Nederland
5.	22 december	16:05	Wanadoo Nederland BV, Nederland
6.	22 december	16:05	Planet Internet, Nederland
7.	22 december	16:05	XS4all Internet B.V., Nederland
8.	22 december	16:05	Quicknet BV, Nederland
9.	22 december	16:05	Tiscali, Nederland
10.	22 december	16:05	Chello, Nederland
11.	22 december	16:05	@Home, Roosendaal, Nederland
12.	22 december	16:05	@Home, Den Bosch, Nederland
13.	22 december	16:05	HCCnet, Nederland
14.	22 december	16:05	Planet Internet, Nederland
15.	22 december	16:05	Planet Internet, Nederland
16.	22 december	16:05	Versatel, Zwolle, Nederland
17.	22 december	16:05	Planet Internet, Nederland
18.	22 december	16:05	XS4all Internet B.V., Nederland
19.	22 december	16:05	Chello, Nederland
20.	22 december	16:05	Het Net, Nederland
21.	22 december	16:05	Planet Internet, Nederland
22.	22 december	16:05	Planet Internet, Nederland
23.	22 december	16:05	Arbo Unie B.V., Amsterdam, Nederland
24.	22 december	16:05	Planet Internet, Nederland
25.	22 december	16:05	Tiscali, Nederland
26.	22 december	16:05	ChevronTexaco Corporation, Canada
27.	22 december	16:05	Casema NV, Nederland
28.	22 december	16:05	Quicknet BV, Nederland
29.	22 december	16:05	@Home, Venlo, Nederland
30.	22 december	16:05	Tiscali, Nederland
31.	22 december	16:05	@Home, Mill, Nederland
32.	22 december	16:05	@Home, Oldenzaal, Nederland
33.	22 december	16:05	Planet Internet, Nederland
34.	22 december	16:05	Cambrium BV, Nederland
35.	22 december	16:05	Chello, Nederland
36.	22 december	16:05	Solcon Holding BV, Nederland
37.	22 december	16:05	Het Net, Nederland
38.	22 december	16:05	Chello, Leeuwarden, Nederland
39.	22 december	16:05	Planet Internet, Nederland
40.	22 december	16:05	Chello, Nederland
41.	22 december	16:05	Chello, Almere, Nederland
42.	22 december	16:05	Planet Internet, Nederland
43.	22 december	16:05	Planet Internet, Nederland
44.	22 december	16:05	Universiteit Maastricht, Maastricht, Nederland
45.	22 december	16:05	@Home, Enschede, Nederland
46.	22 december	16:05	Versatel, Nederland
47.	22 december	16:05	Planet Internet, Nederland
48.	22 december	16:05	Planet Internet, Nederland
49.	22 december	16:05	@Home, Mill, Nederland
50.	22 december	16:05	HCCnet, Nederland

