

TNO-rapport
PG/VGZ/99.069

Steekproefontwerp: *Een praktische handleiding*

TNO Preventie en Gezondheid

Volksgezondheid
Gortergebouw: Wassenaarseweg 56
Postbus 2215
2301 CE Leiden

Telefoon 071 518 18 18
Fax 071 518 19 20

Datum

December 1999

Auteur

K. Oudshoorn

Het kwaliteitssysteem van
TNO Preventie en Gezondheid
voldoet aan ISO 9001.

Alle rechten voorbehouden.
Niets uit deze uitgave mag worden
vermenigvuldigd en/of openbaar
gemaakt door middel van druk, foto-
kopie, microfilm of op welke andere
wijze dan ook, zonder voorafgaande
toestemming van TNO.

Indien dit rapport in opdracht werd
uitgebracht, wordt voor de rechten en
verplichtingen van opdrachtgever en
opdrachtnemer verwezen naar de
Algemene Voorwaarden voor onder-
zoeksovereenkomsten aan TNO, dan wel
de betreffende terzake tussen de
partijen gesloten overeenkomst.
Het ter inzage geven van het
TNO-rapport aan direct belang-
hebbenden is toegestaan.

© 1999 TNO

Auteur

Karin Oudshoorn

Projectnummer

011.40799

ISBN-nummer

90-6743-643-7

Deze uitgave is te bestellen door het overmaken van *f* 27,85 (incl. BTW) op postbankrekeningnummer 99.889 ten name van TNO-PG te Leiden onder vermelding van bestelnummer PG/VGZ/99.069

Samenvatting

Bij de opzet van een onderzoek wordt meestal op het allerlaatst aandacht besteed aan een onderbouwde steekproefomvang berekening (ook wel powerberekening genaamd). Zo'n berekening hangt af van de opzet van het onderzoek, van de te onderzoeken populatie en nog veel meer factoren. Tijdens de opzet van het onderzoek worden vaak keuzes gemaakt die van invloed zijn op de steekproefomvang berekening. Sterker nog, door andere keuzes tijdens de opzet van het onderzoek te maken kan ervoor gezorgd worden dat de minimale steekproefomvang om een bepaalde nauwkeurigheid van de uitspraken te garanderen lager uitvalt dan als men die keuzes niet zou maken. Dit levert dus een besparing op voor het onderzoeksbudget.

Eén van de keuzes die gemaakt kan worden is die van een bepaald steekproefontwerp. Dat is de manier waarop individuen geselecteerd worden voor de steekproef. In dit rapport zijn een aantal van die ontwerpen besproken en met elkaar vergeleken. Vervolgens zijn voor de meest voorkomende situaties en steekproefontwerpen

De theorie die in de eerste hoofdstukken uit de doeken is gedaan wordt in Hoofdstuk 4 geïllustreerd aan de hand van drie cases die allen onderzoek betreffen uit de Sector Milieu van TNO Preventie en Gezondheid.

In Hoofdstuk 5 wordt een zeker niet omvattend overzicht gegeven van relevante literatuur (boeken en wetenschappelijke artikelen) op het gebied van steekproefontwerp. Daarnaast is in dit hoofdstuk een klein overzicht gegeven van relevante statistische software en enkele internetpagina's waar veel informatie te vinden is over powerberekeningen.

Inhoud

Samenvatting.....	3
Inhoud	5
1 Inleiding	7
2 Verschillende typen steekproefontwerpen.....	9
2.1 Belangrijke basisbegrippen	9
2.2 Enkelvoudige aselechte steekproefontwerp.....	13
2.3 Het enkelvoudige systematisch steekproefontwerp.....	14
2.4 Enkelvoudige steekproefontwerp met ongelijke kansen	16
2.5 Stratificeren	18
2.6 Geclusterde steekproef	20
2.7 Tweetrapssteekproefontwerp.....	22
2.8 Verschillen tussen tweetrapssteekproeven en stratificatie.....	22
2.9 Netwerk steekproefontwerp.....	25
2.10 Onbekende populatieomvang	26
2.11 Non-respons.....	26
2.11.1 Survey non-respons	27
2.11.2 Item non-respons	28
2.12 Samenvatting hoofdstuk 2	30
3 Steekproefomvang berekeningen.....	33
3.1 Algemene begrippen en achtergrond bij steekproefgrootte berekeningen.....	33
3.2 Steekproefomvang berekeningen voor enkelvoudige steekproefontwerpen.....	35
3.3 Steekproefomvang berekeningen bij stratificatie.....	40
3.4 Steekproefgrootte berekeningen bij clustering / systematische steekproeven.....	45
3.5 Optimale steekproefomvang bij getrapte steekproeven.....	48
3.6 Relatie tussen nauwkeurigheid en power	49
3.7 Samenvatting hoofdstuk 3	50
4 Uitwerking cases.....	52
4.1 Case 1, Peiling Milieuhinder: Utrecht 1997.....	52
4.2 Case 2, Slaapverstoringsonderzoek Schiphol.....	55
4.3 Case 3, Dosis-effect relaties wegverkeer en geluidshinder in de EU.....	57

5	Inventarisatie literatuur en software over steekproefontwerp	62
5.1	Boeken	62
5.2	Artikelen	65
5.3	Software	67
5.4	Nuttige internetlinks.....	67
	Referenties.....	69

1 Inleiding

Onderzoek van een populatie gebeurt (meestal) op basis van een steekproef uit de populatie. Bij het opstarten van het onderzoek moet er dus nagedacht worden over hoe de steekproef “getrokken” gaat worden. Met behulp van de steekproef moeten de uitkomsten van de onderzoeksvragen valide (d.w.z. overeenkomend met de werkelijkheid), nauwkeurig (precies oftewel reproduceerbaar) en betrouwbaar beantwoord kunnen worden.

De wijze waarop de steekproef genomen wordt kan opgesplitst worden in twee stappen: bepalen van het steekproefontwerp en vervolgens de bepaling van de steekproefomvang. Het steekproefontwerp is het recept waarmee de steekproef getrokken wordt uit de populatie.

Het rapport bestaat uit vier delen. Hoofdstuk 2, gaat in op verschillende typen steekproefontwerpen en de keuze hiertussen. Hoofdstuk 3, behandelt de vraag “hoe groot moet de steekproef zijn om te komen tot nauwkeurige en betrouwbare uitkomsten”. Het vierde hoofdstuk is gewijd aan drie cases die als voorbeeld dienen voor de theorie in hoofdstuk 2 en 3. Deze drie cases zijn gekozen uit de praktijk van de sector Milieu en betreffen:

- **Peiling milieuhinder: Utrecht 1997**
In kader van het Provinciale Milieubeleidsplan 1998-2002 en ter evaluatie van het uitgevoerde beleid heeft TNO-PG in opdracht van de provincie Utrecht onderzoek verricht naar hinder die in de provincie bij het wonen en het recreëren wordt ondervonden door de milieufactoren geluid, geur, trillingen, stof en licht.
Onderzocht zijn de omvang en mate van verstoring door genoemde milieufactoren. Daarnaast is gekeken naar de bezorgdheid over een aantal veiligheidsrisico's. De uitkomsten van het onderzoek hebben betrekking op de provincie Utrecht als geheel en op de vier onderscheiden regio's die de provincie opspannen.
Referentie: Peiling Milieuhinder: Utrecht 1997. J.E.F. van Dongen, H. Vos. TNO rapport 98.026.
- **Slaapverstoringsonderzoek rond Schiphol**
Het slaapverstoringsonderzoek rond Schiphol wordt uitgevoerd binnen het onderzoeksprogramma ‘Gezondheidskundige Evaluatie Schiphol?’. Dit programma wordt gecoördineerd door het RIVM, in opdracht van het Ministerie van VWS, VROM en Verkeer en Waterstaat. Het doel van het slaapverstoringsonderzoek is om informatie te verkrijgen over:
 - a) De relaties tussen blootstelling aan geluid van nachtelijk vliegverkeer enerzijds en mate van slaapverstoring, gezondheid en dagelijks functioneren anderzijds. Voor deze relaties wordt ook het effect van de periode in de nacht op deze relatie ook bekeken;
 - b) Wat nodig is om te schatten tot welke mate door geluid veroorzaakte effecten in de populatie voorkomen in de omgeving van Amsterdam Airport Schiphol.

Het onderzoek is opgesplitst in een pilot en een hoofdonderzoek. De pilot had als doel om het hoofdonderzoek voor te bereiden en goed op te zetten. Zo zijn er ook gegevens verzameld voor een poweranalyse ter voorbereiding van het hoofdonderzoek.

Referentie: Aircraft noise and sleep disturbance Pilotstudy. Passchier-Vermeer W, Vos H, Gils K van, Miedema HME, Eoo F de, Verhoeff EJ, 1999.

- Exposure and community response to noise in Europe
Deze case betreft een onderzoeksvoorstel voor het vijfde kaderprogramma van de Europese Unie. Het onderzoek heeft als doel om meer inzicht te verwerven in de response van de gemeenschap op omgevingsgeluid, dit zeker ook in het kader op de lopende ontwikkelingen van de Europese politiek t.a.v. geluid. Het is de bedoeling om de relatie tussen, enerzijds geluidsblootstelling door vliegtuig-, trein- en wegverkeer en anderzijds de door mensen zelfge-rapporteerde effecten van geluidshinder, slaapverstoring, communicatie interferentie, medische klachten en de algemene gezondheidsstatus te onderzoeken in de landen behorende tot de Europese Unie. Naast deze dosis-effect relatie gaat de interesse ook uit naar de prevalenties van het geluid en de effecten zoals genoemd.

Het laatste hoofdstuk bevat een literatuurverwijzingen met beschrijving van de literatuur om de lezer verder op weg te helpen met een verantwoord steekproefontwerp.

Dit rapport is geschreven in opdracht van de Sector Milieu van de Divisie Volksgezondheid van TNO Preventie en Gezondheid. Het doel van het rapport en het project “Verantwoord Steekproefontwerp bij Milieu” is om een handleiding te geven aan onderzoekers waarmee in de toekomst gewerkt kan worden om een goede steekproefopzet te bewerkstelligen.

Het rapport is tot stand gekomen door bestudering van de bijeengezochte literatuur en door (email-)contacten met enkele steekproefdeskundigen in Nederland. Tevens is nauw overleg geweest met de begeleidingscommissie die bestond uit: Dr. J. van Rijkevorschel, Dr. S. van Buuren, Prof. Dr. P. van der Heijden (allen van de Vakgroep Statistiek van TNO PG) en Dr. H. Miedema, Drs. R. de Jong en Drs. W. Passchier-Vermeer (van de Sector Milieu van TNO PG). Daarnaast wil ik hierbij ook Dr. S.A. Reijneveld bedanken voor het lezen en becommentariëren van het concept.

2 Verschillende typen steekproefontwerpen

In dit hoofdstuk zullen we ingaan op verschillende steekproefontwerpen. Het doel hierbij is om enerzijds een samenvattend overzicht te geven van de meest voorkomende ontwerpen zoals die beschreven zijn in de literatuur. Anderzijds moet de lezer, na het lezen van dit hoofdstuk in staat zijn om voor een specifiek onderzoeksprobleem te bepalen welk steekproefontwerp het meest geschikt is. Zo'n overzicht is natuurlijk nooit volledig en daarom zullen er altijd situaties blijven bestaan die niet binnen het raamwerk passen. Voor moeilijkere situaties kan dit hoofdstuk als een startpunt dienen.

We zullen, aan het eind van dit hoofdstuk, ook kort ingaan op het probleem van non-respons dat ongetwijfeld om de hoek komt kijken bij veel onderzoeken.

Voor een uitgebreide beschrijving van diverse steekproefontwerpen verwijzen we naar Bethlehem (1985), Muilwijk, Snijders & Moors (1992), Kish (1965), Cochran (1978), Thompson (1992) en Lehtonen & Pahkinen (1995). Een uitgebreidere beschrijving van deze (en meer) literatuur is te vinden in Hoofdstuk 5.

De begrippen die in paragraaf 2.1 aan de orde komen vormen de basis voor de rest van het rapport. In de daaropvolgende paragrafen (2.2 tot en met 2.9) worden diverse ontwerpen behandeld. Paragraaf 2.5 over gestratificeerde steekproeven is interessant voor onderzoeken waarbij uitspraken over bepaalde subgroepen van belang zijn. Voor uitspraken over de samenhang tussen sterkte van een determinant en mate waarin gevolgen optreden (zogenaamde dosis-effect relaties) zijn Paragraaf 2.3, 2.4 en ook 2.5 relevant.

2.1 Belangrijke basisbegrippen

Voordat we beginnen met de behandeling van standaard steekproefontwerpen zullen we eerst een aantal basisbegrippen de revue laten passeren die het kader vormen waarin over steekproefontwerpen gepraat wordt.

Een onderzoek beoogt (in de regel) het doen van kwantitatieve uitspraken over een geheel, de *populatie*. Populaties zijn verzamelingen van elementen waaraan je wilt gaan meten, zoals mensen, bedrijven, gemeenten, huishoudens, vliegtuigen etc. Het *steekproefkader* is een administratief overzicht van de fysieke populatie. Dit kan bijvoorbeeld een lijst zijn van alle elementen uit de populatie, een kaartenbak of een geografische kaart.

In principe hoeft er geen steekproefkader te zijn van de te onderzoeken populatie. Steekproefontwerpen gaan er echter vanuit dat er een steekproefkader aanwezig is, dus als dat niet het geval is dan moet er op de een of andere manier een steekproefkader verkregen worden of naar een uitweg gezocht worden, zie hiervoor paragraaf 2.10. Het is bij een steekproefkader van belang

dat elementen uit de fysieke populatie uniek identificeerbaar zijn met elementen uit het steekproefkader. Bijvoorbeeld bij verouderde steekproefkaders kan dit een probleem zijn wegens achterlopende administraties. Andere voorkomende fouten binnen steekproefkaders zijn: meer-voudige vermeldingen, vreemde elementen die niet uit de populatie komen, maar wel in het kader voorkomen, ontbrekende elementen etc. Door een niet goed identificeerbaar steekproefkader worden fouten geïntroduceerd in het onderzoek.

Een *steekproef* uit de populatie is een deelverzameling van de populatie. Dus elk element uit de steekproef moet ook in de populatie (en ook het steekproefkader) voorkomen. Een steekproef wordt uit de populatie “getrokken”, m.b.v. het steekproefkader, aan de hand van een van tevoren vastgelegd *ontwerp*. Een steekproefontwerp is een recept waarmee een steekproef genomen wordt uit de populatie. Bijvoorbeeld, als de waardering van bezoekers over de bloemententoonstelling van het huidige jaar in de Keukenhof het onderwerp van studie is dan is de populatie gelijk aan alle bezoekers van de Keukenhof in dat specifieke jaar. De steekproef is een deel van die populatie bezoekers. Als ontwerp van de steekproef nemen we bijvoorbeeld alle bezoekers waarvan het nummer van hun toegangskaartje eindigt op twee nullen. Dit type ontwerp wordt een systematische steekproef genoemd. Hierop komen we later op terug.

Met behulp van een steekproef uit de populatie worden uitspraken gedaan over de hele populatie. Voorbeelden van uitspraken zijn:

- De geluidshinder, in termen van het percentage ernstig gehinderden, in de grote steden van de Randstad bedraagt 10 % HA (highly annoyed people).
- Gemiddeld genomen besteedt de Nederlandse volwassen bevolking 1,5 % van haar tijd aan recreëren in een stiltegebied.
- De laatste jaren is de overlast van geluidshinder door landende en opstijgende vliegtuigen significant toegenomen.
- Hoger opgeleide individuen gebruiken vaker de trein dan lager opgeleiden.

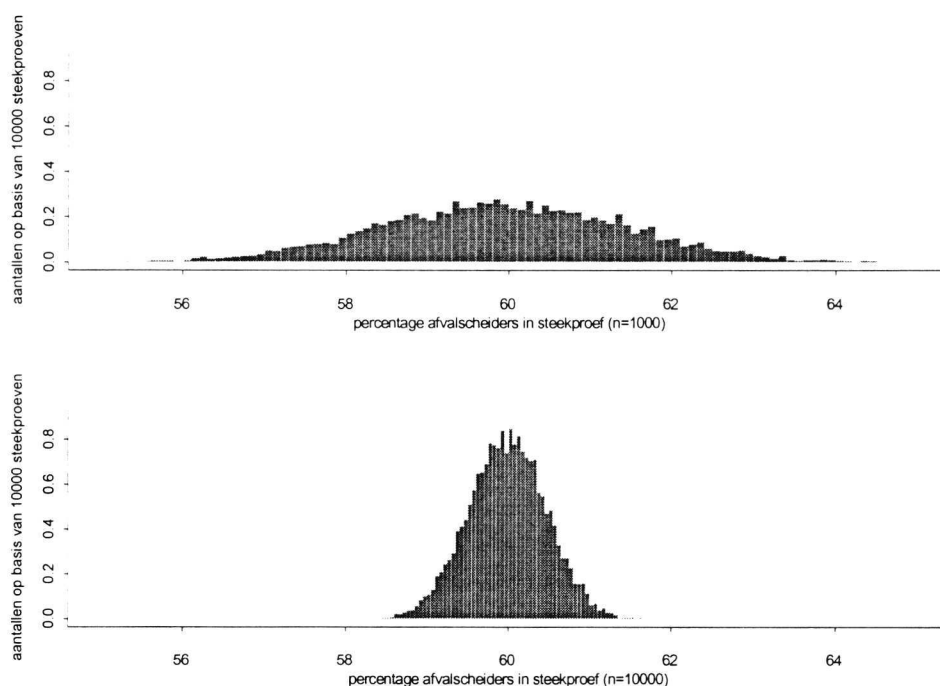
Indien deze uitspraken zijn gebaseerd op een steekproef, in plaats van de gehele onderzochte populatie, dan spreken we bij de eerste twee voorbeelden van *schattingen*. De conclusies van de twee laatste uitspraken zijn gebaseerd op *toetsen*. In feite zijn we bij een onderzoek geïnteresseerd in gegevens (grootheden of, in de statistiek vaak aangeduid met, parameters) die over de gehele populatie gaan. Omdat het tijdrovend en vaak niet mogelijk is om de gehele populatie te onderzoeken, voldoen we dikwijls met schattingen van deze parameters op basis van een steekproef. Zo is het *steekproefgemiddelde* een schatting van het *populatiegemiddelde*. En omdat het steekproefgemiddelde slechts een schatting is van het populatiegemiddelde, hoeven ze niet exact gelijk aan elkaar te zijn. *Berekende* gemiddelden, prevalenties, proporties en percentages op basis van de steekproef zijn allemaal schattingen van de respectievelijk populatiegemiddelden, populatieprevalenties, populatieproporties en populatiepercentages. Deze berekende grootheden, die

verkregen zijn via een steekproef, worden in de statistische literatuur aangeduid als *schatters* van de (populatie)parameters.

Het belangrijkste nadeel van een steekproef boven onderzoek van de gehele populatie is het optreden van steekproeffouten. Een *steekproeffout* ("sampling error") is het verschil tussen de werkelijke situatie (de populatieparameter) en de steekproefschatting dat veroorzaakt wordt door het feit dat de steekproef slechts een deel van de populatie uitmaakt. In de regel, bij een goed steekproefontwerp, zal een schatting van een bepaalde parameter *zuiver* zijn. Hierbij gaan we er vanuit dat er geen systematische doch slechts een toevallige fout aanwezig is. Een schatting van een parameter wordt zuiver genoemd als de *verwachte waarde* van de schatting overeen komt met de populatieparameter. Wat betekent dit nu eigenlijk? Laten we eens naar een voorbeeld kijken.