Table A.6: De laatste vijftig bezoekers, steekproef 3

nr.	datum	tijd	provider
1.	23 december	10:19	Chello, Amsterdam, Nederland
2.	23 december	10:19	Versatel, Nederland
3.	23 december	10:19	ZeelandNet, Terneuzen, Nederland
4.	23 december	10:19	@Home, Roosendaal, Nederland
5.	23 december	10:19	Wanadoo Nederland BV, Nederland
6.	23 december	10:19	Chello, Nederland
7.	23 december	10:19	CAIW, Schiedam, Nederland
8.	23 december	10:19	Chello, Nederland
9.	23 december	10:19	Versatel, Nederland
10.	23 december	10:19	HCCnet, Nederland
11.	23 december	10:19	Planet Internet, Nederland
12.	23 december	10:19	Chello, Nederland
13.	23 december	10:19	Casema NV, Nederland
14.	23 december	10:19	Planet Internet, Nederland
15.	23 december	10:19	Chello, Nederland
16.	23 december	10:19	Versatel, Nederland
17.	23 december	10:19	Priority Telecom, Noorwegen
18.	23 december	10:19	Planet Internet, Nederland
19.	23 december	10:19	Chello, Nederland
20.	23 december	10:19	Wanadoo Nederland BV, Nederland
21.	23 december	10:19	CAIW, Nederland
22.	23 december	10:19	EuroNet Internet, Nederland
23.	23 december	10:19	Planet Internet, Nederland
24.	23 december	10:19	Tiscali, Nederland
25.	23 december	10:19	@Home, Enschede, Nederland
26.	23 december	10:19	Chello, Wormerveer, Nederland
27.	23 december	10:19	Zonnet, Nederland
28.	23 december	10:19	MCI, Nederland
29.	23 december	10:19	HCCnet, Nederland
30.	23 december	10:19	KPN, Nederland
31.	23 december	10:19	Haagse Hogeschool, Den Haag, Nederland
32.	23 december	10:19	Tiscali, Nederland
33.	23 december	10:19	Het Net, Nederland
34.	23 december	10:19	Planet Internet, Nederland
35.	23 december	10:19	Telfort BV, Nederland
36.	23 december	10:19	Tiscali, Nederland
37.	23 december	10:19	Wanadoo Nederland BV, Nederland
38.	23 december	10:19	Jan Zandbergen, Veenendaal, Nederland
39.	23 december	10:19	Nefit Buderus BV, Deventer, Nederland
40.	23 december	10:19	@Home, Enschede, Nederland
41.	23 december	10:19	@Home, Nederland
42.	23 december	10:19	Planet Internet, Nederland
43.	23 december	10:19	Chello, Nederland
44.	23 december	10:19	Chello, Nederland
45.	23 december	10:19	EuroNet Internet, Nederland
46.	23 december	10:19	Casema NV, Nederland
47.	23 december	10:19	XS4all Internet B.V., Nederland
48.	23 december	10:19	CAIW, Naaldwijk, Nederland
49.	23 december	10:19	Swipnet, Zweden
50.	23 december	10:19	Planet Internet, Nederland

Table A.7: De laatste vijftig bezoekers, steekproef 4

nr.	datum	tijd	provider
1.	23 december	12:25	Versatel, Nederland
2.	23 december	12:25	Planet Internet, Nederland
3.	23 december	12:25	Wanadoo Nederland BV, Nederland
4.	23 december	12:25	Planet Internet, Nederland
5.	23 december	12:25	Versatel, Amsterdam, Nederland
6.	23 december	12:25	Chello, Nederland
7.	23 december	12:25	@Home, Den Bosch, Nederland
8.	23 december	12:25	Wanadoo Nederland BV, Nederland
9.	23 december	12:25	Versatel, Nederland
10.	23 december	12:25	Chello, Nederland
11.	23 december	12:25	@Home, Venlo, Nederland
12.	23 december	12:25	Planet Internet, Nederland
13.	23 december	12:25	Hoogheemraadschap Noorderkwartier, Purmerend, Nederland
14.	23 december	12:25	Skynet Belgacom, België
15.	23 december	12:25	EuroNet Internet, Nederland
16.	23 december	12:25	@Home, Den Bosch, Nederland
17.	23 december	12:25	@Home, Den Bosch, Nederland
18.	23 december	12:25	Planet Internet, Nederland
19.	23 december	12:25	Versatel, Nederland
20.	23 december	12:25	Planet Internet, Nederland
21.	23 december	12:25	Wanadoo Nederland BV, Nederland
22.	23 december	12:25	Planet Internet, Nederland
23.	23 december	12:25	KPN, Nederland
24.	23 december	12:25	DaimlerChrysler Nederland, Utrecht, Nederland
25.	23 december	12:25	N.V. Nutsbedrijven Maastricht, Maastricht, Nederland
26.	23 december	12:25	Tiscali, Nederland
27.	23 december	12:25	Chello, Nederland
28.	23 december	12:25	Casema NV, Nederland
29.	23 december	12:25	Quicknet BV, Nederland
30.	23 december	12:25	XB Networks BV, Nederland
31.	23 december	12:25	Tiscali, Nederland
32.	23 december	12:25	Chello, Nederland
33.	23 december	12:25	Planet Internet, Nederland
34.	23 december	12:25	KPN, Nederland
35.	23 december	12:25	Planet Internet, Nederland
36.	23 december	12:25	F. Hoffman-LaRoche Ltd., Zwitserland
37.	23 december	12:25	Casema NV, Nederland
38.	23 december	12:25	Planet Internet, Nederland
39.	23 december	12:25	uicknet BV, Nederland
40.	23 december	12:25	Swipnet, Zweden
41.	23 december	12:25	Het Net, Nederland
42.	23 december	12:25	Planet Internet, Nederland
43.	23 december	12:25	Versatel, Nederland
44.	23 december	12:25	Planet Internet, Nederland
45.	23 december	12:25	Planet Internet, Nederland
46.	23 december	12:25	@Home, Zwolle, Nederland
47.	23 december	12:25	Casema NV, Nederland
48.	23 december	12:25	Planet Internet, Nederland
49.	23 december	12:25	Chello, Nederland
50.	23 december	12:25	@Home, Dokkum, Nederland