Voorbeeld [*Percentage huishoudens dat aan afvalscheiding meewerkt in de stad Utrecht*]:

In dit voorbeeld willen we de hierboven geïntroduceerde begrippen illustreren. In de stad Utrecht werkt 60 % van de huishoudens mee aan het scheiden van afval in "gewoon" afval en "groente-fruitafval". Laten we er even vanuit gaan dat Utrecht bestaat uit 400.000 huishoudens. Als we niet hadden geweten dat 60 % van de huishoudens aan scheiden van afval doet dan had je dat kunnen onderzoeken door het nemen van een steekproef. Bijvoorbeeld door het selecteren van een steekproef van 1000 huishoudens uit de 400.000 huishoudens. Er zijn enorm veel verschillende steekproeven (d.w.z. combinaties van 1000 huishoudens gekozen uit de 400.000 huishoudens) te vormen. In gedachten kunnen we alle mogelijke steekproeven beschouwen, maar ter illustratie beperken we ons tot 10.000 verschillende steekproeven. Per steekproef bepalen we



Figuur 1: 10.000 verschillende steekproeven waarbij het percentage afvalscheiders in Utrecht is bepaald. Boven: steekproeven ter grootte 1000, Onder: steekproeven ter grootte 10.000.

welk percentage huishoudens uit die steekproef aan afvalscheiding doet. Voor de ene steekproef zal dat 58 % zijn, maar voor een andere steekproef kan dat 63 % zijn. Elke steekproef levert een **schatting** van het percentage afval-scheiders in de populatie.

Figuur 1, het bovenste gedeelte, is een histogram van de 10.000 gevonden schattingen van het percentage afvalscheiders in Utrecht, waarbij de steekproefgrootte gelijk is aan 1000. Het blijkt dat de schattingen mooi gecentreerd zijn rond de 60 %, het percentage in de populatie. Het gemiddelde van de steekproefresultaten is gelijk aan de werkelijke waarde, namelijk 60 % en dan wordt de schatting **zuiver** genoemd.

In het bovenste plaatje varieert het percentage afval-scheiders van 55% tot bijna 65 %. Dit is de spreiding die geïntroduceerd wordt door de steekproef. Een maat voor de spreiding rond het gemiddelde is de "variantie". Hoe kleiner de variantie, des te dichter het steekproefpercentage zal liggen bij het populatiepercentage van 60 %. Hoe groter de steekproef, des te kleiner de variantie, en dan zullen de steekproefpercentages dichter rond 60% geconcentreerd zijn. Vergelijk hiervoor maar het boven en onder plaatje (boven: steekproef bestaat uit 1000 huishoudens, onder: steekproef bestaat uit 10.000 huishoudens.).

Tot nu toe hebben we het alleen gehad over bepaalde parameters van de populatie zoals gemiddelden, percentages (prevalenties) en proporties. Naast dit soort beschrijvende parameters kan men in onderzoeken ook geïnteresseerd zijn in relaties tussen verschillende doelvariabelen. Denk bijvoorbeeld aan uitspraken als:

- Hoe hoger de verstedelijkingsgraad, des te groter de lawaaioverlast in huis.
- Hoe groter het aantal overkomende vliegtuigen, des te groter het percentage bewoners met slaapverstoringproblemen rond Schiphol.

Dit type uitspraken wordt een *dosis-effect relatie* (in het engels: dose response relation) genoemd. Een dosis-effect relatie is een relatie tussen een verklarende variabele (bijvoorbeeld het aantal overkomende vliegtuigen) en een afhankelijke variabele (het percentage bewoners met slaapverstoringproblemen).

In het geval van dosis-effect relaties zal men een steekproefontwerp nodig hebben dat gebruik maakt van de kennis over de dosis-effect relatie en in het bijzonder van de onafhankelijke variabele (de dosisvariabele). Denk bijvoorbeeld aan de relatie tussen vliegtuiglawaai en het percentage gehinderden. Om een nauwkeurige dosis-effect relatie aan te tonen, is het zaak om ook genoeg waarnemingen te hebben bij gebieden met veel vliegtuiglawaai. En dit betekent dat je soms bewust niet elk deelgebied uit het onderzoeksgebied een gelijke kans zult geven om in de steekproef voor te komen. Gebieden (of personen) die blootgesteld zijn aan een hoge mate van vliegtuiglawaai zullen "oversampled" moeten worden. Kortom, een dosis-effect relatie vraagt om een speciaal type steekproefontwerp, zoals een systematische steekproef (paragraaf 2.3), een enkel-

voudige steekproef met ongelijke kansen (paragraaf 2.4) of een gestratificeerde steekproef (paragraaf 2.5).

Zoals gezegd introduceren steekproeven dus steekproeffouten. Naast steekproeffouten kunnen zich ook *niet-steekproeffouten* voordoen. Dit type fouten kan ook optreden bij onderzoek van de gehele populatie. Enkele voorbeelden van niet-steekproeffouten zijn: defecten van het steekproefkader, onjuiste manier van trekken van de steekproef, waarnemingsfouten, coderingsfouten bij invoer van de data, typefouten en non-respons. Bij *non-respons* komt de waarneming van een element uit de steekproef niet tot stand door bijvoorbeeld afwezigheid van de respondent, weigering van de respondent, meetinstrumenten die verkeerd (of niet) zijn ingesteld. In paragraaf 2.12 gaan we dieper in op non-respons binnen onderzoeken.

In de volgende paragrafen zullen, zoals gezegd, verschillende ontwerpen aan bod komen. Strikt genomen hoort bij elk ontwerp ook weer (eventueel) verschillende schatters van bijvoorbeeld het populatiegemiddelde en percentages. Dit komt doordat voor het ene ontwerp een schatter wel zuiver is, maar voor het andere ontwerp niet. En het is in de statistiek de gewoonte om zuivere schatters te nemen voor populatieparameters. Formules voor de “juiste”, dus zuivere, schatters in bepaalde situaties zullen we, uitzonderingen daargelaten, niet geven. Hiervoor verwijzen we naar de literatuur (Bethlehem (1985), Muilwijk, Snijders en Moors (1992), Kish (1995) en Cochran (1982)).

2.2 Enkelvoudige aselechte steekproefontwerp

Het eenvoudigste type steekproefontwerp is een enkelvoudige aselechte steekproef (EAS afgekort) of in het engels “*simple random sampling*” genaamd. Dit is een methode van selecteren van n eenheden (mensen, voorwerpen e.d.) uit een groep ter grootte N , waarbij elke combinatie van n elementen uit de populatie een even grote kans heeft om als steekproef genomen te worden. In de praktijk kun je zo’n steekproef trekken door de elementen uit de populatie te nummeren van 1 tot N en vervolgens *zonder* teruglegging n elementen achter elkaar te kiezen uit de getallen 1 tot en met N . Bij trekking van een element moet elk element dat voorkomt in de populatie een gelijke kans hebben om getrokken te worden. Het aselechte kiezen van een element uit de populatie gebeurde vroeger middels tabellen met aselechte getallen erin. Tegenwoordig kun je dit doen met behulp van een *Random Number Generator* in een statistisch pakket of bijvoorbeeld met de functie RNG in Excel. Volgens het Nederlands Normalisatie Instituut heet een trekkingsmethode aselechte als zij onafhankelijk is van alle eigenschappen van de elementen die voor het onderzoek van belang zijn.

Een kleine aanpassing op het enkelvoudige aselechte steekproefontwerp zonder teruglegging is het enkelvoudige steekproefontwerp *met* teruglegging. Zoals aangegeven in de naamgeving is het enige verschil het trekken van de elementen uit de populatie *met* in plaats van *zonder* teruglegging. Trekken zonder teruglegging is te prefereren boven trekken met teruglegging, maar is in de praktijk soms moeilijker uit te voeren. De reden deze voorkeur is dat bij trekken zonder terugleg-

ging, hoewel de schatters hetzelfde zijn, de variantie van de schatters kleiner is dan bij trekken met teruglegging. Dit betekent dat betrouwbaarheidsintervallen smaller zijn en je daardoor nauwkeurigere uitspraken kunt doen. Hoe kleiner de fractie n / N , oftewel de verhouding tussen de steekproefomvang en de populatieomvang, des te kleiner het verschil zal zijn tussen de twee steekproefontwerpen.

2.3 Het enkelvoudige systematisch steekproefontwerp

Het aselekt selecteren van elementen uit een populatie kan best nogal eens wat voeten in de aarde hebben. Denk bijvoorbeeld aan de volgende situatie: om te onderzoeken of recreanten van een zeker natuurgebied tevreden zijn over dit gebied wordt er een enquête onder de recreanten gehouden. Om een EAS te nemen, zoals in de vorige paragraaf beschreven, moet je van tevoren weten hoe groot de populatie van recreanten van het natuurgebied is en daarnaast moet je kunnen beschikken over een steekproefkader van alle recreanten waarin je heen en weer kan springen en elementen uit kan selecteren. Kortom een niet haalbare zaak. In zo'n situatie kun je overstappen op het enkelvoudig systematisch steekproefontwerp.

We gaan er nog steeds vanuit dat de populatiegrootte bekend is. Indien dat niet het geval is dan moet deze geschat worden. Hierop wordt teruggekomen in paragraaf 2.10. Een systematisch steekproefontwerp kenmerkt zich doordat met de keuze van een enkel aselekt getal gelijk de hele steekproef vastligt. Dit gaat als volgt in z'n werk: Als we uit een populatie van omvang N een steekproef willen trekken ter grootte n , bereken dan eerst de staplengte S . De staplengte is gedefinieerd als het gehele getal dat verkregen wordt door N / n naar beneden af te ronden. Kies vervolgens, willekeurig, een startgetal b uit de getallen 1 tot en met S . Selecteer nu het b -de element uit de populatie en elk daarop volgende S -de element, zodat de steekproef gelijk is aan $(b, b+S, b+2S, \dots)$, waarbij de getallen verwijzen naar de nummering in de populatie.

Voorbeeld [een systematische steekproef van recreanten in natuurgebied A]:

Stel je wilt een systematische steekproef trekken van 25 recreanten uit de populatie van alle recreanten die op zondag recreëren in natuurgebied A. Het blijkt dat er gemiddeld zo'n 150 recreanten op zondag naar natuurgebied A komen dus in dit geval is de staplengte S gelijk aan 6 ($= 150/25$). Vervolgens kiezen we willekeurig een startgetal tussen 1 en 6, bijvoorbeeld door het gooien met een dobbelsteen. Dit blijkt 5 te zijn. De steekproef zal nu bestaan uit de vijfde recreant die zondag het natuurgebied binnenkomt en elk zesde recreant die daarop volgt (dus nummer 11, 17, 23 et cetera).

Een speciaal type systematische steekproef is dat van een alfabetisch lexicografische steekproef, oftewel een alfabetische ordening van persoonsnamen. Hierbij wordt de alfabetisch geordende populatie genummerd zodat er vervolgens systematisch uit getrokken kan worden. Dit moet niet verward worden met een *lettersteekproef*, dat is een steekproef die tot stand komt door een trekking van namen door achtereenvolgens het trekken van de eerste letter, de tweede letter et cetera

van een naam. Een lettersteekproef is alleen toelaatbaar als er volstrekt geen verband bestaat tussen de alfabetische ordening en de waarden van de onderzoeksvariabelen. Een lettersteekproef is van het type geclusterde steekproef (zie paragraaf 2.6).

Toevallig hebben we in het bovenstaande voorbeeld de situatie waarbij de fractie N/n gelijk is aan een geheel getal. Als dat niet het geval is dan heeft elk element uit de steekproef strikt genomen niet dezelfde kans om in de steekproef opgenomen te worden. Op de bovenstaande procedure zijn daarom aanpassingen gemaakt om aan deze conditie wel te voldoen (zie Bethlehem (1985)). Dit gaat als volgt:

Procedure: Het trekken van een systematische steekproef

Benodigheden: populatieomvang N (of een schatting daarvan), steekproefomvang n , een willekeurig getal uit het interval $(0,1]$.

Stappen 1. Bereken staplengte $F=N/n$.

2. Trek het getal u (willekeurig dus uniform) uit $(0,1]$. Neem als startgetal $b=u \cdot F$.

3. Bereken de waarden $t_i = b + (i-1) \cdot F$, voor $i=1,2,\dots,n$.

4. Bepaal bij iedere waarde t_i het gehele getal k_i met $k_i-1 < t_i \leq k_i$. Selecteer k_i .

Als er geen verband is tussen de onderzoeksvariabele en de rangorde van de elementen in het steekproefkader dan zijn de elementen van de steekproef, bij benadering, aselekt gekozen.

Naast het feit dat een systematische steekproef makkelijker is om uit te voeren zijn er situaties waarvoor een systematische steekproef te prefereren is boven een EAS. Dit zijn:

- Bij een lineaire trend: wanneer de waarde van de onderzoeksvariabele(n) afhangt van de rangnummers van de populatie-elementen, worden middels een systematische steekproef schattingen verkregen met een kleinere variantie dan bij een EAS. Bij een dosis-effect relatie, waar je een steekproefkader hebt, of kan worden aangebracht, met een rangordening die afhangt van de onafhankelijke variabele, is een systematische steekproef dus een goede keuze.
- Bij periodieke fluctuaties: Als de periode van de fluctuatie in de onderzoeksvariabele gelijk is aan de staplengte die gebruikt wordt bij een systematische steekproef dan introduceer je een vertekening van de steekproef. Aan de andere kant, als de periode niet gelijk is aan de staplengte (en ook geen veelvoud of deler ervan) dan geeft een systematische steekproef juist schatters met een kleinere variantie t.o.v. een EAS.
- Impliciete stratificatie: Het komt nog wel eens voor dat groepen overeenkomende elementen in de populatie in het steekproefkader achtereenvolgende rangnummers hebben. Denk bij-

voorbeeld aan gemeentelijke woningregisters waar de adressen alfabetisch en de nummers oplopend gerangschikt zijn. Indien de groepen ongeveer gelijk van grootte zijn dan vormen groepjes opeenvolgende steekproefelementen zogenoemde strata in de steekproef. Per stratum zijn de elementen op bepaalde aspecten homogener en daarmee krijg je schatters met een kleinere variatie t.o.v. die uit een EAS. Op stratificatie komen we in paragraaf 2.6 terug.

Het nadeel van een systematische steekproef t.o.v. een EAS is dat de variantie van bijvoorbeeld het populatiegemiddelde niet zuiver te schatten is. De variantie van het populatiegemiddelde is een maat voor de spreiding van de populatie rond het gemiddelde. Wanneer er geen reden is om aan te nemen dat de onderzoeksvariabele gerelateerd is aan de rangorde in het steekproefkader, wordt de variantie geschat door net te doen alsof de gegevens d.m.v. een EAS verkregen zijn.

2.4 Enkelvoudige steekproefontwerp met ongelijke kansen

Op het eerste gezicht lijkt het voor de hand te liggen om een steekproef zodanig vast te stellen zodat elk element uit de populatie gelijke kans heeft om in de steekproef voor te komen. De hiervoor behandelde steekproefontwerpen waren zo ook ingericht. Maar deze strategie hoeft helemaal niet optimaal voor een onderzoek te zijn. Er zijn tal van situaties te bedenken waarbij een “oneerlijke” steekproef, een steekproef met ongelijke kansen, preciezere schattingen oplevert, of minder groot hoeft te zijn om dezelfde nauwkeurigheid te bereiken. Preciezere schattingen betekenen o.a. smallere betrouwbaarheidsintervallen.

Stel je wilt bijvoorbeeld het gemiddelde inkomen van de Leidse bevolking bepalen en je weet de leeftijd van elk individu in de populatie. Leeftijd hangt samen met inkomen en daarom is de variabele leeftijd geschikt om als hulpvariabele te gebruiken om de steekproef samen te stellen. Als het inkomen van een individu perfect voorspeld wordt uit zijn leeftijd en voor de hele populatie de leeftijd bekend is dan hebben we dus ook een zeer precieze schatting van het gemiddelde inkomen van de Leidse bevolking te pakken. In de praktijk is de relatie tussen leeftijd en inkomen niet één op één en daarom zal er altijd nog onzekerheid over blijven.

Hoe gaat het trekken van elementen met ongelijke kansen nu te werk? Stel je hebt de beschikking over een relevante hulpvariabele waarvan de waarde bekend is voor elk element uit de populatie. Laten we de waarden noteren met $X_1, X_2, \dots, X_N$. Dus X_1 is bijvoorbeeld de leeftijd van het eerste element uit de populatie. Bereken $X = X_1 + X_2 + \dots + X_N$. De trekkingskans van het k -de element uit de populatie om in de steekproef voor te komen wordt evenredig gekozen aan X_k , en is gelijk aan $p_k = X_k / X$. Dit betekent dat elementen met een hogere waarde voor de hulpvariabele een grotere kans hebben om in de steekproef te komen. Met een eenvoudig computerprogramma is het mogelijk de steekproef uit het kader te trekken. De formule van de schatter voor het populatiegemiddelde (bijvoorbeeld het gemiddelde inkomen van de Leidse bevolking) is gelijk aan

$$y_{HT} = \frac{X}{nN} \sum_{i=1}^n y_i / x_i.$$

De variabelen x_i en y_i zijn de getrokken waarden die in de steekproef zitten. De afkorting HT staat voor Horvitz-Thompson schatter. Deze schatter speelt een belangrijke rol in de survey-achtige (onderzoeks)wereld. Intuïtief gezien is het vreemd dat, bijvoorbeeld in het geval van de inkomens, hogere inkomens een grotere kans hebben om in de steekproef voor te komen. Door echter in formule voor het steekproefgemiddelde omgekeerde gewichten mee te nemen als waarmee de steekproef getrokken is, wordt op zo'n manier toch een zuivere schatting verkregen van het populatiegemiddelde! Daarnaast is de variantie van de schatter **kleiner** dan (of gelijk aan) bij een EAS en dat is het belangrijkste voordeel van elke HT-schatter. En daardoor worden betrouwbaarheidsintervallen smaller bij dit ontwerp. Dit is dan ook een belangrijk voordeel van het steekproefontwerp met ongelijke kansen t.o.v. een EAS. Door de beschikbare informatie over de populatie op een juiste manier te gebruiken worden nauwkeurigere schattingen verkregen. Mutatis Mutandis, kan men met een kleinere steekproef volstaan.

De winst van een ontwerp met ongelijke kansen ten opzichte van een EAS kan geïllustreerd worden met behulp van de volgende simulatie. Stel Y is een onderzoeksvariabele die in de populatie een gemiddelde heeft van 10. X is een hulpvariabele, die hoog correleert met Y (correlatiecoëfficiënt ρ). Met behulp van een random number generator trekken we 10.000 verschillende waarden voor X en Y . Dit is de simulatiepopulatie. Vervolgens trekken we 500 waarnemingen voor Y uit de gesimuleerde populatie met ongelijke kansen gebaseerd op de X 'en. Omdat we enerzijds de populatie kennen en anderzijds een steekproef eruit getrokken hebben kunnen we nu het populatiegemiddelde μ , het steekproefgemiddelde y_{HT} (met behulp van de formule op de vorige pagina) en de schatter s voor de standaardafwijking van het steekproefgemiddelde uitrekenen. Daarnaast trekken we ook een EAS van 500 elementen uit dezelfde populatie. Hiervoor berekenen we ook het gemiddelde μ_{EAS} en de geschatte standaarddeviatie van het gemiddelde s_{EAS} .

We voeren deze berekening uit voor verschillende gesimuleerde datasets die variëren in correlatie tussen Y en X . Uit de tabel kunnen we concluderen dat voor een correlatie van hoger dan 0,7 het inderdaad loont om met ongelijke kansen te trekken. De variantie neemt aanzienlijk af. Bijvoorbeeld in het geval van een correlatie van 0,9 kun je met een vier keer zo kleine steekproef voldoen om dezelfde standaardafwijking in het gemiddelde te hebben.

Tabel 1: Vergelijking van een ontwerp met ongelijke kansen en een EAS. In de kolommen staat respectievelijk de correlatie tussen X en Y , het populatiegemiddelde, het geschatte gemiddelde met de ongelijke kansen steekproef, de standaardafwijking van het geschatte gemiddelde, en vervolgens het geschatte gemiddelde en standaardafwijking van de EAS. In de laatste kolom is de winst zichtbaar van het ongelijke kansen ontwerp t.o.v. de EAS.

ρ	μ	y_{HT}	s	μ_{EAS}	s_{EAS}	s_{EAS} / s
0,7	10,00	9,90	0,047	10,03	0,043	0,91
0,8	10,00	9,93	0,0236	10,01	0,044	1,86
0,9	10,01	10,00	0,0108	10,06	0,043	3,98
0,99	9,99	9,99	0,00090	9,97	0,044	47,8

De keuze van de gewichten p_k als trekkingkansen van de elementen uit de populatie geeft bijna een kant en klaar recept voor het steekproefontwerp. Er zijn enkele trekkingschema's bekend voor steekproeven (met teruglegging) met ongelijke kansen. Deze schema's gaan ervan uit dat de hulpvariabelen zodanig gedefinieerd zijn dat ze geheel en positief zijn. In Bethlehem (1985) wordt onder meer de cumulatieve methode beschreven. Tegenwoordig is het met de beschikbare geavanceerde statistische programmatuur (bijvoorbeeld S-PLUS) heel eenvoudig om een steekproef met ongelijke kansen te trekken uit een populatie.

Een steekproefontwerp met ongelijke kansen ligt ook voor de hand bij onderzoek naar een dosis-effect relatie. In zo'n geval is het logisch om per "dosering" (de onafhankelijke variabele) evenredig veel waarnemingen te doen. Dit betekent dat er ook voldoende waarnemingen genomen worden in het gebied van een hoge (of lage) "dosis". Elementen uit de populatie komen dan met gewogen kansen in de steekproef voor, waarbij de kansen evenredig zijn met de dosis (de onafhankelijke variabele, die dus dienst doet als hulpvariabele). Van tevoren moet je dus weten, voor de elementen uit de populatie, wat de waarde van de onafhankelijke variabele is.

Een andere situatie waarbij een steekproefontwerp met ongelijke kansen om de hoek komt kijken is het gestratificeerd steekproefontwerp. Dit is een veel voorkomend steekproefontwerp en daarom besteden we hieraan in de volgende paragraaf meer aandacht. De moraal van het steekproefontwerp met ongelijke kansen is dat door alle beschikbare kennis over de populatie op een goede manier te betrekken bij het ontwerpen van de steekproef je kunt komen tot preciezere schattingen van de "werkelijke" situatie.

Naast een (enkelvoudig) steekproef met ongelijke kansen en een systematische steekproef komt ook een mengeling hiervan voor: een systematische steekproef met ongelijke kansen. Dit ontwerp combineert de voordelen van beiden. Wederom levert het schatten van de variantie problemen op. Voor schema's waarmee een systematische steekproef met ongelijke kansen verkregen kunnen worden verwijzen we naar Bethlehem (1985).

2.5 Stratificeren

Het opdelen van de populatie in homogene groepen wordt stratificeren genoemd. Een deelpopulatie noemt men een stratum. Het trekken van een steekproef waarbij de populatie gestratificeerd is, gaat in z'n werk door per stratum een enkelvoudige steekproef te trekken. Een steekproef met strata erin wordt een gestratificeerde steekproef genoemd. Stratificatie wordt veel toegepast in onderzoek. Een belangrijk kenmerk van stratificatie is dat je de populatie opdeelt in strata met behulp van een bepaald criterium en vervolgens uit *elk* stratum een steekproef neemt. Een reden hiervoor kan bestaan als je per stratum een schatting wilt kunnen geven met een bepaalde nauwkeurigheid. Laten we naar een voorbeeld gaan kijken waar verschillende stratificaties mogelijk zijn.

Stel je wilt onderzoek doen naar geluidshinder door nachtelijke vliegbewegingen, onder inwoners van Zwanenburg. Mensen verschillen van elkaar en daarom zal de mate van ondervonden geluidshinder tijdens de nacht per persoon ook sterk verschillen. Het verschil in ondervonden geluidshinder is te wijten aan een scala van factoren. Sommige mensen slapen makkelijk en hebben helemaal geen last van de vliegtuigen. Anderen hebben wel last van de vliegtuigen maar werken zelf ook op Schiphol en zijn dus economisch afhankelijk van de vliegbewegingen. Ze zullen daarom minder snel toegeven dat ze last hebben van geluidsoverlast door vliegtuigen. Aan de andere kant zijn er ook milieuactivisten die sowieso tegen elke (nachtelijke) vliegbeweging zijn en dus snel geluidshinder ondervinden. Kortom de dosis-effect relatie tussen ervaren geluidshinder bij mensen en het geluidsniveau dat bij mensen gemeten wordt zal variëren door diverse factoren.

In zo'n geval is het interessant om de populatie op te splitsen in bepaalde categorieën waarvan men vermoedt dat er een behoorlijke correlatie bestaat tussen het behoren tot die categorie en de meetuitkomst. In het geval van de geluidshinderkwestie in Zwanenburg kan men een opsplitsing maken in de Zwanenburgse huishoudens naar bijvoorbeeld economische afhankelijkheid van Schiphol (wel of niet economisch afhankelijk). Door het opsplitsen van de huishoudens van Zwanenburg naar wel of niet economische afhankelijkheid krijg je twee deelpopulaties, de zogenoemde strata, die elk, naar verwachting, homogenere populaties vormen ten aanzien van ervaren geluidshinder. Door opsplitsing in strata kun je enerzijds uitspraken doen over de strata apart. Bijvoorbeeld: "bij economisch gebonden huishoudens is het percentage geluidsgehinderde huishoudens gelijk aan 10 %". Anderzijds kun je door het combineren van de uitkomsten uit de homogene strata tot nauwkeurigere uitspraken komen over de totale populatie dan met een eenvoudige aselechte steekproef van de populatie zonder de opsplitsing. Dit komt doordat er voordeel getrokken wordt uit het feit dat de elementen uit hetzelfde stratum op elkaar lijken. In het extreme geval, als per stratum alle elementen identiek zijn, is één element uit elke deelpopulatie voldoende om de gehele populatie te beschrijven. In dit geval is het populatiegemiddelde dus exact te bepalen omdat uit elk stratum een element waargenomen is. In de praktijk zal zich zo'n situatie niet voordoen, maar het gaat eerder om het onderliggende principe in dit voorbeeld: naarmate de homogeniteit van de strata toeneemt, des te nauwkeuriger kun je met dezelfde steekproefomvang de populatieparameters schatten.

Een ander voorbeeld is een onderzoek naar de prevalentie van roken per stadsdeel in Amsterdam voor meerdere leeftijdscategorieën. Door a priori te stratificeren op stadsdeel en leeftijdscategorie levert het onderzoek nauwkeurigere schattingen op van de gevraagde prevalenties dan met een EAS uit de hele bevolking van Amsterdam tegelijkertijd verkregen zouden zijn

In het voorbeeld over Zwanenburg was er sprake van een onderverdeling van de populatie van huishoudens op grond van economische afhankelijkheid. Als we nu naar geluidshinder in drie verschillende gemeenten (bijvoorbeeld Zwanenburg, Amsterdam Bijlmer en Rijsenhout) zouden kijken dan kun je stratificeren naar gemeente en vervolgens binnen een stratum van een gemeente kun je eventueel weer een onderverdeling naar economische afhankelijkheid maken. Op deze manier is de steekproef in twee stappen verdeeld in een aantal verschillende strata. Er zitten als

het ware twee lagen in de steekproef. De reden om meerdere lagen aan te brengen in dit geval is omdat je verwacht dat er verschillen bestaan in de relatie tussen geluid en hinder tussen de gemeenten, bijvoorbeeld omdat de bevolkingsopbouw per gemeente verschilt. Het hangt helemaal van de onderzoeksvraag af hoeveel lagen je in een steekproef moet bouwen en daarnaast natuurlijk ook van de beschikbare informatie. Bij het ontwerpen van een steekproef is het belangrijk om steeds de volgende vraag te stellen: “lijken alle elementen uit een (sub)populatie op elkaar of is er een nuttige stratificatie aan te brengen zodat de elementen per stratum homogener zijn ten aanzien van de onderzoeksvraag?”.

Bij onderzoek naar dosis-effect relaties wordt vaak ook een gestratificeerd ontwerp gebruikt. Door de populatie te verdelen in deelpopulaties waarvoor de (verwachte) dosis zo gelijk mogelijk is wordt er gegarandeerd dat er voor alle niveaus van de dosis elementen in de steekproef zitten. Voor een dosisvariabele met een continue schaal kun je strata maken door de dosisvariabele te verdelen in een aantal categorieën en vervolgens te stratificeren volgens de categorieklassen. Zoals eerder gezegd zijn er een aantal goede alternatieven voor een ontwerp bij dosis-effect relaties. Thompson (1992, pagina 87) adviseert om een gestratificeerd ontwerp te nemen waarbij de stratificatie is gebaseerd op de dosisvariabele en waar de verdeling van de steekproefaantallen per stratum evenredig is met de (verwachte) spreiding van de expositievariabele binnen in een stratum.

Bij het gestratificeerde steekproefontwerp wordt uit *elk* stratum een (deel)steekproef getrokken. De manier waarop de deelsteekproef getrokken wordt hoeft niet enkelvoudig aselekt te zijn. Men kan ook per stratum een systematische steekproefontwerp of een ontwerp met ongelijke kansen toepassen. Sterker nog, men kan het ontwerp per deelsteekproef variëren. Dit laatste brengt wel complicaties met zich mee, dat men goed in de gaten moet houden wat in zo'n geval de juiste formules zijn voor zuivere schattingen van de populatieparameters.

2.6 Geclusterde steekproef

Om een steekproef te kunnen nemen uit de populatie van alle Nederlandse huishoudens is het nodig om een steekproefkader van deze huishoudens ter beschikking te hebben. Dit bestaat echter niet. Een steekproefkader van huishoudens is alleen beschikbaar op gemeenteniveau. In zo'n geval kun je dus niet zonder meer een enkelvoudig aselekte steekproef nemen omdat het steekproefkader ontbreekt. Je zou nu als volgt te werk kunnen gaan:

1. Trek een steekproef van gemeenten.
2. Trek een steekproef van postcodegebieden uit elke gekozen gemeente.
3. Trek een steekproef van huizenblokken uit elk gekozen postcodegebied.
4. Betrek alle huishoudens uit het gekozen huizenblok in het onderzoek.

Lijsten van gemeenten en postcodegebieden zijn gemakkelijk te verkrijgen. Vanaf dat punt kan de steekproeftrekking indien nodig bijvoorbeeld gebaseerd worden op een lokale kaart. Er is dus geen landelijke lijst van huishoudens nodig. Bovendien komen de geselecteerde huishoudens voor in een huizenblok en kunnen dus gemakkelijk door één interviewer ondervraagd worden. Dit reduceert ook de kosten van het onderzoek.

Het laatste voorbeeld is een voorbeeld van een geclusterde of getrapte steekproef. Bij een geclusterde steekproef worden er, middels een afgesproken criterium, clusters gevormd. Vervolgens worden er aselekt (volgens de regels van een eenvoudige aselekte steekproef) clusters gekozen uit alle aanwezige clusters. Alle elementen uit de gekozen clusters worden in de steekproef meegenomen.

Een geclusterde steekproef lijkt veel op een gestratificeerde steekproef. Het verschil is dat bij een gestratificeerde steekproef er uit *alle* strata een enkelvoudige steekproef getrokken wordt en bij een geclusterde steekproef wordt er op het niveau van de clusters een enkelvoudige steekproef getrokken en vervolgens worden *alle* elementen uit de getrokken clusters meegenomen. Een geclusterde steekproef kan dus eigenlijk worden opgevat als een EAS op geaggregeerd niveau doordat in plaats van elementen uit de populatie clusters worden getrokken. Het procédé van de geclusterde steekproef kan gegeneraliseerd worden door clusters binnen clusters te definiëren. Dit hadden we ook al gedaan in het voorbeeld van de gemeenten en de postcodegebieden.

De voordelen van een geclusterde steekproef zijn: eenvoudiger te realiseren doordat er bijvoorbeeld geen steekproefkader van de hele populatie aanwezig hoeft te zijn, maar een steekproefkader van de clusters en een steekproefkader binnen de clusters. Daarnaast is een geclusterde steekproef vaak goedkoper dan een “gewone” steekproef omdat, bijvoorbeeld, in het geval van enquêtes de reiskosten van de enquêteurs aanzienlijk verminderd worden. Bij een EAS van de Nederlandse bevolking zullen de te bezoeken adressen verspreid zijn over het hele land. Als er gemeenten als clusters gekozen zijn en we onderzoeken enkele gemeenten in het geheel dan zal het aantal te bezoeken adressen binnen een klein gebied geconcentreerd liggen.

Een geclusterde steekproef lijkt uit praktisch oogpunt dus aantrekkelijk te zijn. Echter, door de selectie van een cluster worden tegelijk een aantal min of meer op elkaar gelijkende elementen geselecteerd. Hierdoor levert zo'n steekproef minder informatie op dan een even grote enkelvoudige steekproef. In het zeer extreme geval zijn namelijk alle elementen uit een cluster gelijk. Het onderzoeken van slechts één element per cluster levert dan evenveel informatie op als het onderzoeken van alle elementen uit het cluster. Dit verlies aan informatie wordt het *clustereffect* genoemd. Om evenveel informatie uit een geclusterde steekproef te krijgen als uit een enkelvoudige steekproef moet je dus een grotere steekproef nemen (wat neer komt op meer clusters). We komen hierop terug in het volgende hoofdstuk.

De precisie van de schatters bij een geclusterde steekproef hangt af van het aantal elementen in een cluster en van de grootte van het clustereffect. De variantie zal klein zijn als zowel de aantallen elementen per cluster als ook de clustergemiddelden van de clusters elkaar niet veel ontle-

pen. Als vooral de clusters nogal in omvang verschillen en de clustergemiddelden elkaar niet zoveel ontlopen dan kan de invloed van de omvang van de clusters op de variantie worden verminderd door de clusters niet enkelvoudig aselekt te kiezen, maar door de trekkingskansen van de clusters te laten afhangen van de omvang van de clusters. De trekkingskansen van cluster i wordt dan evenredig gekozen aan de omvang van de clusters, dus met kans $p_i = \text{omvang cluster } i / \text{totale omvang populatie}$.

2.7 Tweetrapssteekproefontwerp

Het geclusterde steekproefontwerp is dus qua prestatie niet zo erg goed. Hoe groter het clustereffect is, des te meer waarnemingen (dus clusters) er nodig zijn voor een gelijkblijvende precisie. In de praktijk wordt een geclusterde steekproef dan ook niet zonder meer toegepast. Aan een, nogal op de geclusterde steekproef lijkend, ontwerp, namelijk het tweetrapssteekproefontwerp wordt de voorkeur gegeven.

Bij een tweetrapssteekproef worden er eerst clusters gevormd, de zogenaamde *primaire eenheden*. De elementen uit de primaire eenheden worden de *secundaire eenheden* genoemd. Vervolgens worden er eerst primaire eenheden geselecteerd en uit de geselecteerde primaire eenheden worden er secundaire eenheden geselecteerd. In het voorbeeld van de steekproef onder Nederlandse huishoudens verandert het geclusterde steekproefontwerp in een tweetrapssteekproefontwerp door de laatste stap te vervangen door “trek een steekproef van huishoudens uit elk gekozen huizenblok”. In dit geval is er zelfs sprake van een viertrapssteekproef.

Voor de onderzoeker geeft dit ontwerp veel vrijheid. De primaire eenheden kunnen met gelijke of ongelijke kansen gekozen worden, net zoals de secundaire eenheden binnen de primaire eenheden. Daarnaast kan de onderzoeker beslissen of hij weinig primaire eenheden wil en veel secundaire eenheden binnen de primaire eenheden of andersom, dus veel primaire eenheden maar weinig secundaire eenheden. Veel primaire eenheden en weinig secundaire eenheden geeft meestal hogere kosten voor het onderzoek maar ook preciezere schattingen van de populatie parameters. Net zoals bij het geclusterde steekproefontwerp zal het vaak effectiever zijn om de primaire eenheden te selecteren met kansen die evenredig zijn aan de omvang van de eenheden. Dit laatste levert ook het voordeel op dat de steekproefomvang niet meer afhangt van de getrokken primaire eenheden en dus kan de steekproefomvang beter van tevoren gepland worden.

2.8 Verschillen tussen tweetrapssteekproeven en stratificatie

Voor een gestratificeerde steekproefontwerp en een tweetrapssteekproefontwerp geldt dat de totale populatie in groepen wordt verdeeld middels een bepaald criterium. Het verschil tussen de twee ontwerpen is dat bij een gestratificeerde steekproef uit alle groepen (strata) een enkelvoudige (aselecte, symmetrische of met ongelijke kansen) steekproef wordt getrokken, bij een tweetrapssteekproef wordt eerst een enkelvoudige steekproef getrokken uit de groepen en vervolgens

wordt per groep een enkelvoudige steekproef genomen uit de elementen van de groep. Analyses van gegevens verzameld m.b.v. tweetrapssteekproeven en het daarop lijkende geclusterde steekproef zijn fundamenteel anders dan die van een gestratificeerde steekproef. Maar een eerste belangrijke stap in de richting van een goede analyse van de gegevens is al gezet als de structuur van de steekproef helder is.

Door de selecties (clusteringen) die plaatsvinden bij het nemen van een tweetrapssteekproef kun je spreken van verschillende niveaus van de steekproef. In het voorbeeld van de Nederlandse huishoudens is er eerst een steekproef genomen uit de populatie van Nederlandse gemeenten, vervolgens op het niveau van postcodegebieden, huizenblokken binnen een postcodegebied en als laatste huishoudens binnen een huizenblok met dezelfde postcode en uit dezelfde gemeente. Stel je hebt aan de geselecteerde huishoudens gevraagd om bij te houden hoeveel afval ze produceren per week. Omdat er een aselechte steekproef getrokken is uit de Nederlandse gemeenten kun je met de getrapte steekproef nu ook uitspraken doen op het geaggregeerde niveau van gemeenten. Je zou bijvoorbeeld kunnen kijken naar de hoeveelheid geproduceerd afval per gemeente, door de geregistreerde hoeveelheden per huishouden en per gemeente te aggregeren. Omdat de steekproef van de gemeenten aselekt is gekozen kun je de uitkomsten van de steekproef **generaliseren** naar de populatie van Nederlandse gemeenten, bijvoorbeeld door een betrouwbaarheidsinterval uit te rekenen van het gemiddeld geproduceerd afval per gemeente. In het geval van een gestratificeerde steekproef waarbij bijvoorbeeld alle gemeenten in Zuid-Holland zijn meegenomen kun je uitkomsten uit een onderzoek niet generaliseren naar heel Nederland. De keuze van gemeenten is in zo'n geval niet willekeurig maar juist heel systematisch: alle gemeenten uit Zuid-Holland zijn meegenomen. Het zijn dan alleen de Zuid-Hollandse gemeenten waarover je een uitspraak kunt doen.