Table A.8: De laatste vijftig bezoekers, steekproef 5

nr.	datum	tijd	provider
1.	23 december	13:08	Chello, Nederland
2.	23 december	13:08	Planet Internet, Nederland
3.	23 december	13:08	Chello, Nederland
4.	23 december	13:08	Planet Internet, Nederland
5.	23 december	13:08	Planet Internet, Nederland
6.	23 december	13:08	Planet Internet, Nederland
7.	23 december	13:08	EuroNet Internet, Zaltbommel, Nederland
8.	23 december	13:08	Planet Internet, Nederland
9.	23 december	13:08	Demon Internet, Nederland
10.	23 december	13:08	Planet Internet, Nederland
11.	23 december	13:08	Planet Internet, Nederland
12.	23 december	13:08	Chello, Nederland
13.	23 december	13:08	Planet Internet, Nederland
14.	23 december	13:08	XS4all Internet B.V., Nederland
15.	23 december	13:08	XS4all Internet B.V., Nederland
16.	23 december	13:08	@Home, Almelo, Nederland
17.	23 december	13:08	Wanadoo Nederland BV, Nederland
18.	23 december	13:08	Wanadoo Nederland BV, Amersfoort, Nederland
19.	23 december	13:08	Quicknet BV, Nederland
20.	23 december	13:08	@Home, Emmen, Nederland
21.	23 december	13:08	Planet Internet, Nederland
22.	23 december	13:08	Planet Internet, Nederland
23.	23 december	13:08	Planet Internet, Nederland
24.	23 december	13:08	Chello, Purmerend, Nederland
25.	23 december	13:08	Quicknet BV, Nederland
26.	23 december	13:08	XS4all Internet B.V., Nederland
27.	23 december	13:08	Chello, Nijmegen, Nederland
28.	23 december	13:08	Sticht. Studenten Huisvesting, Nijmegen, Nederland
29.	23 december	13:08	HCCnet, Nederland
30.	23 december	13:08	Universiteit Leiden, Leiden, Nederland
31.	23 december	13:08	Cubic Circle, Nederland
32.	23 december	13:08	Verenigde Staten (alaskaair.com)
33.	23 december	13:08	Chello, Nederland
34.	23 december	13:08	Het Net, Nederland
35.	23 december	13:08	Wanadoo Nederland BV, Nederland
36.	23 december	13:08	XS4all Internet B.V., Nederland
37.	23 december	13:08	@Home, Groningen, Nederland
38.	23 december	13:08	XS4all Internet B.V., Nederland
39.	23 december	13:08	HCCnet, Nederland
40.	23 december	13:08	Planet Internet, Nederland
41.	23 december	13:08	Scarlet Network, Nederland
42.	23 december	13:08	Wanadoo Nederland BV, Nederland
43.	23 december	13:08	Chello, Lelystad, Nederland
44.	23 december	13:08	Het Net, Nederland
45.	23 december	13:08	Versatel, Nederland
46.	23 december	13:08	Chello, Nederland
47.	23 december	13:08	Wanadoo Nederland BV, Nederland
48.	23 december	13:08	Versatel, Nederland
49.	23 december	13:08	Het Net, Nederland
50.	23 december	13:08	@Home, Groningen, Nederland