Een andere reden om een clustering aan te brengen is wegens het (vrij vertaald) “door de bomen het bos niet meer zien” effect. Net als elementen binnen een cluster variëren zullen clusters onderling ook variëren, maar dan op geaggregeerd niveau. De tussen-cluster-spreiding en de binnen-cluster-spreiding zijn twee verschillende typen spreidingen die uitgesplitst en dus geïdentificeerd worden door de, in de steekproef, bewust geïntroduceerde clustering. Als er in een populatie een natuurlijke clustering aanwezig is, die van invloed is op de onderzoeksvariabele(n), en je houdt geen rekening met deze clustering in de steekproef dan zullen de spreidingen tussen de natuurlijk clusters en de spreiding binnen de clusters niet meer uit elkaar te halen zijn. Het gevolg is dat je alleen de totale spreiding ziet van de steekproef en die is dan zo groot dat je weinig meer kan zeggen over de onderzoeksvariabele(n). Dit effect komt naar voren bij bijvoorbeeld dosis-effect relaties die over de totale populatie niet echt zichtbaar zijn, maar als er naar deze dosis-effect relatie wordt gekeken per cluster (eenheid) er wel duidelijk sprake is van een verband.

Een voorbeeld hiervan kwam naar voren tijdens het pilotonderzoek naar geluidshinder bij Schiphol. Om de dosis-effect relatie te onderzoeken tussen geluid en hinder heeft men gedurende een maand lang bij een aantal geselecteerde mensen elke nacht enerzijds het geluidsniveau gemeten en anderzijds een aantal parameters waaraan af te leiden is of mensen last hebben gehad van het

geluid (bijvoorbeeld het aantal keer wakker worden in een nacht). De steekproef bestaat in dit geval dus uit een aantal (ongeveer 30) metingen per persoon van het geluid en de bijbehorende parameters van de slaapkwaliteit. De personen worden nu dus gezien als primaire eenheden, de clusters. Uit de totale populatie van personen die wonen rond Schiphol is een aselechte steekproef genomen en vervolgens is er bij de personen uit de steekproef een aantal dagen gemeten. Als je nu een plaatje zou maken met op de horizontale as het gemeten geluidsniveau en op de verticale as de slaapkwaliteit van alle personen en alle nachten bij elkaar dan krijg je een enorme puntenwolk, zonder enig zichtbare relatie. Er is echter reden om aan te nemen dat er een dosis-effect relatie bestaat tussen geluidsniveau en ervaren hinder, maar doordat mensen nu eenmaal verschillen van elkaar en daarmee ook verschillen vertonen op het gebied van slapen zie je als het ware door de bomen het bos niet meer. Je moet daarom corrigeren voor de niveauverschillen tussen de personen.

Het steekproefontwerp van het slaapverstoringsonderzoek met herhaalde metingen per persoon is misschien een gek soort getrapte steekproef. De populatie waar de steekproefelementen uit getrokken worden is de tijd. Toch wordt binnen de statistiek een herhaalde metingen steekproefopzet gezien als een getrapte steekproef. Belangrijk hierbij is wel dat je bij de analyse van de gegevens er óf zeker van moet zijn dat de herhaalde metingen onafhankelijk van elkaar zijn, óf je moet de correlaties en tijdseffecten tussen de herhaalde metingen meenemen in de analyse.

Een voorbeeld van een getrapte steekproef bij een dosis-effect relatie betreft een steekproef van scholieren om de relatie aan te tonen tussen IQ van Vwo-scholieren en hun score, bijvoorbeeld, op het eindexamen bij Nederlands. Als je een steekproef wilt nemen uit de populatie van scholieren ligt het voor de hand om eerst scholen aselekt te selecteren uit alle 6-VWO klassen van Nederland, vervolgens enkele klassen aselekt per school te kiezen en uiteindelijk leerlingen aselekt te selecteren uit de gekozen klassen. Zo'n getrapte steekproef van leerlingen is veel eenvoudiger te realiseren dan een eenvoudige aselechte steekproef van leerlingen uit de populatie van alle 6-VWO leerlingen in Nederland. Maar naast het feit dat een getrapte steekproef in deze situatie eenvoudiger te trekken is, is ze, in statistisch opzicht, ook te prefereren boven een eenvoudige aselechte steekproef. Laten we er even vanuit gaan dat elke klas een andere leraar (of lerares) heeft. In eerste instantie zou je zeggen dat het IQ van leerlingen verschilt en daarom ook de score op het eindexamen Nederlands. Echter, het zou natuurlijk ook kunnen dat door de manier van lesgeven van een leraar de klas in het geheel beter dan gemiddeld scoort op het eindexamen. Dit is dan niet te danken aan het IQ van de leerlingen, maar komt ook door de leraar. De leraar verschilt per geselecteerd cluster en dus hebben we hier te maken met een andere relatie per cluster doordat er een niveauverschil is tussen de leraren! Zo zou je ook nog kunnen redeneren op het niveau van de school. Als je de gemiddelde score voor Nederlands op het eindexamen per school uitrekent dan kunnen die gemiddelde scores verschillen van elkaar doordat er bijvoorbeeld per school een andere lesmethode is gebruikt. Kortom, in dit voorbeeld komt naar voren dat enerzijds een geclusterde steekproef makkelijk is en daarom te prefereren is, maar anderzijds ook *noodzakelijk* is om een goede dosis-effect relatie aan te tonen tussen IQ en score door te corrigeren voor verschillen tussen de scholen en verschillen tussen de klassen.

In het scholierenvoorbeeld was het dus noodzakelijk om een getrapte steekproef te trekken omdat de spreiding tussen de leerlingen niet alleen veroorzaakt wordt doordat de leerlingen een verschillend IQ hebben maar ook omdat de spreiding tussen de leerlingen veroorzaakt wordt door verschillen tussen de scholen en tussen de klassen. Een eenvoudige aselechte steekproef van leerlingen (uit de populatie van alle Vwo-leerlingen van Nederland) zou in dat geval niet voldoen aan de eis dat alle elementen uit de steekproef “evenveel op elkaar lijken”. Twee leerlingen uit dezelfde klas, die allebei voorkomen in de steekproef, lijken nu eenmaal meer op elkaar dan twee leerlingen uit verschillende scholen. Bij een eenvoudige aselechte steekproef wordt er bij de analyse vanuit gegaan dat elk steekproefelement uit “dezelfde verdeling” komt waarin je geïnteresseerde bent. In het scholierenvoorbeeld is de verdeling, waaruit je trekt, “vervuild” met ongewenste invloeden (klaseffecten, schooleffecten).

Naast steekproefontwerpen met een stratificatie of een clustering bestaan er ook mengvormen hiervan. Zo zou je in het slaapverstoringsonderzoek rond Schiphol kunnen stratificeren op gemeenten (alle gemeenten binnen een straal van 15 kilometer rond Schiphol), vervolgens personen kunnen kiezen per gemeente en per persoon een aantal herhaalde metingen kunnen uitvoeren.

2.9 Netwerk steekproefontwerp

In een onderzoek waar gezocht wordt naar de prevalentie van bijvoorbeeld een zeldzame ziekte is het zaak om veel respondenten te hebben anders zal er nooit een relevante (significant groter dan nul) prevalentie gevonden worden. Een eenvoudige en goedkope manier om de steekproef te vergroten is gebaseerd op het idee van ‘netwerk sampling’. Hierbij wordt, aan de geselecteerde (volwassen) respondenten in het onderzoek, gevraagd om niet alleen over henzelf te beantwoorden of ze een bepaalde ziekte, eigenschap hebben, maar ook of hun nakomelingen (of kennissen e.d.) hieraan onderworpen zijn. Het gevolg hiervan is dat personen met veel nakomelingen, die wonen in meerdere huishoudens, een grotere kans hebben om in de steekproef voor te komen. Daarom is het gewone steekproefgemiddelde een onzuivere schatter voor het populatiegemiddelde. Een schatter voor het populatiegemiddelde die rekening houdt met de ongelijke kansen waarmee bepaalde populatie-elementen in de steekproef voorkomen is echter wel zuiver (zie Thompson (1992), hoofdstuk 15).

Netwerk sampling is ontstaan doordat het onvermijdelijk naar voren kwam in steekproefsituaties zoals patiënten wiens records in meerdere getrokken medische centra aanwezig zijn. Normale schatters geven dan onzuivere uitkomsten. Daarom moest men rekening houden met de multipliteit van de personen in de steekproef, dat is het aantal geselecteerde units waaraan een persoon “gelinked” is om dit vervolgens te vertalen naar de ongelijke insluitkansen van de elementen in de steekproef. Later zag men in dat netwerk sampling een manier is om het aantal waarneming, en dus de opbrengst (in termen van nauwkeurigheid, efficiency) van de steekproef, te vergroten. Vergroting van de steekproef door individuen uit de steekproef zelf nieuwe individuen te laten aandragen wordt ook wel Snowball sampling genoemd.

2.10 Onbekende populatieomvang

We gaan er in het hele rapport vanuit dat de populatieomvang bekend is. Dit hoeft natuurlijk niet het geval te zijn. Als ze niet bekend is dan zal er een schatting gemaakt moeten worden van de steekproefomvang. Dit kost altijd meer moeite en tijd en gaat dus ten koste van het onderzoeksbudget.

Een manier om de populatieomvang te schatten is met behulp van capture-recapture. Dit gaat als volgt te werk: neem een kleine steekproef met omvang n uit de populatie en noteer de populatie-elementen zodat later duidelijk is dat deze elementen geselecteerd zijn geweest. Neem vervolgens nog een keer een steekproef, ter grootte m , uit de populatie en kijk vervolgens hoeveel elementen uit de tweede steekproef ook in de eerste steekproef voorkwamen. Als dit laatste aantal gelijk is aan x en als de tweede steekproef onafhankelijk verkregen is van de eerste steekproef kun je de steekproefomvang schatten door $m * n / x$. In Thompson (1992) wordt een uitgebreide beschrijving gegeven van (single) capture-recapture.

Een andere manier om het probleem aan te pakken is de volgende: Stel dat de populatieomvang zelf niet bekend is, maar je weet wel dat de populatie verdeeld is over verschillende clusters, die in omvang kunnen verschillen maar waarvan het aantal wel bekend is of simpel geteld kan worden. In zo'n geval kun je de populatieomvang schatten door eerst een aantal clusters aselekt zonder teruglegging te selecteren en hiervan de clusteromvang te bepalen door simpelweg alle elementen uit het cluster te tellen. We noteren het totaal aantal clusters met L en het aantal geselecteerde clusters met a . Bereken vervolgens M , de som van de gevonden clusteromvangs. De populatieomvang kan nu geschat worden met behulp van $L * M / a$.

Als de populatie veel groter is dan de beoogde steekproefomvang n is het niet zo belangrijk om de omvang te schatten. In de meeste formules komt namelijk de populatieomvang N alleen voor in termen van de eindigheidscorrectie $(1 - n / N)$. Deze wordt dan gewoon op één gesteld, en er wordt dan net gedaan alsof de populatie oneindig groot is. In de praktijk wordt aangeraden om deze benadering alleen toe te passen als de eindigheidscorrectie groter is dan 0,90.

2.11 Non-respons

Het ontbreken van gegevens is een veel voorkomend en vaak lastig probleem bij het nemen van steekproeven en bij het analyseren van de data. Het grootste gevaar van non-respons is de mogelijkheid van een ernstige vertekening van de uitkomsten als de non-respons en de daarmee samenhangende ondervertegenwoordiging van bepaalde groepen uit de populatie blijkt samen te hangen met een belangrijke uitkomstvariabele. In Reijneveld & Stronks (1999) wordt voor een studie in Amsterdam beschreven wat de vertekening van non-respons is op de schattingen voor gebruik van de gezondheidszorg.

Non-respons is op te splitsen in twee typen, te weten: survey non-respons en item non-respons. We spreken van survey non-respons als van een deel van de steekproef geen gegevens bekend zijn. Voorbeelden zijn: mensen die niet thuis zijn bij telefonische enquêtes, het niet terugsturen van een schriftelijke enquête, ontbrekende geluidsmetingen doordat een meetapparaat niet goed afgesteld staat etc.

In het geval van item non-respons zijn er cases binnen de steekproef waarvoor bepaalde items (vragen of variabelen) niet bekend zijn. Denk hierbij bijvoorbeeld aan (geluidniveau-meet)apparatuur dat het begeeft tijdens het onderzoek en waar men pas later achter komt, geïnterviewden die bepaalde vragen in een enquête niet ingevuld hebben, mensen die halverwege een longitudinaal onderzoek verdwijnen uit de steekproef (door verhuizing, overlijden etc).

2.11.1 Survey non-respons

Over survey non-respons is al veel geschreven. De redenen voor dit type non-respons kunnen zeer divers zijn. Uit onderzoek blijkt onder andere dat de bereidwilligheid om deel te nemen aan (telefonische) enquêtes lager is in de Randstad en in grote steden t.o.v. de rest van Nederland, dat bewoners van dure woningen vaker bereid zijn om deel te nemen, dat ouderen minder bereid zijn om deel te nemen en dat de hoger opgeleiden, hogere sociale klassen volgens eigen opgave vaker bereid zijn deel te nemen aan een onderzoek. Het blijkt ook dat jongeren en respondenten uit hogere sociale klassen vaker bereidwillig zijn, maar minder bereikbaar zijn (zie de Leeuw & Hox (1998) voor een samenvattend artikel).

Het blijkt dat de non-respons lager is naarmate de beoogde populatie specifiek is. Dit komt doordat de relevantie voor de respondent in de regel dan groter is en de onderzoeker in zo'n situatie meer energie zal steken in het bereiken van de populatie. Bij regulier veldwerkprocedures ligt ze tussen de 50 á 60 %, in tegenstelling tot andere Europese landen waar de respons bij overheidsonderzoeken wel 80 % bedraagt. In het algemeen levert 'face to face' onderzoek het beste resultaat qua respons. De non-respons bij telefonische enquêtes kan verhoogd worden door het aantal contactpogingen te verhogen.

Bij survey non-respons is het van belang om bij de analyses een statistische correctie aan te brengen om het effect van de non-respons te verkleinen. Hiervoor zijn verschillende technieken voor gebruikelijk, zoals *weging* en *poststratificatie*.

Bij weging wordt aan de ondervertegenwoordigde groepen een groter gewicht toegekend, voor het schatten van populatie parameters, dan aan de oververtegenwoordigde groepen. Wanneer we van een aantal variabelen weten wat de verdeling is in de populatie, bijvoorbeeld op grond van een ander onderzoek, en we kennen de verdeling van die variabelen in de gerealiseerde steekproef, dan kunnen we de ratio als wegingvariabele gebruiken. Stel we een onderzoek houden onder huishoudens van de provincie Utrecht die voorkomen in het PTT-telefoonbestand. We trekken hiertoe een steekproef uit het PTT-telefoonbestand van de provincie Utrecht en het blijkt dat het percentage huishoudens, met telefoon, uit de steden van Utrecht dat in de steekproef

voorkomt zo'n 60 procent is. Als nu ook bekend is dat in de Utrechtse populatie, met telefoon, het percentage huishoudens dat in de steden woont 65 procent bedraagt dan wordt de weegfactor voor de huishoudens met telefoon uit de steden $w = (\text{populatie proportie}) / (\text{steekproef proportie})$ gelijk aan 1.08. Met behulp van deze weegfactor kunnen we de steekproef, wat betreft het aantal huishoudens uit een stad, representatiever maken voor de populatie. Het kan zijn dat ook nog een correctie nodig is om de juiste populatieomvang te krijgen. We gaan ervan uit dat de steekproef ook voor wat betreft andere variabelen die van belang zijn in het onderzoek meer representatief is.

Het effect van het wegen wordt groter naarmate de samenhang van de doelvariabelen voor het onderzoek en de wegingvariabelen groter is. Het meest ideaal is de situatie waarbij binnen de categorieën van de wegingvariabelen weinig variatie is van de onderzoeksvariabelen. Dit is eigenlijk precies hetzelfde fenomeen als verkleining van de variatie binnen subpopulaties door stratificatie. Het enige verschil is dat bij stratificatie men van tevoren hierover nadenkt en het steekproefontwerp zodanig neemt dat er bewust een stratificatie plaatsvindt. In het geval van weging wordt het achteraf gestratificeerd, vandaar ook wel poststratificatie genoemd.

Het nadeel van wegen is dat het bepalen van varianties en correlaties ingewikkelder wordt door de weging. SPSS kan zelfs foute resultaten geven hierbij!

Weging wordt moeilijker naarmate meerdere variabelen voor de weging gebruikt moeten worden. Vaak zijn namelijk alleen de marginale kansen bekend van de variabelen en voor een juiste weging moet men over de gezamenlijke kansen van de variabelen beschikken. In dat geval worden ingewikkeldere technieken gebruikt, zoals lineair of multiplicatief wegen (zie Bethlehem en Kersten (1987)). Multiplicatief wegen staat ook bekend als iteratief proportioneel wegen of raking (zie Brick & Kalton (1996)). Het programma Bascula (ontwikkeld bij het CBS) kan gebruikt worden om wegingstechnieken toe te passen zoals poststratificatie, lineair wegen en multiplicatief wegen (zie Bascula, reference manual).

2.11.2 Item non-respons

De mate waarin item non-respons een probleem vormt binnen een onderzoek is afhankelijk van de achterliggende reden van het ontbreken van de gegevens en de mate waarin er gegevens ontbreken. Het ligt voor de hand dat hoe selectiever de steekproef wordt door het ontbreken van gegevens en hoe groter het aantal ontbrekende gegevens des te groter het probleem wordt en er ook statistisch geavanceerdere technieken uit de kast getrokken moeten worden. In het ergste geval is het probleem zo groot dat er niets meer aan te doen is.

Een onderverdeling van item non-respons is gebaseerd op het onderliggende responsgedrag of de respons-status van mensen. Als alle personen uit een steekproef met dezelfde kans een non-respons zijn dan spreken we van 'missing completely at random', afgekort MCAR. In dat geval heb je in feite een kleinere steekproef als dat je van tevoren gepland had, wat een lagere nauwkeurigheid geeft van de schattingen. De conclusies uit een onderzoek waar de non-respons van het type MCAR zijn zuiver en dus niet vertekend door de non-respons.

Helaas is MCAR toch wel een ideale, niet veel voorkomende situatie. De situatie waarbij de non-respons op een bepaalde vraag of variabele uit een onderzoek, gegeven de uitkomsten op andere vragen uit hetzelfde onderzoek willekeurig is noemen we 'missing at random' (MAR). In deze situatie is dus in feite alle informatie over de reden van non-respons bekend. Het gewoon toepassen van standaard statistische technieken op het compleet geobserveerde deel van de data geeft onzuivere schattingen.

Als het responsmechanisme niet MAR is dan is het 'missing not at random' (MNAR). Dit is bijvoorbeeld het geval als bij een vraag naar inkomen juist de respondenten die een hoog inkomen genieten dit inkomen niet op willen geven. Als er nu andere vragen in zo'n zelfde onderzoek gesteld zouden zijn waarmee het onbekende inkomen geschat zou kunnen worden heb je in feite weer een MAR-situatie terug. In de epidemiologie spreekt men niet van MNAR maar van het feit dat de respons-status een effect-modifier is voor de relatie tussen kenmerk en uitkomst. Het CBS corrigeert voor non-respons in de gezondheidsenquête door aan te nemen dat het de respons-status MAR is.