Table A.9: De laatste vijftig bezoekers, steekproef 6

nr.	datum	tijd	provider
1.	23 december	13:37	Planet Internet, Nederland
2.	23 december	13:37	CAIW, Naaldwijk, Nederland
3.	23 december	13:37	@Home, Nederweert, Nederland
4.	23 december	13:37	Planet Internet, Nederland
5.	23 december	13:37	Planet Internet, Nederland
6.	23 december	13:37	Scarlet Network, Nederland
7.	23 december	13:37	Casema NV, Nederland
8.	23 december	13:37	Wanadoo Nederland BV, Nederland
9.	23 december	13:37	Wanadoo Nederland BV, Nederland
10.	23 december	13:37	Wanadoo Nederland BV, Nederland
11.	23 december	13:37	@Home, Tilburg, Nederland
12.	23 december	13:37	Chello, Nederland
13.	23 december	13:37	Chello, Nederland
14.	23 december	13:37	@Home, Harlingen, Nederland
15.	23 december	13:37	Scarlet Network, Nederland
16.	23 december	13:37	Planet Internet, Nederland
17.	23 december	13:37	KPNQwest, Duitsland
18.	23 december	13:37	Endemol Entertainment Holding, Hilversum, Nederland
19.	23 december	13:37	Wanadoo Nederland BV, Nederland
20.	23 december	13:37	Wanadoo Nederland BV, Nederland
21.	23 december	13:37	@Home, Groningen, Nederland
22.	23 december	13:37	Planet Internet, Nederland
23.	23 december	13:37	Wanadoo Nederland BV, Nederland
24.	23 december	13:37	Tiscali, Nederland
25.	23 december	16:37	Casema NV, Nederland
26.	23 december	13:37	Chello, Amsterdam, Nederland
27.	23 december	13:37	Planet Internet, Nederland
28.	23 december	13:37	Wanadoo Nederland BV, Nederland
29.	23 december	13:37	XS4all Internet B.V., Nederland
30.	23 december	13:37	Novaxess BV, Nederland
31.	23 december	13:37	Planet Internet, Nederland
32.	23 december	13:37	Demon Internet, Nederland
33.	23 december	13:37	@Home, Den Bosch, Nederland
34.	23 december	13:37	Planet Internet, Nederland
35.	23 december	13:37	@Home, Tilburg, Nederland
36.	23 december	13:37	Versatel, Nederland
37.	23 december	13:37	Chello, Nederland
38.	23 december	13:37	Versatel, Nederland
39.	23 december	13:37	Quicknet BV, Nederland
40.	23 december	13:37	Wanadoo Nederland BV, Nederland
41.	23 december	13:37	Planet Internet, Nederland
42.	23 december	13:37	Tiscali, Nederland
43.	23 december	13:37	Planet Internet, Nederland
44.	23 december	13:37	HCCnet, Nederland
45.	23 december	13:37	Wanadoo Nederland BV, Nederland
46.	23 december	13:37	@Home, Den Bosch, Nederland
47.	23 december	13:37	Wanadoo Nederland BV, Nederland
48.	23 december	13:37	Chello, Helmond, Nederland
49.	23 december	13:37	Planet Internet, Nederland
50.	23 december	13:37	HCCnet, Nederland