De meeste statistische pakketten zullen al protesteren bij een enkele ontbrekende waarde. Voor bepaalde statistische analyses zijn er dan nog wel wat reparaties ingebouwd die ingesteld kunnen worden (bijvoorbeeld in SPSS bij multiële regressie analyse). De eenvoudigste oplossing om het probleem van item non-respons te omzeilen is door verwijdering van die cases uit de steekproef waar één of meer ontbrekende gegevens in voorkomen. Dit wordt in de literatuur aangeduid als *list-wise deletion*. List-wise deletion zal leiden, zoals eerder gezegd, tot vertekening van de uitkomsten als de volledig waargenomen cases systematisch verschillen van de slechts deels geobserveerde cases, dus als er geen MCAR-situatie is. Daarnaast kan list-wise deletion, zeker bij multivariate analyses waarbij dus alleen die cases worden meegenomen waarvoor alle variabelen waargenomen zijn, leiden tot een grote verspilling van de aanwezige data. List-wise deletion is daarom alleen een goede aanpak als de er geen systematische verschillen zijn tussen de volledig waargenomen cases en de slechts deels waargenomen cases én onder de conditie dat de fractie ontbrekende gegevens klein is (kleiner dan 5 á 10 %).

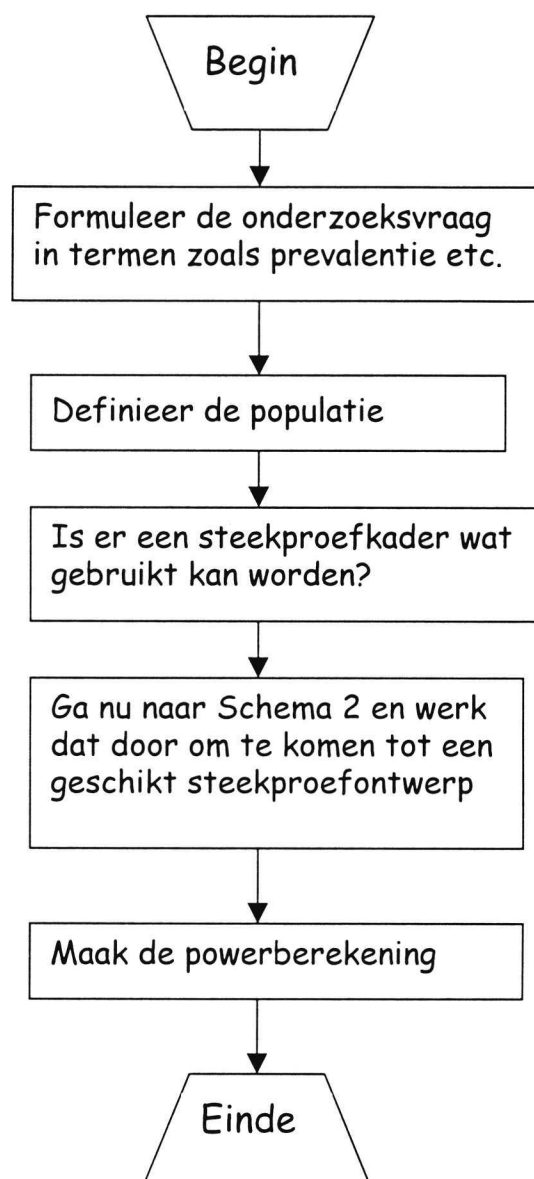
In de epidemiologie wordt in multiële regressie vaak gecorrigeerd voor missings values door een afzonderlijke dummy-variabele toe te voegen voor de ontbrekkenden. Uit onderzoek is gebleken dat deze methode kan leiden tot onzuivere resultaten, zie Greenland en Finkle, (1995).

Andere oplossingen voor item non-respons zijn (zie Little & Rubin (1987)) *available case deletion*, *mean imputation*, *single imputation*, *hot-deck imputation* en *multiple imputation*. Van al deze methoden is multiële imputatie de beste en meestal de enige juiste oplossing. Gelukkig komt er steeds meer goede software voor multiële imputatie beschikbaar. Binnen TNO is een procedure in het statistisch pakket S-PLUS ontwikkeld waarmee geïmputeerde datasets verkregen kunnen worden en waarmee vervolgens de gewenste analyses uitgevoerd kunnen worden. Een beschrijving van dit algoritme, genaamd MICE, is te vinden in van Buuren & Oudshoorn (1999). Oudshoorn et al. (1999) behandelt de toepassing van MICE op het AVO-95 bestand (het Aanvullend Verzoeken Onderzoek, uitgevoerd door het SCP in 1995).

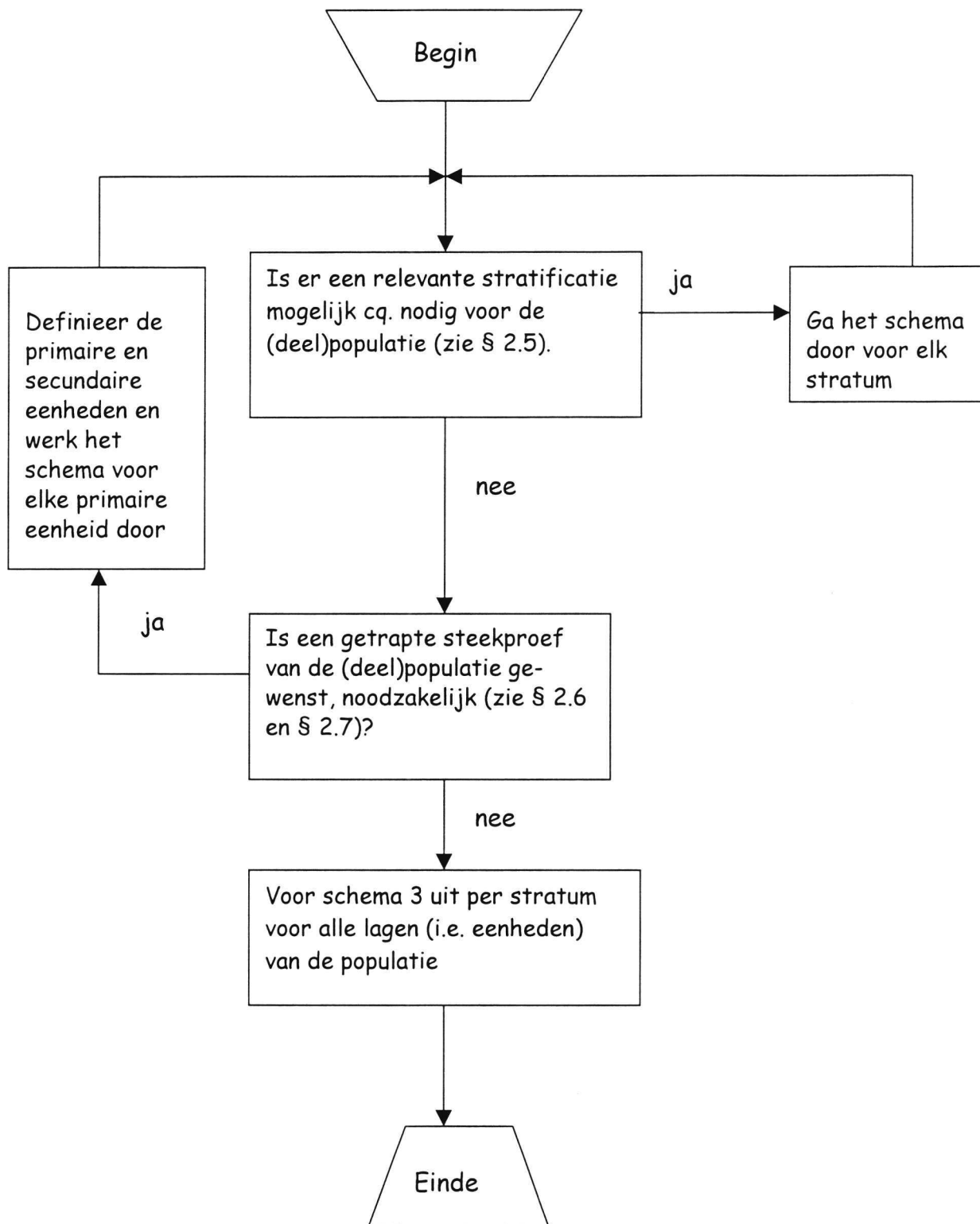
2.12 Samenvatting hoofdstuk 2

De inhoud van hoofdstuk 2 is samengevat in drie schema's die te zien zijn op de volgende pagina's.

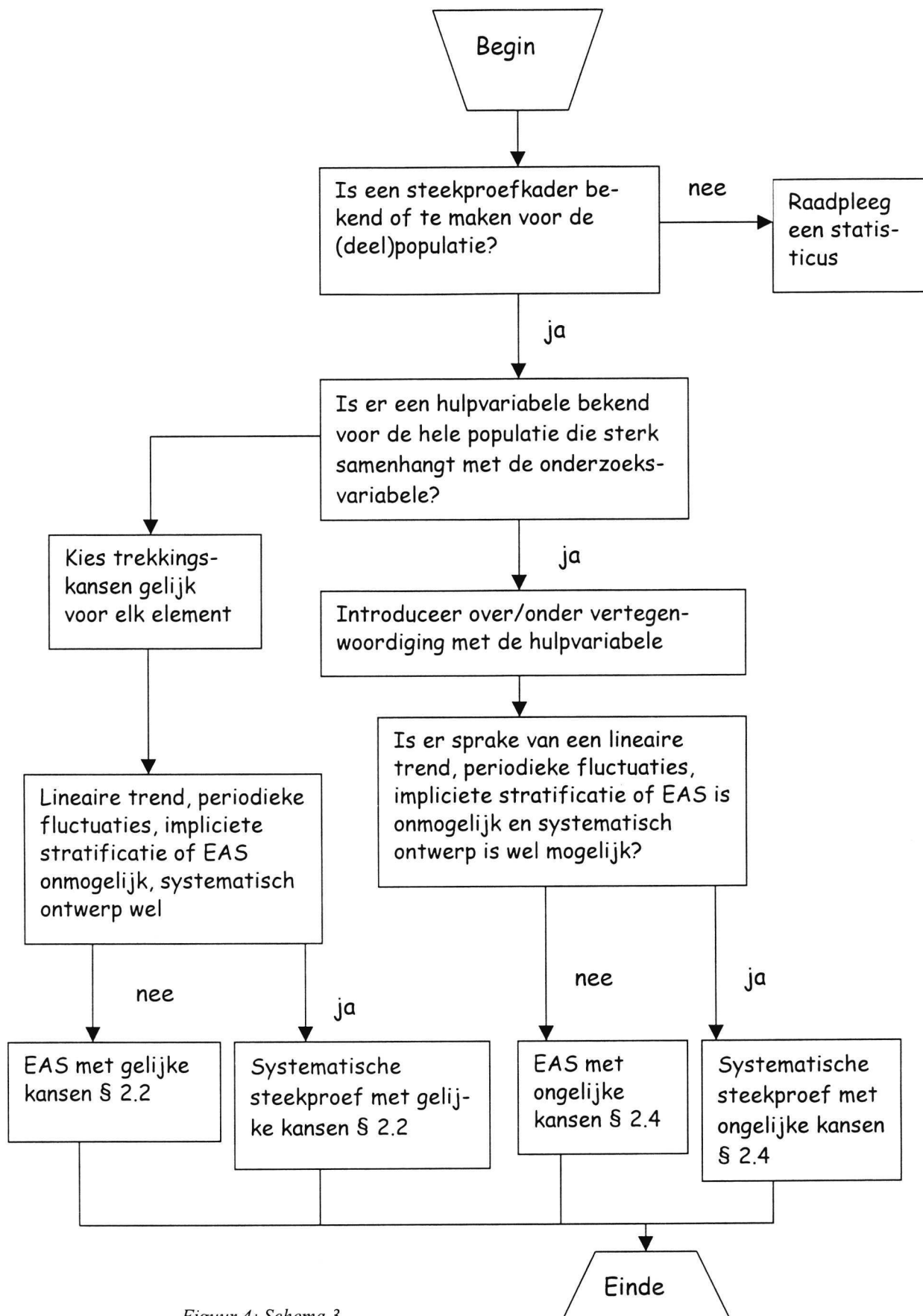
Om van de onderzoeksvraag te komen tot een verantwoord steekproefontwerp met daarbij als element een onderbouwde steekproefomvang berekening moet er een aantal stappen genomen worden. Bij deze stappen horen vragen die beantwoord dienen te worden. Het eerste schema geeft globaal weer hoe men komt van een onderzoeksvraag tot een uiteindelijke steekproefomvang berekening of een zogenaamde powerberekening. Binnen schema 1 wordt schema 2 aangeroepen. Schema 3 wordt aangeroepen binnen Schema 2.



Figuur 2: Processchema 1 voor steekproefontwerp bij een onderzoeksopzet.



Figuur 3: Schema 2



Figuur 4: Schema 3

3 Steekproefomvang berekeningen

In dit hoofdstuk gaan we in op het berekenen van de steekproefomvang die nodig is om een bepaalde nauwkeurigheid te bereiken. Eerst zullen we een aantal basisbegrippen uitleggen die noodzakelijk zijn. Vervolgens behandelen we het geval van steekproefomvang berekeningen voor parameters zoals het gemiddelde, het populatietotaal, populatiefractie, percentage en prevalentie in het geval van een EAS. De berekeningen uit deze paragraaf vormen de basis voor de meer ingewikkeldere steekproefontwerpen die daarna aan bod komen. In dit hoofdstuk laten we het steekproefontwerp met ongelijke kansen en netwerk sampling links liggen. Het doel van dit hoofdstuk is om een kader te scheppen en de basisideeën uit te leggen van waaruit de onderzoeker, (eventueel) met hulp van een statisticus, verder op weg kan om voor zijn specifieke situatie de minimaal benodigde steekproefomvang te bepalen. In dit hoofdstuk zullen de formules niet overal geschuwd worden. Wel is gepoogd deze formules te illustreren aan de hand van eenvoudige rekenvoorbeelden. Voor de begripsvorming is met name Paragraaf 3.1 van belang.

Bij een dosis-effect relatie wordt een gestratificeerd ontwerp geadviseerd. Hieraan wordt dan ook aandacht besteed aan het einde van paragraaf 3.3, wat gaat over steekproefomvang berekeningen bij stratificatie.

Steekproefomvang berekeningen, zoals die hierna beschreven worden, zijn geformuleerd in termen van een gewenste nauwkeurigheid. Dit is typisch een benadering voor beschrijvende onderzoeken, zoals bij de vakgroep Milieu van TNO Preventie en Gezondheid veel het geval is. In onderzoeken waar de nadruk meer ligt op het toetsen van hypothesen gaat men daarentegen uit van een bepaalde power die minimaal gehaald dient te worden. In paragraaf 3.6 zullen we op deze tweespalt in aanpak van steekproefomvang berekeningen dieper ingaan. Er is namelijk een sterke relatie tussen de nauwkeurigheid en het onderscheidend vermogen (power) van een uitspraak.

3.1 Algemene begrippen en achtergrond bij steekproefgrootte berekeningen

Om de discussie te vereenvoudigen voeren we eerst enige terminologie, notatie en definities in.

N	De omvang van de populatie
n	De steekproefomvang
p	De fractie populatie-elementen met een zekere eigenschap (bijvoorbeeld het gedeelte van de populatie dat zwaar gehinderd is door vliegverkeer); dan is $P = 100 p \%$ het bijbehorende percentage, oftewel de prevalentie. p wordt de populatiefractie genoemd.
$\hat{p}$	De fractie elementen uit de steekproef met een zekere eigenschap.

- μ Het populatiegemiddelde, dat wil zeggen het gemiddelde van alle y -waarden in de gehele populatie: $\mu = (y_1 + y_2 + \dots + y_N) / N$. Denk bijvoorbeeld aan de situatie waarbij y_i het inkomen is van een persoon uit de populatie, μ is dan het gemiddelde inkomen van alle personen in de populatie.
- $\bar{y}$ Het steekproefgemiddelde, dat wil zeggen het gemiddelde van alle y -waarden in de steekproef, dus $\bar{y} = (y_1 + y_2 + \dots + y_n) / n$.
- τ Het populatietotaal, dit is de som van alle y -waarden in de gehele populatie. In formule: $\tau = (y_1 + y_2 + \dots + y_N) = N\mu$.
- $\hat{\tau}$ De schatter voor het populatietotaal, $\hat{\tau} = N * (y_1 + y_2 + \dots + y_n) / n$

In feite kun je populatiefractie p ook opvatten als een populatiegemiddelde. De redenatie is als volgt: Definieer $y_i = 1$, als (populatie)element i de eigenschap wel heeft en $y_i = 0$, als element i de eigenschap niet heeft. Dan geldt dat $p = (y_1 + y_2 + \dots + y_N) / N = \mu$. Met dezelfde definitie van y_i geldt dan ook dat $\hat{p}$ eigenlijk een steekproefgemiddelde is.

Het type uitspraak dat men doet in onderzoek is vaak van de volgende (theoretische) vorm:

- ‘De (onbekende) fractie p ligt tussen de getallen $\hat{p}_L$ en $\hat{p}_U$, in formule, $\hat{p}_L \leq p \leq \hat{p}_U$ ’. Bijvoorbeeld: de fractie ernstig door wegverkeer gehinderden in de provincie Utrecht ligt tussen de 0,03 en 0,08.
- ‘De (onbekende) prevalentie P ligt tussen de getallen $\hat{P}_L$ en $\hat{P}_U$, in formule, $\hat{P}_L \leq P \leq \hat{P}_U$ ’. Bijvoorbeeld: de prevalentie van last bij in slaap komen door nachtelijk vliegverkeer rond Schiphol ligt tussen de 8 en 16 procent.
- ‘Het (onbekende) populatiegemiddelde μ ligt tussen de getallen $\hat{\mu}_L$ en $\hat{\mu}_U$, in formule, $\hat{\mu}_L \leq \mu \leq \hat{\mu}_U$ ’. Bijvoorbeeld: het aantal uren dat inwoners van Utrecht gemiddeld per jaar doorbrengen in een recreatiegebied ligt tussen de 20 en 35.

Hoe dichter de getallen $\hat{p}_L$, $\hat{p}_U$ bij elkaar liggen des te *nauwkeuriger* de uitspraken zijn. Een uitspraak, bijvoorbeeld, dat de fractie gehinderden rond Schiphol tussen de 0,30 en 0,70 ligt is veel minder nauwkeurig dan de uitspraak dat deze fractie tussen de 0,45 en de 0,55 ligt. In de statistiek wordt de halve afstand tussen de ondergrens $\hat{p}_L$ en de bovengrens $\hat{p}_U$ de *nauwkeurigheid*, genoteerd met d , genoemd. In formule geldt:

$$d = (\hat{p}_U - \hat{p}_L) / 2.$$

Dezelfde grootte kan men definiëren voor percentages, prevalenties en gemiddelden. Merk op, dat hoe sterker of nauwkeuriger de uitspraak is, hoe kleiner de nauwkeurigheid d is.

Wanneer een uitspraak zoals ‘de fractie ernstig geluidsgehinderden rond Schiphol ligt tussen de 0,30 en de 0,70’ wordt gedaan dan is deze juist of onjuist. Zekerheid hierover heb je niet en kan slechts verkregen worden door de hele populatie te onderzoeken, hetgeen nu juist niet de bedoeling is. De onzekerheid over de uitspraak kan echter wel gekwantificeerd worden in termen van de *onbetrouwbaarheid* α van de gebruikte methode om tot de uitspraak te komen (of in termen van $1 - \alpha$, de *betrouwbaarheid* van de methode). De onbetrouwbaarheid van een methode is de fractie onjuiste uitspraken die gedaan worden als dezelfde methode (oneindig) vaak wordt toegepast. Zo geeft een onbetrouwbaarheid van 0,05 (vijf procent) aan dat gemiddeld 1 op de 20 uitspraken niet waar is. In het engels wordt de onbetrouwbaarheid ‘*significance level*’ genoemd. In het algemeen geldt dat hoe sterker of **nauwkeuriger** een uitspraak is, hoe **onbetrouwbaarder** ze is, als de uitspraak gebaseerd is op dezelfde steekproef. Of andersom, hoe betrouwbaarder de uitspraak is, des te onnauwkeuriger ze zal zijn. Aan de andere kant geldt dat als een uitspraak gedaan wordt aan de hand van een grotere steekproef dan neemt de nauwkeurigheid en/of de betrouwbaarheid van de uitspraak toe.

Dit principe is de basis van een goede steekproefgrootte berekening of power-analyse. Hoe groter een steekproef, des te nauwkeuriger en betrouwbaarder uitspraken gedaan worden. Een steekproefgrootte berekening is dan ook een wisselwerking tussen nauwkeurigheid, betrouwbaarheid en minimale steekproef die daarvoor nodig is.

Standaard keuzen voor de betrouwbaarheid zijn 90 % (gemiddeld één op de tien uitspraken is niet waar), 95 % (gemiddeld één op de twintig uitspraken is niet waar) of 99 % (gemiddeld één op de honderd uitspraken is niet waar). Deze waarden komen overeen met respectievelijk $\alpha = 0,10$, $\alpha = 0,05$ en $\alpha = 0,01$. Voor de nauwkeurigheid worden de waarden $d = 0,10$, $d = 0,05$ of $d = 0,01$ genomen. Standaard combinaties van betrouwbaarheid en nauwkeurigheid zijn de paren $\alpha = 0,10$, $d = 0,1$, en $\alpha = 0,05$, $d = 0,05$.

3.2 Steekproefomvang berekeningen voor enkelvoudige steekproefontwerpen

In deze paragraaf bespreken we hoe de optimale steekproefomvang berekend wordt voor een enkelvoudig steekproefontwerp. De berekening hangt af van de parameter die geschat moet worden of waarvan een betrouwbaarheidsinterval gewenst is. We beperken ons tot de volgende parameters: het populatiegemiddelde μ , de populatiefractie p , de prevalentie (of percentage) P en het populatietotaal z . We geven, naast de formules voor de optimale steekproefomvang ook een schatter voor de variantie van de desbetreffende parameter en een formule voor het betrouwbaarheidsinterval. Tevens zullen we voor standaardkeuzen van de betrouwbaarheid en de nauwkeurigheid tabellen presenteren waaruit de optimale steekproefomvang afgelezen kan worden.

Een aantal begrippen zijn van belang voor het goed kunnen toepassen van de tabellen. In een aantal formules voor de steekproefgrootte komt S^2 voor. Dit is een maat voor de spreiding in de populatie. Om precies te zijn, voor de waarden $y_1, y_2, \dots, y_N$ van de populatie, geldt dat de populatievariantie σ^2 gelijk is aan:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2.$$

S^2 is gelijk aan $\sigma^2 * N/(N-1)$. We gebruiken S^2 i.p.v. de populatievariantie σ^2 omdat de formules dan iets eenvoudiger zijn. Voor grote N is het verschil hiertussen verwaarloosbaar. De populatievariantie σ^2 en dus ook S^2 kan slechts berekend worden als de waarden van alle populatie-elementen bekend zijn. Maar die zijn juist onbekend, anders zou immers het nemen van een steekproef overbodig zijn.

Kennis over de spreiding in de populatie is nodig om een steekproefomvang berekening uit te voeren. Dit volgt bijvoorbeeld uit het feit dat hoe groter de spreiding in de data is, des te groter de steekproef moet zijn om het populatiegemiddelde met een bepaalde nauwkeurigheid te schatten. Soms is een schatting van S^2 bekend uit eerder onderzoek. Deze waarde kan gebruikt worden als schatting voor S^2 . Als er geen schatting van S^2 bekend is, dan moet ze geschat worden. Indien er wel iets bekend is over de spreidingsbreedte* in de populatie dan is de volgende regel handig: *Voor dalende of unimodale† begrensde verdelingen van de data geldt dat de standaardafwijking $\sigma^\ddagger$ kleiner is dan $0,29 * \text{spreidingsbreedte in de populatie}$, zie Muilwijk (1992).* Voor verdelingen met een niet te smalle en hoge top geldt dat σ in de regel groter is dan $0,20 * \text{spreidingsbreedte}$. Daarom kan $0,25 * \text{spreidingsbreedte}$ als ruwe schatting genomen worden voor σ . Uitschieters kunnen het beste weggelaten worden uit de berekening van de spreidingbreedte.

Een nog niet gedefinieerde grootheid, is $u_{\alpha/2}$. Dit getal is een waarde van de Student-t verdeling en hangt af van de gekozen betrouwbaarheid α en van de steekproefomvang n . Voor standaardwaarden van de betrouwbaarheid is $u_{\alpha/2}$ af te lezen uit Tabel 1. In de literatuur wordt $u_{\alpha/2}$ ook wel aangeduid met de t -waarde. Voor grote n kan de t -waarde benaderd worden door de z -waarde. Dit is precies die waarde van $u_{\alpha/2}$ bij $n > 120$. In dat geval is het een waarde uit de normale verdeling. We gebruiken de notatie $u_{\alpha/2}$ i.p.v. t om te benadrukken dat ze afhangt van de betrouwbaarheid α .

Tabel 2: waarden voor $u_{\alpha/2}$ bij gegeven betrouwbaarheid (horizontaal) en steekproefomvang (verticaal)

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 10$	1,81	2,23	3,17
$n = 60$	1,67	2,00	2,66
$n > 120$	1,64	1,96	2,58

*. De spreidingsbreedte is gedefinieerd als de grootste waarde in de populatie min de kleinste.

† Een unimodale verdeling van de data is een verdeling met slechts één top. De normale verdeling is een voorbeeld van een unimodale verdeling.

‡ De standaardafwijking σ is de wortel uit σ^2 . In de literatuur wordt ze ook wel met standaardfout aangeduid. S is de wortel uit S^2 .

De waarde van $u_{\alpha/2}$ hangt af van de steekproefomvang die nog niet gekozen is en wat nu juist hetgeen is waarnaar we op zoek zijn. Omdat de onzekerheid in de steekproefomvang berekening aanzienlijk is (alleen al omdat de variantie meestal geschat wordt en in de regel is die schatting niet nauwkeurig) kun je in de praktijk uitgaan van de laatste regel van de kolom, dus $n > 120$. Als er vervolgens een optimale steekproefomvang berekend is die klein is ten opzichte van 120, dan moet de berekening over gedaan worden met een grotere waarde voor $u_{\alpha/2}$.

In de praktijk van sociaal wetenschappelijk en preventieve gezondheidsonderzoek zoals bij TNO Preventie en Gezondheid zullen de te onderzoeken populaties veelal groot zijn ten opzichte van de steekproef die men gaat nemen. De verhouding van de steekproef n en de populatieomvang N wordt genoteerd met f . Als f niet klein is dan moet er een eindige populatiecorrectie toegepast worden op de schattingen. In Tabel 3 (derde rij) staat in de formules voor de verschillende variantieschatters de term $(1-f)$. Dit is de eindige populatiecorrectie. In de berekening van de optimale steekproefomvang is de eindige populatiecorrectie verwaarloosd. Er is echter wel een vuistregel om, indien nodig, hiervoor toch te corrigeren. Zie hiervoor de voetnoot bij de optimale steekproefomvang in de tabel.

Nu hebben we alle nog onbekende begrippen de revue laten passeren die nodig zijn om Tabel 3 met alle optimale steekproefomvangs, schatters en varianties te kunnen gebruiken. Merk op dat in de berekening van de optimale steekproefomvang (zesde regel in Tabel 3) gewerkt wordt met (een ruwe schatting van) S^2 en niet met s^2 , de schatting van S^2 (vierde regel in Tabel 3), die uitgerekend kan worden als de steekproef eenmaal genomen is. Deze variantieschatting is nodig om bijvoorbeeld een betrouwbaarheidsinterval uit te kunnen rekenen (zie de vijfde regel in Tabel 3).

Voorbeeld [Optimale steekproefomvangberekening voor een onderzoek naar afvalscheiding]:

*In de provincie Utrecht wil men door middel van een onderzoek achterhalen hoeveel GFT-afval de inwoners aan gewicht gemiddeld per week produceren. De parameter waar men in geïnteresseerd is dus een populatiegemiddelde. Deskundigen op het gebied van afval gaan ervan uit dat het gewicht per gezin varieert van 3 tot 10 kilo per week. Daarom nemen we $7 * 0,25 = 1,75$ als schatting voor de standaardafwijking. Het kwadraat hiervan, 3,06, is een schatting voor S^2 . Uit Tabel 2 volgt dat voor een betrouwbaarheid van 0,05 we in eerste instantie voor $u_{\alpha/2}$ de waarde 1,96 kunnen kiezen. De gemeente wil het gemiddeld gewicht met een nauwkeurigheid van 0,5 kilo achterhalen. Dit vullen we dus als waarde voor d in. Nu zijn alle benodigde gegevens op een rijtje gezet. M.b.v de formule voor de minimale steekproefomvang met absolute d voor het populatiegemiddelde uit Tabel 3 volgt de minimaal benodigde steekproefomvang 47 bedraagt (want $(1,96)^2 * 3,06 / (0,5)^2 = 47$). Omdat Utrecht meer dan 100.000 inwoners heeft hoeven we geen rekening te houden met de eindige populatiecorrectie. Als de gemeente niet op 0,5 kilo nauwkeurig maar op 0,1 kilo nauwkeurig het gewicht wil weten dan is de minimale benodigde steekproefomvang gelijk aan 1176.*

Parameter	Populatiegemiddelde μ	Populatietotaal τ	populatiefractie* p
Schatter	$\bar{y} = (y_1 + \dots + y_n) / n$	$\hat{\tau} = N * (y_1 + \dots + y_n) / n$	$\hat{p} = \text{aantal enen in de steekproef} / \text{totaal aantal cases}$
Variantie schatter†	$s^2(\bar{y}) = \frac{1-f}{n * (n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$	$s^2(\hat{\tau}) = \frac{N^2 * (1-f)}{n * (n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$	$s^2(\hat{p}) = \frac{(1-f)}{(n-1)} \hat{p} * (1-\hat{p})$
Betrouwbaarheids-interval	$\bar{y} - u_{\alpha/2} * s(\bar{y}) \leq \mu \leq \bar{y} + u_{\alpha/2} * s(\bar{y})$	$\hat{\tau} - u_{\alpha/2} * s(\hat{\tau}) \leq \tau \leq \hat{\tau} + u_{\alpha/2} * s(\hat{\tau})$	$\hat{p} - u_{\alpha/2} * s(\hat{p}) \leq p \leq \hat{p} + u_{\alpha/2} * s(\hat{p})$
Minimale steekproefomvang (met absolute d)‡	$n = \frac{u^2_{\alpha/2} * S^2}{d^2}$	$n = \frac{u^2_{\alpha/2} * S^2 * N^2}{d^2}$	$n = \frac{u^2_{\alpha/2} * p * (1-p)}{d^2} \leq \frac{0,25 * u^2_{\alpha/2}}{d^2}$ §
Minimale steekproefomvang (met relatieve d)	$n = \frac{u^2_{\alpha/2} * S^2}{(r * \mu)^2}$	$n = \frac{u^2_{\alpha/2} * S^2 * N^2}{(r * \tau)^2}$	$n = \frac{u^2_{\alpha/2} * p * (1-p)}{(r * p)^2}$

Tabel 3: formules voor de minimale steekproefomvang voor verschillende parameters.

* Voor percentages en proporties, gebruik de laatste kolom waarbij voor fracties steeds prevalenties of percentages worden ingevuld in plaats van fracties.

† Als de eindige populatiefractie f klein is dan kan de term $1-f$ weggelaten worden uit de formule.‡ De eindige populatie correctie is hierbij verwaarloosd. Als de berekende minimale steekproefomvang n groot blijkt te zijn t.o.v. de totale populatie omvang N dan moet als uiteindelijke steekproefomvang niet n genomen worden maar $n / (1 + n / N)$.§ De minimale steekproefomvang bij fracties kan afgeschat worden door voor $p(1-p)$, het maximum 0,25 in te vullen. Dan hangt de formule niet meer af van de (onbekende) fractie. Deze afchatting is voor fracties kleiner dan 0,20 en groter dan 0,80 aan de veilige kant. In dat geval is de werkelijke minimale steekproefomvang minder groot. Het voordeel van de formule is wel dat er geen kennis van tevoren nodig is van de grootte van de fractie.** De nauwkeurigheid d is hier relatief aan de waarde van de parameter gekozen, dus bijvoorbeeld voor het gemiddelde is de relatieve nauwkeurigheid gelijk aan r/μ , waarbij r de relatieve fout is in het schatten van μ . Voor de minimale steekproefomvang met relatieve d geldt dezelfde opmerking over de eindige steekproefcorrectie als bij de minimale steekproef met absolute d .

Voor fracties (en dus ook voor percentages en prevalenties) geldt dat de formule van de optimale steekproefomvang geen populatievariantie of fractie zelf bevat, indien de benaderende formule wordt gebruikt (zie de vierde voetnoot bij Tabel 3 en de uitdrukking na het klein gelijk teken). Dat wil zeggen dat voor een gegeven nauwkeurigheid en betrouwbaarheid de minimale steekproefomvang alleen nog maar afhangt van de populatieomvang. In Tabel 4 kun aflezen wat de minimale steekproefomvang is bij gegeven populatieomvang en betrouwbaarheid en nauwkeurigheid voor fracties, percentages en prevalenties*. Voor percentages kleiner dan 10,0 % en groter dan 90 % zijn deze aantallen niet zonder meer toepasbaar. In dat geval is het aan te raden om de hulp van een statisticus te vragen.

Voor populaties die minstens uit 100.000 elementen bestaan geldt dat een steekproef van 405 voor een betrouwbaarheid en nauwkeurigheid van 5 % altijd voldoende is. Dit komt doordat de eindige populatie correctie vanaf 100.000 afgerond nul is. De invloed van deze correctie is voor kleinere populaties wel merkbaar, en daarom is de verhouding steekproefomvang versus populatieomvang groter naarmate de populatieomvang kleiner is, dit zie je ook terugkomen in de optimale steekproefaantallen gegeven in Tabel 4.

Tabel 4: De minimale steekproefomvang bij fracties, prevalenties en percentages bij gegeven N , α en d .

Populatieomvang N	$\alpha = 10 \%, d = 10 \%$	$\alpha = 5 \%, d = 5 \%$
100	47	84
250	62	160
500	69	229
750	72	268
1000	74	293
2000	76	340
5000	78	377
7000	78	384
8000	78	387
10.000	78	390
50.000	79	402
100.000	79	405

* Deze aantallen zijn exact uitgerekend en daarom kan het voorkomen dat een powerberekeningspakket als Nquery Advisor iets andere aantallen geeft omdat die gebruik maakt van benaderingen.

3.3 Steekproefomvang berekeningen bij stratificatie

Eén van de redenen om een gestratificeerde steekproef te gebruiken is als er verschillende groepen in de populatie aan te wijzen zijn die homogeen zijn ten aanzien van de parameter waarin we geïnteresseerd zijn, homogener dan in de populatie in het geheel. Statistisch gezien betekent dit dat er groepen in de populatie zijn waarbinnen de standaardafwijking (als maat van spreiding) in de parameter kleiner is dan de standaardafwijking van de parameter over de hele populatie. In deze paragraaf zullen we aangeven hoe de steekproefomvang per stratum gekozen dient te worden zodat we uiteindelijk een zo nauwkeurig mogelijke schatting krijgen van de parameter met zo min mogelijk waarnemingen.

Eerst voeren we wat notatie in. Stel de populatie bestaat uit N elementen. L is het aantal strata waarin we de populatie verdelen. Zij $N_1, N_2, \dots, N_L$ de aantallen behorende bij de L strata (dus N_1 is het aantal elementen uit de populatie in het eerste stratum enzovoorts). Logischerwijs is N gelijk aan de som van N_1 tot en met N_L . Vervolgens onderscheiden we twee situaties:

1. De standaardafwijking per stratum is bekend of een ruwe schatting ervan is bekend. Deze noteren we nu met $S_1, \dots, S_L$.
2. De standaardafwijking per stratum is niet bekend of kan gelijk verondersteld worden voor elk stratum. We gaan er dan vanuit dat de standaardafwijking van de gehele populatie wel bekend is of geschat kan worden. Deze noteren we met S .

De keuze van de steekproefomvang bij een gestratificeerde steekproef bestaat uit twee stappen. Ten eerste moet de grootte van de totale steekproefomvang, n , bepaald worden en ten tweede moet de meest optimale allocatie (verdeling) van deze n steekproefelementen over de verschillende strata gekozen worden. We gaan er in eerste instantie vanuit dat de totale steekproefomvang vast ligt en gelijk is aan n .

Voor de eerste hierboven geschetste situatie berekenen we de optimale steekproefgrootte met behulp van het zogenoemde *Neyman allocatie schema* (Cochran 1977). In feite is dit schema het meest optimale schema, maar het zal blijken dat de winst van dit schema boven de schema's die verder aan bod komen in praktische situaties minimaal zal zijn in termen van nauwkeurigheid of steekproefomvang (en dus budget). Voor de volledigheid behandelen we dit schema wel.

Het Neyman allocatieschema zegt het volgende: als de totale grootte van de steekproef gelijk is aan n dan moeten we de steekproefomvang per stratum (genoteerd met $n_1, n_2, \dots, n_L$) zodanig kiezen dat de steekproefgrootte n_i van het i -de stratum gelijk is aan:

$$n_i = n \frac{N_i S_i}{\sum_{k=1}^L N_k S_k}.$$

Het Neyman allocatie schema komt erop neer dat n_i groter wordt naarmate de omvang van stratum i , N_i (ten opzichte van de totale populatieomvang) en/of de standaardafwijking in het i -de stratum S_i groter is (ten opzichte van standaardafwijking in de totale populatie). Om het Neyman allocatie schema te kunnen toepassen moet men, per stratum, beschikken over de grootte van het stratum en de standaardafwijking binnen het stratum. Zoals gezegd in paragraaf 3.2 kun je door de spreiding binnen een stratum te nemen en deze met 0,25 te vermenigvuldigen een ruwe schatting krijgen van de standaardafwijking binnen een stratum.

Als er van tevoren al strata bekend zijn waar de spreiding veel groter is dan in de andere strata dan is het aan te bevelen om het desbetreffende stratum geheel mee te nemen in de steekproef of hierin veel waarnemingen te doen. Bijvoorbeeld als de Nederlandse bevolking gestratificeerd is naar inkomensklassen dan zal de spreiding in eigen vermogen per Nederlander, binnen het stratum van de hoogste inkomens veel groter zijn dan in de andere klassen en daarom wordt geadviseerd om het stratum van de hoogste inkomens te “oversamplen”. Zo laat het CBS bij onderzoek naar de economische activiteit van bedrijven de grootste bedrijven in Nederland buiten de steekproeftrekking, en neemt de gegevens van deze bedrijven standaard mee in de analyse. De grootste bedrijven kun je in dit geval zien als een stratum dat geheel wordt meegenomen in de steekproef.

Het nadeel van het Neyman allocatie schema is dat er verondersteld wordt dat alle individuele standaardafwijkingen per stratum bekend zijn of in ieder geval een ruwe schatting ervan. Dit zal in de praktijk meestal niet het geval zijn. Een simpel en populaire alternatieve methode voor het berekenen van de optimale steekproefgrootte is het *proportionele allocatie schema*, waarbij men voor elk stratum een steekproefgrootte kiest die proportioneel is met de fractie N_i / N . Dit betekent in formulevorm dat $n_i = n * N_i / N$. Enige voorzichtigheid is geboden bij het afronden van n_i naar een geheel getal. Hoe groter het verschil tussen de afronding en de optimale n_i zelf, des te groter zal de vertekening zijn van bijvoorbeeld het gemiddelde. Deze vertekening wordt in de praktijk meestal verwaarloosd. Strikt genomen moet je bij de berekening van bijvoorbeeld het gemiddelde een gewogen gemiddelde gebruiken waarbij gewogen wordt met gewichten N_i / N .