Table A.10: De laatste vijftig bezoekers, steekproef 7

nr.	datum	tijd	provider
1.	23 december	14:02	Chello, Amsterdam, Nederland
2.	23 december	14:02	Chello, Nederland
3.	23 december	14:02	Demon Internet, Nederland
4.	23 december	14:02	Planet Internet, Nederland
5.	23 december	14:02	@Home, Emmen, Nederland
6.	23 december	14:02	@Home, Landgraaf, Nederland
7.	23 december	14:02	Wanadoo Nederland BV, Nederland
8.	23 december	14:02	Planet Internet, Nederland
9.	23 december	14:02	Planet Internet, Nederland
10.	23 december	14:02	Casema NV, Nederland
11.	23 december	14:02	@Home, Roden, Nederland
12.	23 december	14:02	Tiscali, Nederland
13.	23 december	14:02	@Home, Tilburg, Nederland
14.	23 december	14:02	Planet Internet, Nederland
15.	23 december	14:02	Planet Internet, Nederland
16.	23 december	14:02	Planet Internet, Nederland
17.	23 december	14:02	@Home, Mill, Nederland
18.	23 december	14:02	HCCnet, Nederland
19.	23 december	14:02	Chello, Nederland
20.	23 december	14:02	Chello, Nederland
21.	23 december	14:02	Chello, Zaanstad, Nederland
22.	23 december	14:02	ZeelandNet, Goes, Nederland
23.	23 december	14:02	Planet Internet, Nederland
24.	23 december	14:02	Tiscali, Nederland
25.	23 december	14:02	Chello, Nederland
26.	23 december	14:02	Wanadoo Nederland BV, Nederland
27.	23 december	14:02	Chello, Nederland
28.	23 december	14:02	Casema NV, Nederland
29.	23 december	14:02	Cubic Circle, Nederland
30.	23 december	14:02	Wanadoo Nederland BV, Nederland
31.	23 december	14:02	Vertis BV, Nederland
32.	23 december	14:02	CAIW, Nederland
32.	23 december	14:02	Planet Internet, Nederland
34.	23 december	14:02	@Home, Roosendaal, Nederland
35.	23 december	14:02	Quicknet BV, Nederland
36.	23 december	14:02	Planet Internet, Nederland
37.	23 december	14:02	@Home, Venray, Nederland
38.	23 december	14:02	XB Networks BV, Nederland
39.	23 december	14:02	Wanadoo Nederland BV, Nederland
40.	23 december	14:02	Wanadoo Nederland BV, Nederland
41.	23 december	14:02	Demon Internet, Nederland
42.	23 december	14:02	Chello, Nederland
43.	23 december	14:02	@Home, Tilburg, Nederland
44.	23 december	14:02	Fiberworld BV, Nederland
45.	23 december	14:02	Planet Internet, Nederland
46.	23 december	14:02	CAIW, Nederland
47.	23 december	14:02	Quicknet BV, Nederland
48.	23 december	14:02	Chello, Nederland
49.	23 december	14:02	Versatel, Nederland
50.	23 december	14:02	Wanadoo Nederland BV, Nederland

Table A.11: De laatste vijftig bezoekers, steekproef 8

nr.	datum	tijd	provider
1.	23 december	14:51	Chello, Nederland
2.	23 december	14:51	Chello Broadband GmbH, Oostenrijk
3.	23 december	14:51	Wanadoo Nederland BV, Nederland
4.	23 december	14:51	Fiberworld BV, Leeuwarden, Nederland
5.	23 december	14:51	@Home, Roermond, Nederland
6.	23 december	14:51	Quicknet BV, Nederland
7.	23 december	14:51	CAIW, Naaldwijk, Nederland
8.	23 december	14:51	Planet Internet, Nederland
9.	23 december	14:51	Wanadoo Nederland BV, Nederland
10.	23 december	14:51	@Home, Mill, Nederland
11.	23 december	14:51	Versatel, Nederland
12.	23 december	14:51	Planet Internet, Nederland
13.	23 december	14:51	Chello, Nederland
14.	23 december	14:51	Chello, Leeuwarden, Nederland
15.	23 december	14:51	Het Net, Nederland
16.	23 december	14:51	@Home, Almelo, Nederland
17.	23 december	14:51	XS4all Internet B.V., Nederland
18.	23 december	14:51	Wanadoo Nederland BV, Nederland
19.	23 december	14:51	EuroNet Internet, Nederland
20.	23 december	14:51	Planet Internet, Nederland
21.	23 december	14:51	Chello, Nederland
22.	23 december	14:51	BBeyond, Nederland
23.	23 december	14:51	Planet Internet, Nederland
24.	23 december	14:51	Speedlinq, Nederland
25.	23 december	14:51	@Home, Den Bosch, Nederland
26.	23 december	14:51	Planet Internet, Nederland
27.	23 december	14:51	Wanadoo Nederland BV, Nederland
28.	23 december	14:51	Planet Internet, Nederland
29.	23 december	14:51	@Home, Den Bosch, Nederland
30.	23 december	14:51	Planet Internet, Nederland
31.	23 december	14:51	Swipnet, Zweden
32.	23 december	14:51	Chello, Nederland
33.	23 december	14:51	Planet Internet, Nederland
34.	23 december	14:51	Planet Internet, Nederland
35.	23 december	14:51	EUnet Finland, Finland
36.	23 december	14:51	Chello, Zaanstad, Nederland
37.	23 december	14:51	@Home, Roermond, Nederland
38.	23 december	14:51	Internet Online BV, Nederland
39.	23 december	14:51	Planet Internet, Nederland
40.	23 december	14:51	Chello Broadband GmbH, Oostenrijk
41.	23 december	14:51	Chello, Leeuwarden, Nederland
42.	23 december	14:51	XS4all Internet B.V., Nederland
43.	23 december	14:51	Chello, Nederland
44.	23 december	14:51	Planet Internet, Nederland
45.	23 december	14:51	Planet Internet, Nederland
46.	23 december	14:51	Enertel N.V., Nederland
47.	23 december	14:51	Planet Internet, Nederland
48.	23 december	14:51	@Home, Groningen, Nederland
49.	23 december	14:51	Wanadoo Nederland BV, Nederland
50.	23 december	14:51	Planet Internet, Nederland

Table A.12: De laatste vijftig bezoekers, steekproef 9

nr.	datum	tijd	provider
1.	23 december	15:55	Chello, Amsterdam, Nederland
2.	23 december	15:55	KPN, Nederland
3.	23 december	15:55	Versatel, Nederland
4.	23 december	15:55	Planet Internet, Nederland
5.	23 december	15:55	Planet Internet, Nederland
6.	23 december	15:55	InterConnect Services BV, Nederland
7.	23 december	15:55	XS4all Internet B.V., Nederland
8.	23 december	15:55	Chello, Nederland
9.	23 december	15:55	Chello Broadband GmbH, Oostenrijk
10.	23 december	15:55	HCCnet, Nederland
11.	23 december	15:55	Planet Internet, Nederland
12.	23 december	15:55	@Home, Zwolle, Nederland
13.	23 december	15:55	Casema NV, Nederland
14.	23 december	15:55	Chello, Nederland
15.	23 december	15:55	Planet Internet, Nederland
16.	23 december	15:55	XS4all Internet B.V., Nederland
17.	23 december	15:55	iscali, Nederland
18.	23 december	15:55	Planet Internet, Nederland
19.	23 december	15:55	@Home, Venray, Nederland
20.	23 december	15:55	Home, Den Bosch, Nederland
21.	23 december	15:55	Wanadoo Nederland BV, Nederland
22.	23 december	15:55	Quicknet BV, Nederland
23.	23 december	15:55	SURFnet, Nederland
24.	23 december	15:55	Concepts, Nederland
25.	23 december	15:55	Planet Internet, Nederland
26.	23 december	15:55	Telenet, België
27.	23 december	15:55	Quicknet BV, Nederland
28.	23 december	15:55	Planet Internet, Nederland
29.	23 december	15:55	Casema NV, Nederland
30.	23 december	15:56	Chello, Amsterdam, Nederland
31.	23 december	15:56	InterNLnet, Nederland
32.	23 december	15:56	XS4all Internet B.V., Nederland
33.	23 december	15:56	Tiscali, Nederland
34.	23 december	15:56	Planet Internet, Nederland
35.	23 december	15:56	Demon Internet, Nederland
36.	23 december	15:56	Quicknet BV, Nederland
37.	23 december	15:56	Casema NV, Nederland
38.	23 december	15:56	Demon Internet, Nederland
39.	23 december	15:56	@Home, Nederland
40.	23 december	15:56	@Home, Landgraaf, Nederland
41.	23 december	15:56	XS4all Internet B.V., Nederland
42.	23 december	15:56	Telenet, België
43.	23 december	15:56	Chello, Nederland
44.	23 december	15:56	@Home, Landgraaf, Nederland
45.	23 december	15:56	@Home, Den Bosch, Nederland
46.	23 december	15:56	XS4all Internet B.V., Nederland
47.	23 december	15:56	Chello, Helmond, Nederland
48.	23 december	15:56	Planet Internet, Nederland
49.	23 december	15:56	Planet Internet, Nederland
50.	23 december	15:56	Tiscali, Nederland