Als de standaardafwijkingen per stratum gelijk zijn dan is het Neyman allocatie schema gelijk aan het proportionele allocatie schema. Des te meer de standaardafwijkingen per stratum van elkaar verschillen, des te beter is het om het Neyman allocatie schema te gebruiken. Helaas is dit praktisch niet altijd mogelijk. Toch is gebruik van het proportionele allocatie schema niet slecht, temeer ze, in vergelijking met een enkelvoudige aselechte steekproef (dus zonder strata), kleinere betrouwbaarheidsintervallen geeft voor de parameters (Rice [1988], hoofdstuk 7.5). Daarnaast is het zo dat voor fracties (en dus percentages en prevalenties) het verschil in efficiency tussen het Neyman allocatie schema en het proportionele schema pas echt relevant wordt als de gezochte fractie zeer klein of zeer groot is. Voor de relatieve precisie (genoteerd met $RP(\%)$), uitgedrukt in de verhouding tussen de variantie bij het schatten van een fractie met het Neyman allocatie schema en de variantie bij het schatten van dezelfde fractie maar dan met behulp van het proportionele schema geldt het volgende: als de fracties in de strata variëren tussen de 0,40 en de 0,60 dan is $RP(\%)$ gelijk aan 99,99 %. Er is dus bijna geen verschil tussen beide schema's. Voor een fractie tussen 0,20 en 0,80 is de $RP(\%)$ minstens 98,77 %. Voor een fractie variërend tussen 0,1

en 0,9 is RP(%) minstens 93,75 % en voor een fractie tussen de 0,05 en 0,95 is ze tenminste 84,57 % (zie Moors (1973)). Door de wortel uit de RP(%) te nemen krijg je de relatieve precisie in termen van de standaardafwijking van de geschatte fractie. Uit de getallen hierboven volgt dat gezochte fractie erg klein of groot moeten zijn voordat het Neyman allocatie schema echt de voorkeur verdiend boven het proportionele allocatie schema.

Voor het schatten van een gemiddelde (of een populatie totaal) kan een zelfde verhaal gehouden worden. Afhankelijk van het maximale verschil tussen de standaardafwijkingen per stratum kan een afschatting geven worden van de relatieve precisie zoals hierboven is gedefinieerd. Als U het quotiënt is van de maximale voorkomende standaardafwijking bij een stratum en de minimale voorkomende standaardafwijking bij een stratum (dus $U = \max S_i / \min S_i$) dan geldt dat voor een U van 1,5 de relatieve precisie 96,0 % is, voor U gelijk aan 2 is ze 88,9 %, voor U gelijk aan 2,5 is ze 81,6 % en voor U gelijk aan 3 is ze 75 %. In termen van de standaardafwijking moet je bij deze getallen de wortel nemen. Een U van 2 betekent dat de maximale voorkomende standaardafwijking twee keer zo groot is als de minimale standaardafwijking. In dat geval is het verlies bij gebruik van het proportionele allocatie schema in plaats van het Neyman allocatie schema, in termen van de nauwkeurigheid van de schatting hoogstens 2 % (bij een α van 5 %).

Naast het proportionele allocatie schema en het Neyman allocatieschema zijn er nog twee andere standaard manieren om de totale steekproefomvang te verdelen over de verschillende strata. Het meest eenvoudige is het *gelijke allocatieschema*. Hierbij wordt de totale steekproefomvang gelijk verdeeld over de L verschillende strata, dus $n_i = n / L$. In het geval waar de standaardafwijkingen per stratum gelijk zijn en de strata ook even groot zijn komt het gelijke allocatieschema overeen met de twee andere schema's. Als dat echter niet zo is dan is het gelijke allocatieschema minder efficiënt dan de andere twee schema's. Qua efficiency ligt dit schema tussen een EAS en het proportionele allocatie schema in, waarbij het verlies van efficiency groter is naarmate de strata qua grootte en homogeniteit verschillen. Het voordeel van het gelijke allocatieschema is dat er voor dit schema geen informatie nodig is over de grootte van alle strata. Als men dus in een situatie zit waarbij de grootte van de strata niet (ongeveer) bekend zijn, dan is het afhankelijk van de verschillen in de strata in termen van de grootte en de homogeniteit of men óf het gelijke allocatie schema moet gebruiken óf dat men eerst energie moet steken in het schatten van de grootte van de strata (en eventueel ook de standaardafwijkingen binnen een stratum) zodat vervolgens het proportionele schema toegepast kan worden (of het Neyman allocatieschema).

Zoals gezegd is er nog een vierde standaard manier om de allocatie over de strata aan te pakken. Dit gaat met het *optimale allocatie schema*. Bij dit schema wordt uitgegaan van een tevoren gegeven budget waarmee de steekproef getrokken moet. In de meest eenvoudige situatie, met een budget ter grootte C , wordt uitgegaan van een lineaire kostenfunctie, $C = c_0 + n_1 c_1 + \dots + n_L c_L$. Hierbij zijn c_0 de vaste kosten die voor het hele onderzoek gelden en is c_i , $i=1, \dots, L$, de kosten van het trekken en waarnemen van één element uit het i -de stratum. Gegeven deze kostenfunctie geldt dat de steekproefomvang per stratum als volgt gekozen dient te worden :

$$n_i = \frac{C - c_0}{\sum_{i=1}^L N_i S_i \sqrt{c_i}} \frac{N_i S_i}{\sqrt{c_i}}$$

De totale steekproefomvang is gelijk aan $n_1 + \dots + n_L =: n$; de totale kosten bedragen precies C (een kwestie van invullen van de formule voor n_i in de lineaire kostenfunctie).

Ter verduidelijking van deze formule zullen we een voorbeeld doorrekenen. Stel er is sprake van drie strata, het totale budget C is 10.000 gulden en daarvan is 2.000 gulden bestemd als vaste kosten voor het onderzoek (c_0). Verder geldt dat de kosten per individu voor het eerste stratum 25 gulden bedragen en voor het tweede en derde stratum 50 gulden. De overige gegevens (standaardafwijking S_i per stratum en de omvang N_i van het stratum) staan hieronder in de tabel gegeven. In de vijfde en zesde kolom van Tabel 5 staan de benodigde tussenresultaten van de berekening. De eerste factor uit de formule voor de optimale verdeling over de strata bestaat uit de teller $C - c_0$, wat in het voorbeeld gelijk is aan 8.000 gulden en uit een noemer die gelijk is aan de som van de vijfde kolom van Tabel 5, 1.383.883,5. Door deze getallen op elkaar te delen krijg je de waarde voor de eerste factor van de formule, 0,0057808. Vermenigvuldiging van deze factor en de zesde kolom van Tabel 5 geeft de steekproefomvang voor het desbetreffende stratum (gegeven in de zevende kolom).

Tabel 5: Voorbeeld voor de berekening van de optimale steekproefomvang bij een vast budget m.b.v. het optimale allocatieschema.

	Kosten per individu c_i	Standaard-afwijking per stratum S_i	Grote van stratum N_i	het $N_i S_i \sqrt{c_i}$	$N_i S_i / \sqrt{c_i}$	n_i
Stratum 1	25	10	10.000	500.000	20.000	116
Stratum 2	50	20	5.000	707106,8	14142,1	82
Stratum 3	50	10	2.500	176776,7	3535,5	20

De steekproefomvang in een stratum wordt groter gekozen naarmate de omvang van het stratum zelf groter is, de homogeniteit kleiner is (dus de standaardafwijking binnen een stratum is groter) en/of de kosten (in termen van c_i) binnen dat stratum kleiner is. Als de kosten van trekking en waarneming per element in een stratum voor de verschillende strata gelijk zijn dan komt het optimale allocatie schema precies overeen met het Neyman allocatie schema.

Als het uitgangspunt van een steekproefomvang berekening niet het beschikbare budget is, maar een minimale nauwkeurigheid die bereikt moet worden dan staan daar minimale kosten tegenover die voor de steekproef nodig zijn. Het kan dus zijn dat gegeven het budget het niet mogelijk is om een bepaalde nauwkeurigheid te garanderen.

De formule voor de optimale allocatie gebruikt o.a. de standaardafwijking binnen een stratum. Omdat de standaardafwijking meestal onbekend is wordt hiervan een grove schatting gemaakt en deze schatting wordt dan gebruikt in de steekproefomvang berekening. Daarom zal de berekende optimale steekproefomvang, hierna de actuele steekproefomvang genoemd, verschillen van de theoretische optimale steekproefomvang. Dit veroorzaakt een verlies in nauwkeurigheid van de schattingen. In termen van het verschil tussen de actuele steekproefomvang en de (theoretische) optimale omvang gelden de volgende richtlijnen: als het verschil 10 % is dan is er een verlies van 1,0 % in de nauwkeurigheid, voor een verschil van 20 % is dit verlies 3,9 %, voor 30 % bedraagt het verlies 8,3 %, voor 50 % daalt de nauwkeurigheid met 20,0 % en een verschil van 100 % geeft een daling van 50 % in de nauwkeurigheid (overgenomen uit Moors (1987)). Kortom, in het geval als het verschil in de actuele en optimale steekproefomvang niet al te fors is dan valt het verlies van nauwkeurigheid mee. Een zelfde plaatje is bekend voor de onzekerheid in de kosten per stratum. Als er een factor van ruim twee (2,25 om precies te zijn) tussen de grootte van de kosten zit per stratum dan levert het gebruik van het Neyman allocatie schema, ten opzichte van het optimale allocatie schema een verlies in relatieve precisie op van 4 %. Als er een factor 4 zit in de grootte van kosten per stratum is de relatieve precisie gelijk aan 88,9 % (Moors (1987)).

We zijn er, behalve bij het optimale allocatie schema, vanuit gegaan dat de totale steekproefomvang vast ligt. In de praktijk is deze ook nog vrij te kiezen, gegeven de totale gewenste nauwkeurigheid d_0 . In tabel 6 staan de formules voor de optimale steekproefomvang uitgesplitst naar de verschillende populatieparameters, voor de verschillende allocatieschema's. Ook hiervoor geldt weer dat men van tevoren een idee of schatting moet hebben van de standaardafwijkingen per stratum. Omdat het in de praktijk niet nodig is om het Neyman allocatieschema te gebruiken zijn de formules voor de optimale steekproefomvang hiervoor niet gegeven. Net als bij Tabel 3 is de eindige populatiecorrectie verwaarloosd. Indien blijkt dat dit incorrect is dan kan formule (5.46) uit Cochran (1977) gebruikt worden.

Tabel 6: totale steekproefomvang gegeven de nauwkeurigheid d_0 en de betrouwbaarheid a in termen van de parameter $u_{\alpha/2}$. De eindige populatiecorrectie is verwaarloosd.

	Populatiegemiddelde	Populatietotaal	Fractie*
Proportioneel allocatie schema	$n = \frac{u_{\alpha/2}^2 \sum_{i=1}^L N_i S_i^2 / N}{d_0^2}$	$n = \frac{u_{\alpha/2}^2 N^2 \sum_{i=1}^L N_i S_i^2 / N}{d_0^2}$	$n = \frac{u_{\alpha/2}^2 \sum_{i=1}^L N_i S_i^2 / N}{d_0^2}$
Gelijk allocatie schema	$n = \frac{u_{\alpha/2}^2 L \sum_{i=1}^L N_i^2 S_i^2 / N^2}{d_0^2}$	$n = \frac{N^2 u_{\alpha/2}^2 L \sum_{i=1}^L N_i S_i^2 / N^2}{d_0^2}$	$n = \frac{u_{\alpha/2}^2 L \sum_{i=1}^L N_i^2 S_i^2 / N^2}{d_0^2}$

3.4 Steekproefgrootte berekeningen bij clustering / systematische steekproeven

De nauwkeurigheid van een schatting hangt af van de betrouwbaarheid, de populatievariantie en de steekproefomvang. Door gebruik van clustering zal de nauwkeurigheid afnemen, wat bijvoorbeeld resulteert in een breder betrouwbaarheidsinterval voor de gezochte grootte zoals het gemiddelde. Deze bewering is af te leiden door te kijken naar twee extreme situaties. Stel dat alle elementen binnen een cluster precies hetzelfde zijn, dan geeft observatie van één element uithet cluster evenveel informatie als observatie van alle elementen uit het cluster. Dit betekent dat we bij observatie van a clusters met elk m elementen evenveel informatie verkrijgen als bij een EAS van a elementen. De effectieve steekproefomvang is dus niet $n = a * m$ maar slechts a . Door deze afname van de (effectieve) steekproefomvang neemt de nauwkeurigheid van de schatters af ten opzichte van een EAS uit de hele populatie zonder stratificatie. De andere extreme situatie is de volgende: stel dat de elementen binnen een cluster evenveel of net zo weinig op elkaar lijken als de elementen uit de andere clusters. Een geclusterde steekproef heeft dan dezelfde eigenschappen als een EAS. Daarom is de effectieve steekproefomvang van een geclusterde steekproef gelijk aan $n = a$. Dit betekent dat, ten opzichte van een EAS de nauwkeurigheid hetzelfde is.

De praktijk zal tussen de twee extreme situaties in zitten. Er zal altijd een zekere mate van homogeniteit zijn tussen de elementen binnen een cluster. En daarom zal de nauwkeurigheid lager zijn bij een geclusterde steekproef dan bij een EAS. Naarmate de homogeniteit binnen een cluster groter is, ten opzichte van de spreiding tussen de clusters, zal men meer waarnemingen nodig hebben om dezelfde nauwkeurigheid te bereiken.

* De formules voor de optimale steekproefomvang voor fracties zijn hetzelfde als die voor het populatiegemiddelde. De standaardafwijking in de formule, S_i , is in dit geval echter gelijk aan $N_i P_i Q_i / (N_i - 1)$, waarbij P_i de fractie in stratum i is met de onderzochte eigenschap; Q_i is de fractie elementen in stratum i dat niet die eigenschap heeft.

Een maat voor de homogeniteit binnen een cluster is de intraklasse correlatie (ook wel rate of homogeneity genaamd, zie Kish [1965], blz. 161). Dit is de correlatiecoëfficiënt tussen elementen uit hetzelfde cluster. Voor de effectieve steekproefgrootte n_{eff} geldt de volgende relatie tussen de intraklasse correlatie ρ , de steekproefomvang n , het aantal clusters a , de clusteromvang m :

$$n_{eff} = (N-1) (1 + (m-1) \rho) / (m(a-1)) \approx n / [1 + (m-1)\rho].$$

Dit betekent dat we bij een geclusterde steekproef, als we een even effectieve (bijvoorbeeld met dezelfde nauwkeurigheid) steekproef willen als bij een EAS (met steekproefgrootte n), we voor de geclusterde steekproef een steekproef ter grootte $n \cdot (1 + (m-1) \rho)$ moeten selecteren. Als de clusters niet even groot zijn, maar toch ook niet al te extreem qua grootte verschillen dan kun je voor a de gemiddelde clustergrootte invullen. Bij een hoge intraklasse correlatie zal de effectieve steekproef dramatisch kunnen afnemen, zoals dadelijk zal blijken uit een voorbeeld. De mate waarin de effectieve steekproefomvang afneemt hangt ook af van de clusteromvang. Als deze niet al te groot is ten opzichte van de steekproefomvang dan zal het verlies in effectieve steekproefomvang mee kunnen vallen.

Een correlatiecoëfficiënt kan ook negatief zijn. Een negatieve intraklasse correlatie betekent dat de elementen uit een cluster inhomogeen zijn. De meest extreme situatie is een aantal precies dezelfde clusters die elk bestaan uit verschillende elementen. In dat geval is de intraklasse correlatie minimaal. Als de intraklasse correlatie negatief is dan is de effectieve steekproefomvang dus juist groter dan n , omdat het methode-effect, oftewel de factor $1 + (m-1) \rho$, kleiner dan 1 is. In de praktijk komt dit weinig voor.

Tabel 7 geeft een indruk van de gevolgen van geclusterde steekproeven. We gaan uit van het voorbeeld van onderzoek in scholen. Stel er zijn 25 leerlingen per klas. En er zijn 500 leerlingen in totaal die getest worden. Dit betekent dat $m = 25$ (aantal leerlingen per klas), $n = 500$ (totaal aantal leerlingen in het onderzoek) en dus $a = 20$ (=aantal klassen). Voor dit voorbeeld blijkt dus o.a. dat bij een intraklasse correlatie van 0,25 de effectieve steekproefomvang slechts een zevende is van de oorspronkelijke steekproefomvang. Het verschil in de gekozen steekproefomvang en de effectieve steekproefomvang is de prijs die er betaald moet worden voor de clustering. In de laatste twee kolommen wordt het methode-effect en de effectieve steekproefomvang gegeven als er maar 12 leerlingen per klas zitten. Door de situatie van 25 leerlingen per klas te vergelijken met die van 12 leerlingen per klas kunnen we concluderen dat hoe kleiner de clusters zijn des te minder groot het verlies is in effectieve steekproefomvang.

Tabel 7: De effectieve steekproefomvang in het geval van 20 klassen, 25 leerlingen per klas ($n = 500$).

	25 leerlingen per klas		12 leerlingen per klas	
	Methode effect ($1+(m-1) \rho$)	Effectieve steekproef	Methode effect ($1+(m-1) \rho$)	Effectieve steekproef
$\rho = 0.075$	2,8	179	1,8	277
$\rho = 0.10$	3,4	147	2,1	238
$\rho = 0.15$	4,6	109	2,65	188
$\rho = 0.25$	7	71	3,75	133

Een systematische steekproef kun je opvatten als een geclusterde steekproef waarbij er slechts één cluster getrokken wordt uit $F = N/n$ clusters van elk omvang n . Voor een systematische steekproef geldt daarom dat de effectieve steekproefomvang, bij het schatten van het gemiddelde, gelijk is aan $(1+(n-1) \rho) (N-1) / (N-n)$. Dit betekent dat systematische steekproeven een hogere effectieve steekproefomvang hebben dan een EAS ter grootte n , indien

$$-1 / (n-1) \leq \rho < -1 / (N-1).$$

Aan deze voorwaarde is voldaan als de variantie binnen de systematische steekproef groter is dan de populatievariantie. Als voorbeeld kun je denken aan de situatie als de onderzoeksvariabele bij benadering lineair afhankelijk is van het rangnummer in het steekproefkader. In zo'n geval komen er zowel grote als gemiddelde en kleine waarden van de onderzoeksvariabelen in de steekproef voor, zelfs als de meeste waarden van de onderzoeksvariabele vlakbij het gemiddelde liggen. Het getrokken cluster is dan relatief veel heterogener dan de populatie in het algemeen, wat resulteert in nauwkeurigere schattingen dan als er een EAS getrokken zou zijn*. Dit is ook de reden dat een systematische steekproef een goede keuze is bij dosis-effect relaties als de volgorde in het steekproefkader samenhangt met de dosisvariabele.

Alle formules die tot nu toe gegeven zijn, zijn alleen maar toepasbaar als de intraklasse correlatie bekend is, of als er een schatting van gemaakt kan worden. Indien dit niet het geval is kun je door verschillende waarden van ρ in te vullen in de formules wel een indruk krijgen van de invloed van de intraklasse correlatie op de effectieve steekproefomvang.

Voor een geclusterde of systematische steekproef bestaat de bepaling van de optimale steekproefomvang dus uit twee stappen (voor gegeven betrouwbaarheid en nauwkeurigheid)

* Als het getrokken cluster homogener is dan de populatie in het algemeen (dus $\rho > 0$) dan is een systematische steekproef minder nauwkeurig dan een EAS.