Table A.13: De laatste vijftig bezoekers, steekproef 10

nr.	datum	tijd	provider
1.	23 december	16:01	Planet Internet, Nederland
2.	23 december	16:01	Planet Internet, Nederland
3.	23 december	16:01	Infopact, Nederland
4.	23 december	16:01	Planet Internet, Nederland
5.	23 december	16:01	Demon Internet, Nederland
6.	23 december	16:01	XS4all Internet B.V., Nederland
7.	23 december	16:01	Demon Internet, Nederland
8.	23 december	16:01	CAIW, Zevenbergen, Nederland
9.	23 december	16:01	@Home, Den Bosch, Nederland
10.	23 december	16:01	Planet Internet, Nederland
11.	23 december	16:01	XS4all Internet B.V., Nederland
12.	23 december	16:01	Planet Internet, Nederland
13.	23 december	16:01	Chello, Nederland
14.	23 december	16:01	Chello, Nederland
15.	23 december	16:01	Essent Kabelcom, Nederland
16.	23 december	16:01	Planet Internet, Nederland
17.	23 december	16:01	Chello, Nederland
18.	23 december	16:01	Planet Internet, Nederland
19.	23 december	16:01	MCI, Rotterdam, Nederland
20.	23 december	16:01	XS4all Internet B.V., Nederland
21.	23 december	16:01	Wanadoo Nederland BV, Nederland
22.	23 december	16:01	Chello, Nederland
23.	23 december	16:01	Quicknet BV, Nederland
24.	23 december	16:01	Wanadoo Nederland BV, Nederland
25.	23 december	16:01	Planet Internet, Nederland
26.	23 december	16:01	Planet Internet, Nederland
27.	23 december	16:01	Wanadoo Nederland BV, Nederland
28.	23 december	16:01	Wanadoo Nederland BV, Nederland
29.	23 december	16:01	EuroNet Internet, Nederland
30.	23 december	16:01	Chello, Nederland
31.	23 december	16:01	Casema NV, Nederland
32.	23 december	16:01	Chello, Nederland
33.	23 december	16:01	Quicknet BV, Nederland
34.	23 december	16:01	Chello, Nederland
35.	23 december	16:01	Speedling, Nederland
36.	23 december	16:01	Wanadoo Nederland BV, Nederland
37.	23 december	16:01	XS4all Internet B.V., Nederland
38.	23 december	16:01	Chello, Nederland
39.	23 december	16:01	Planet Internet, Nederland
40.	23 december	16:01	@Home, Geleen, Nederland
41.	23 december	16:01	Versatel, Nederland
42.	23 december	16:01	Planet Internet, Nederland
43.	23 december	16:01	Wanadoo Nederland BV, Nederland
44.	23 december	16:01	Chello, Nederland
45.	23 december	16:01	Planet Internet, Nederland
46.	23 december	16:01	@Home, Venray, Nederland
47.	23 december	16:01	@Home, Landgraaf, Nederland
48.	23 december	16:01	KPN, Nederland
49.	23 december	16:01	Planet Internet, Nederland
50.	23 december	16:01	Planet Internet, Nederland

Table A.14: De laatste vijftig bezoekers, steekproef 11

nr.	datum	tijd	provider
1.	23 december	17:00	@Home, Geleen, Nederland
2.	23 december	17:00	Chello, Lelystad, Nederland
3.	23 december	17:00	Casema NV, Nederland
4.	23 december	17:00	@Home, Oldenzaal, Nederland
5.	23 december	17:00	Chello, Nederland
6.	23 december	17:00	Wanadoo Nederland BV, Nederland
7.	23 december	17:00	Wanadoo Nederland BV, Nederland
8.	23 december	17:00	Solcon Holding BV, Nederland
9.	23 december	17:00	Planet Internet, Nederland
10.	23 december	17:00	Casco Products, Zweden
11.	23 december	17:00	Wanadoo Nederland BV, Nederland
12.	23 december	17:00	Planet Internet, Nederland
13.	23 december	17:00	Planet Internet, Nederland
14.	23 december	17:00	Planet Internet, Nederland
15.	23 december	17:00	Planet Internet, Nederland
16.	23 december	17:00	Planet Internet, Nederland
17.	23 december	17:00	Planet Internet, Nederland
18.	23 december	17:00	@Home, Den Bosch, Nederland
19.	23 december	17:00	@Home, Landgraaf, Nederland
20.	23 december	17:00	Planet Internet, Nederland
21.	23 december	17:00	@Home, Enschede, Nederland
22.	23 december	17:00	HCCnet, Nederland
23.	23 december	17:00	XS4all Internet B.V., Nederland
24.	23 december	17:00	Casema NV, Nederland
25.	23 december	17:00	Wanadoo Nederland BV, Nederland
26.	23 december	17:00	Speedlinq, Nederland
27.	23 december	17:00	Planet Internet, Nederland
28.	23 december	17:00	HCCnet, Nederland
29.	23 december	17:00	@Home, Dokkum, Nederland
30.	23 december	17:00	Chello, Nederland
31.	23 december	17:00	@Home, Venray, Nederland
32.	23 december	17:00	Planet Internet, Nederland
33.	23 december	17:00	Planet Internet, Nederland
34.	23 december	17:00	Chello, Nederland
35.	23 december	17:00	@Home, Venlo, Nederland
36.	23 december	17:00	Chello Broadband GmbH, Oostenrijk
37.	23 december	17:00	XS4all Internet B.V., Nederland
38.	23 december	17:00	Planet Internet, Nederland
39.	23 december	17:00	Chello, Nederland
40.	23 december	17:00	@Home, Den Bosch, Nederland
41.	23 december	17:00	@Home, Mill, Nederland
42.	23 december	17:00	Wanadoo Nederland BV, Nederland
43.	23 december	17:00	@Home, Landgraaf, Nederland
44.	23 december	17:00	Speedlinq, Nederland
45.	23 december	17:00	@Home, Tilburg, Nederland
46.	23 december	17:00	Planet Internet, Nederland
47.	23 december	17:00	@Home, Enschede, Nederland
49.	23 december	17:00	Wanadoo Nederland BV, Nederland
50.	23 december	17:00	Wanadoo Nederland BV, Nederland

Table A.15: De laatste vijftig bezoekers, steekproef 12

Bibliography

- [1] Microsoft excel. <http://www.microsoft.com/netherlands/office/excel/default.aspx>.
- [2] Spss incorporated. <http://www.spss.com>.
- [3] Statistics.com. <http://www.statistics.com>.
- [4] Studiegids psychologie. 2005 / 2006.
- [5] Webstats for you. <http://www.webstats4u.com>.
- [6] Howard Rosenthal David K. Hildebrand and James D. Laing. *Analysis of Ordinal Data*. Sage Publications Ltd, 1977.
- [7] Alphons de Vocht. *Basishandboek SPSS 13*. Bijleveld Press, 2005.
- [8] Jules L. Ellis. *Statistiek voor de psychologie deel 1, Verdeling van een variabele, samenhang tussen twee variabelen*. Uitgeverij Boom, 2004.
- [9] Jules L. Ellis. *Statistiek voor de psychologie deel 2, Toetsen voor twee gemiddelden en toetsingstheorie*. Uitgeverij Boom, 2004.
- [10] Perry Hinton and Charlotte Brownlow. *SPSS Explained*. Routledge, 2004.
- [11] Burkhard Lenze. Mathematics and neural networks - a glance at some basic connections. *Acta Applicandae Mathematicae*, Volume 55(Number 3), 1999.
- [12] McClave and Sincich. *Statistics, Eight Edition*. Prentice Hall, 2000.
- [13] Jo Segers. *Methoden voor de Maatschappijwetenschappen*. Van Gorcum, 2002.
- [14] Bernd Souvigner. Syllabus statistiek voor informatiekunde. 2005.
- [15] Index van de cursussen Informatiekunde aan de Radboud Universiteit Nijmegen. <http://www.niii.ru.nl/onderwijs/opleidingen/>.