1. Bepaal met behulp van de formules in paragraaf 3.2 wat voor steekproefomvang n je zou kiezen als je een EAS zou hebben.
2. Bereken wat de effectieve steekproefomvang n_{eff} is met behulp van de hierboven gegeven formules. Kies de uiteindelijke steekproefomvang gelijk aan $n * n / n_{eff}$.

3.5 Optimale steekproefomvang bij getrapte steekproeven

Voor getrapte steekproeven behandelen we alleen het geval van een tweetrapssteekproef. Hierbij gaan we ervan uit dat de populatie opgedeeld is in L primaire eenheden die elk weer bestaan uit M secundaire eenheden. Denk bijvoorbeeld aan ziekenhuizen als primaire eenheden en patiënten als secundaire eenheid. De steekproef zal bestaan uit l geselecteerde primaire eenheden waarbinnen weer m secundaire eenheden per primaire eenheid gekozen wordt. We gaan ervan uit dat de steekproefomvang berekend moet worden onder de randvoorwaarde van een gegeven budget ter grootte C .

De optimale steekproefomvang bij getrapte steekproeven hangt af van de gebruikte kostenfunctie. Een eenvoudig en vaak gebruikte kostenfunctie is $C = c_1 * l + c_2 * l * m$, waarbij c_1 de kosten om een extra primaire eenheid op te nemen in de steekproef en c_2 de kosten om een extra secundaire eenheid op te nemen in de steekproef zijn. De optimale steekproefomvang, die de nauwkeurigheid van het gemiddelde maximaliseert, is gelijk aan (zie het artikel van Cohen (1998)):

$$m_{opt} = \sqrt{\frac{c_1}{c_2} \times \frac{1-\rho}{\rho}}.$$

Hierbij is ρ de intraklasse correlatie van de elementen binnen de primaire eenheid*. Combinatie van de kostenfunctie en de optimale omvang van de secundaire eenheid geeft dat het optimaal aantal primaire eenheden gelijk is aan $C / (c_1 + c_2 * m_{opt})$.

Ter illustratie van de formule: Stel dat de verhouding tussen c_1 en c_2 gelijk is aan 10. Dit betekent dat het tien keer zo duur is om een extra primaire eenheid aan de steekproef toe te voegen dan een secundaire eenheid. Veronderstel verder dat de intraklasse coëfficiënt gelijk is aan 0,25. Dan volgt uit de formule dat het optimaal is om vijf á zes secundaire eenheden te samplen. Indien het totale budget 10,000 gulden bedraagt, en de kosten voor een extra primaire eenheid 100 gulden is (wat betekent dat de kosten voor een extra secundaire eenheid 10 gulden is) dan is het aantal primaire eenheden gelijk aan $10,000 / (100 + 10 * 5) = 66,7$. Indien de kosten voor een extra secundaire eenheid niet tien maar één gulden bedragen is de verhouding tussen de kosten voor

* Indien de intraklasse correlatie niet (ongeveer) bekend is maar wel de verhouding van de standaard afwijking binnen de secundaire eenheid (genoteerd met v) en de primaire eenheid en het aantal primaire eenheden in de populatie (L) dan geldt bij benadering dat $(1-\rho)/\rho$ gelijk is aan $v^2/(1-v^2/L)$ (zie Cochran (1977), pagina 282).

een extra primaire eenheid en een secundaire eenheid gelijk aan 100. In dat geval is het optimale aantal primaire eenheden (bij een gelijkblijvende intraklasse correlatie van 0,25) gelijk aan 17 en het aantal primaire eenheden (bij gelijkblijvend totaal budget van 10,000 gulden) is gelijk aan 85.

In de berekening van de optimale steekproefomvang zoals hierboven is uitgegaan van een vast totaal budget wat niet overschreden mag worden. Hoe lager het budget is, des te minder nauwkeurig zullen de schattingen zijn en des te breder de betrouwbaarheidsintervallen die uitgerekend worden. Het kan dus ook zo zijn dat gegeven het budget het niet mogelijk is om een bepaalde nauwkeurigheid te garanderen.

3.6 Relatie tussen nauwkeurigheid en power

In dit hoofdstuk zijn de steekproefomvang berekeningen gegeven als functie van een vooraf afgesproken nauwkeurigheid. Deze aanpak is klassiek voor beschrijvend onderzoek. In toetsend onderzoek is het gangbaar om de optimale steekproefomvang te bepalen aan de hand van een minimale power. De *power* (de eigenlijke Nederlandse term is “onderscheidingsvermogen”) is de kans om statistische significantie te bereiken wanneer de alternatieve hypothese inderdaad juist is.

De reden voor de tweespalt in aanpak bij optimale steekproefomvang berekeningen is te wijten aan het volgende: bij beschrijvend onderzoek is men (meestal) geïnteresseerd in schattingen van populatieparameters als gemiddelden, aantallen en proporties. Om de onzekerheid in de schattingen hiervan aan te geven worden naast de schattingen zelf ook betrouwbaarheidsintervallen geconstrueerd. Hoe smaller de intervallen, des te nauwkeuriger de schattingen van de parameters zijn. Daarom wordt de optimale steekproefomvang in termen van de nauwkeurigheid (dit is helft van de breedte van het betrouwbaarheidsinterval) berekend. Indien voor de betrouwbaarheid een α van 0,05 gekozen is dan is de nauwkeurigheid de helft van de breedte van het 95 %-betrouwbaarheidsinterval. In toetsend onderzoek ligt de nadruk op het toetsen van hypothesen die betrekking hebben op (populatie)parameters. In de toetsingstheorie is het gangbaar om toetsen te construeren op basis van een significantieniveau α en de power $1-\beta$. Het *significantieniveau* van een toets is de kans op verwerpen van de nulhypothese als deze waar is. Het significantieniveau is een equivalent van de betrouwbaarheid van een betrouwbaarheidsinterval.

De standaard nulhypothese voor een populatieparameter θ is $H : \theta = \theta_0$. Neem als voorbeeld twee patiëntengroepen van hartpatiënten. Eén groep ondergaat een behandeling en de andere groep krijgt een placebo medicijn. De uitkomstvariabele is de bloeddruk. De nulhypothese zou dan zijn: er is geen verschil tussen de groepen qua bloeddruk, dus de behandeling heeft geen effect. θ is in dit geval het verschil in bloeddruk tussen de twee groepen en θ_0 is gelijk aan nul (geen verschil). De alternatieve hypothese zou als volgt kunnen zijn: $H : \theta > \theta_0$. In ons voorbeeld betekent dit niets anders dan: de bloeddruk van de behandelingsgroep is lager dan die van de controle groep, oftewel er is een positief effect waargenomen wegens de behandeling. De power van de toets die gebruikt wordt om de nulhypothese te verwerpen hangt af van het gekozen

alternatief. Hoe groter het verschil is, in bijvoorbeeld bloeddruk, dat je minimaal wilt aantonen des te groter zal de power van de bijbehorende toets zijn omdat grote verschillen nu eenmaal makkelijker aan te tonen zijn dan kleine verschillen. Andersom: hoe kleiner het verschil is dat je wilt aantonen des te groter de steekproefomvang moet zijn om een zelfde power te garanderen. In de praktijk wordt vaak uitgegaan van een power van 0,80 en wordt dan vervolgens, gegeven het minimale verschil dat men wil aantonen, de bijbehorende steekproefomvang berekend.

Redenaties ten aanzien van de optimale steekproefomvang met behulp van het onderscheidingsvermogen gaan op dezelfde manier als met een vast gekozen nauwkeurigheid. Voor standaard situaties wordt in Lemeshow, Hosmer, Klar en Lwanga (1990) aanwijzingen gegeven hoe je de juiste steekproefomvang kan berekenen, met behulp van handige tabellen. Voor ingewikkeldere situaties met bijvoorbeeld stratificaties en clusteringen kun je dezelfde principes hanteren die we in de vorige paragrafen behandeld hebben.

3.7 Samenvatting hoofdstuk 3

Dit hoofdstuk geeft voor de diverse steekproefontwerpen handvaten en formules om de benodigde steekproefomvang te berekenen. Voor een steekproefomvang berekening zijn meestal de volgende ingrediënten nodig:

- De gewenste *betrouwbaarheid* α , meestal wordt hiervoor 0,05 gekozen. Dat betekent dat 1 op de 20 uitspraken die je doet niet correct is. Indien er altijd 95 % betrouwbaarheidsintervallen worden gegeven moet men er van uit gaan dat 1 op de 20 betrouwbaarheidsintervallen niet de gezochte parameter bevat.
- De gewenste *nauwkeurigheid* d . De nauwkeurigheid geeft de mate van precisie aan waarmee de uitspraken gedaan worden. Hoe groter de nauwkeurigheid, des te kleiner zal het betrouwbaarheidsinterval zijn.
- De *standaardafwijking* S . Dit is een maat voor de spreiding van de doelvariabele in de populatie. Meestal is ze onbekend. Voor de meeste verdelingen van de dat geldt dat de standaardafwijking afgeschat kan worden door 0,25 keer de spreidingsbreedte (het databereik) van de doelvariabele. Voor prevalenties, proporties en percentages is de standaardafwijking niet nodig om te kennen.

Gegeven de betrouwbaarheid, de nauwkeurigheid en de standaardafwijking kunnen met de formules uit dit hoofdstuk (of met diverse beschikbare software) voor diverse situaties de benodigde steekproefomvang berekend worden. Men dient wel de structuur van de populatie en van de steekproef in ogenschouw te nemen. Een steekproefomvang berekening staat nooit op zichzelf en is van hoge mate afhankelijk van de onderliggende structuur in de populatie en het gekozen steekproefontwerp.

Indien er in het steekproefontwerp gekozen is voor een geclusterde steekproef dan zal de effectieve steekproefomvang, afhankelijk van de intraklasse correlatie, veel lager kunnen uitvallen dan de oorspronkelijke steekproefomvang.

Een powerberekening is een berekening van de benodigde steekproefomvang om een bepaald onderscheidingsvermogen te kunnen garanderen. In feite gaat de manier waarop een powerberekening uitgevoerd dient te worden analoog aan een steekproefomvang berekening waarbij een bepaalde nauwkeurigheid gegarandeerd wordt. Bij beschrijvend onderzoek zal men veelal een nauwkeurigheid vereisen. Toetsend onderzoek gaat vaak uit van een minimaal onderscheidend vermogen.

4 Uitwerking cases

4.1 Case 1, Peiling Milieuhinder: Utrecht 1997

In verband met het opstellen van het Provinciale Milieubeleidsplan 1998-2002 en de evaluatie van het uitgevoerde beleid heeft de provincie Utrecht onderzoek laten verrichten naar hinder die in de provincie bij het wonen en het recreëren wordt ondervonden door milieufactoren. Deze peiling van milieuhinder is in 1997 uitgevoerd door TNO PG (zie van Dongen & Vos[2]). Onderzocht zijn de omvang en mate van verstoring door geluid, geur, trillingen, stof en licht. Ook is gekeken naar de bezorgdheid over een aantal veiligheidsrisico's

De uitkomsten van het onderzoek worden weergegeven voor de provincie als geheel en voor de vier regio's, Stadsgewest Utrecht, Stadsgewest Amersfoort, West Utrecht en Zuid-Oost Utrecht afzonderlijk. De doelpopulatie van de studie betrof inwoners van de provincie Utrecht van 18 jaar en ouder. Omdat de interesse ook uitging naar uitkomsten uitgesplitst naar de vier regio's ligt een steekproefontwerp, gestratificeerd naar regio, voor de hand. Dit is dan ook gebeurd.

Het onderzoek is verricht door middel van een telefonische enquête. Als steekproefkader is het PTT telecom bestand genomen waaruit aselekt een steekproef is getrokken binnen 4-cijferige postcodegebieden in de provincie Utrecht. In feite is er dus op twee niveaus gestratificeerd: eerst per regio en vervolgens op postcodegebied binnen de regio's. Per postcodegebied of gemeente is er een taakstelling, afhankelijk van het aantal inwoners per gemeente, gehanteerd voor wat betreft het aantal geslaagde enquêtes dat bereikt moest worden.

De gesprekken zijn in principe gevoerd met de eerst jarige persoon van 18 jaar of ouder binnen het huishouden behorende bij het getrokken telefoonnummer. Als het telefoongesprek niet slaagde is een volgend nummer van de lijst gebeld. Dit resulteerde in een non-respons van 66 % (3117 getrokken telefoonnummers bestaande uit 1074 respondenten, 1002 weigeringen voor het gesprek, bij 301 huishoudens was de juiste persoon niet aanwezig, 391 nummers waren in gesprek of gaven geen gehoor en de overige 309 nummers vielen af wegens andere redenen). Volgens een steekproefdeskundige van het CBS (persoonlijke communicatie met Dr. R. Renssen, CBS Heerlen) hanteert het CBS bij dit soort telefonische enquêtes de richtlijn dat middels een aantal (6 of meer komt in de praktijk zeker voor) herhaalde pogingen men moet proberen om de gezochte persoon (de eerst jarige van 18 jaar of ouder) te pakken te krijgen. Hierbij wordt geprobeerd de "pakkans" te vergroten door de tijd van telefoneren te variëren. Als er slechts een enkele poging ondernemen wordt kan de steekproef selectief zijn, er zullen meer mensen die vaak thuis zijn in de steekproef voorkomen dan er in de populatie voorkomen. Daarnaast is het gebruikte steekproefkader niet dekkend voor de doelpopulatie omdat er uitgegaan wordt van mensen met een (niet geheim) telefoonnummer dat voorkomt in het PTT telecom bestand. Aan de andere kant is het aantal huishoudens zonder telefoon dalend.

Omdat uit eerder onderzoek is gebleken dat leeftijd een factor is die van invloed kan zijn op hinder zijn de resultaten uit het onderzoek gewogen naar leeftijd. Het blijkt namelijk dat er verschillen bestaan, qua leeftijdsopbouw, tussen de steekproef en CBS-gegevens over ervaren hinder. De leeftijdsgroep 35 tot en met 49 jaar is oververtegenwoordigd en de 18 tot 24 jarigen zijn ondervertegenwoordigd. Vooraf had men natuurlijk kunnen besluiten om te stratificeren naar leeftijd, maar dan had het aanzienlijk veel meer moeite gekost om de steekproef te verkrijgen, omdat er geen steekproefkader aanwezig is waarin gestratificeerd is naar leeftijd.

Aan de hand van de procedures in Hoofdstuk 3 bepalen we de optimale steekproefomvang per stratum. We concentreren ons op het niveau van de regio's omdat de resultaten van de studie ook op dit niveau verlangd werden. De studie heeft als doel om de omvang en mate van verstoring door geluid, geur, trillingen, stof en licht te onderzoeken in termen van percentages "erg gehinderden", "gehinderden" en "in enigerlei mate gehinderden". De respondent kon bij beantwoording van de vragen kiezen op een vier-puntsschaal. Als kosten geen, of niet van doorslaggevende rol zijn, of als de kosten niet te sterk variëren per regio, dan is Neyman allocatie de meest efficiënte manier om de benodigde interviews te verdelen over de regio's. Indien de onderzoeksparameters percentages (proporties, prevalenties) zijn is gebleken dat het verschil in efficiëntie tussen Neyman allocatie en evenredige allocatie minimaal is. Daarom voldoet een evenredige allocatie over regio's naar inwoneraantallen.

Uit Tabel 4, paragraaf 3.2, blijkt dat voor een onbetrouwbaarheid α van 0,05 en een nauwkeurigheid van 0,05 een steekproefomvang van 405 mensen voldoende is bij een populatieomvang van minstens 100.000 en dat laatste is het geval in de provincie Utrecht. Voor een onbetrouwbaarheid en een nauwkeurigheid van elk 0,01 zullen er minstens 14.314 interviews nodig zijn. Nu is het zo dat deze aantallen gebaseerd zijn op de "worst-case" situatie dat de gezochte percentages in de buurt van de 50 % liggen. De gezochte percentages gehinderden zullen meestal ruim lager liggen dan de 50 % en dus zullen uitspraken hierover nauwkeuriger uitvallen.

Tabel 8: Inwonertallen Provincie Utrecht, uitgesplitst naar regio. (bron: CBS. Statistisch Jaarboek 1999).

Regio	Gemeenten	Inwoneraantal
Stadsgewest Utrecht	Bunnik, De Bilt, Vleuten-De Meern, Houten, Maarssen, Maartensdijk, Nieuwegein, Utrecht, IJsselstein en Zeist	526.743
Stadsgewest Amersfoort	Amersfoort, Baarn, Bunschoten, Eemnes, Leusden, Soest en Woudenberg	255.011
Zuid-Oost Utrecht	Amerongen, Doorn, Driebergen/Rijsenburg, Leersum, Maarn, Renswoude, Rhenen, Veenendaal en Wijk bij Duurstede	151.078
West Utrecht	Abcoude, Breukelen, De Ronde Venen, Harmelen, Loenen, Loosdrecht, Lopik, Montfoort, Oudewater en Woerden	155.789

We gaan er vanuit dat een onbetrouwbaarheid en nauwkeurigheid van 0,05 voldoende is. In Tabel 7 is af te lezen hoe de inwoners over de provincie Utrecht verdeeld zijn naar regio. In deze aantallen inwoners zijn echter niet de personen onder de 18 jaar weggelaten. De informatie ontbreekt

dus over hoe groot het aantal volwassenen is per regio. We gaan er, voor de eenvoud, even van uit dat het percentage volwassenen per regio gelijk is. Volgens het statistisch jaarboek geldt voor de provincie Utrecht dat 24,7 % van de inwoners tot de categorie 0-19 jarigen behoort. Dat betekent dat het aantal inwoners per regio van 18 jaar of ouder in ieder geval minstens 100.000 bedraagt. Omdat we veronderstellen dat het percentage volwassenen per regio gelijk verdelen we de totale optimale steekproefomvang van 405 in verhouding naar de inwonersaantallen per regio over de regio's: Stadsgewest Utrecht 196 (48,4 %), Stadsgewest Amersfoort 95 (23,4 %), Zuid-Oost Utrecht 56 (13,9 %) en West Utrecht 58 (14,3 %) (tussen haakjes staan de verhoudingen).

De verhoudingen die in het onderzoek zijn genomen komen overeen met de optimale verhoudingen (Stadsgewest Utrecht 47,7 %, Stadsgewest Amersfoort 23,0 %, Zuid-Oost Utrecht 14,3 % en West Utrecht 15,0 %). Het verschil is waarschijnlijk te danken aan enerzijds het feit dat we uitgaan van de inwonersaantallen van 1999 i.p.v. 1997 en anderzijds de aanname over gelijkheid van leeftijdsverdeling per regio. Bij een aantal van 1074 respondenten, zoals in het onderzoek het geval was, geldt dat de minimale nauwkeurigheid (bij $\alpha = 0,05$) gelijk is aan 0,0306.

Omdat er ook uitspraken gedaan zijn op het niveau van regio is het interessant om de nauwkeurigheid per regio naast elkaar te zetten, zie Tabel 9.

Tabel 9: De minimale bereikte nauwkeurigheid in het onderzoek bij een betrouwbaarheid van 5 %.

Regio	Aantal respondenten	Minimale Nauwkeurigheid, (betrouwbaarheid van 5 %)*
Stadsgewest Utrecht	512	0,0443
Stadsgewest Amersfoort	247	0,0645
Zuid-Oost Utrecht	161	0,0805
West Utrecht	154	0,0824
Provincie Utrecht	1074	0,0306

Als het uitgangspunt was geweest om per regio een nauwkeurigheid van 0,05 te garanderen dan had men per regio minstens 405 respondenten moeten selecteren. En dat had geresulteerd in een grotere totale steekproef (en daarmee een hogere nauwkeurigheid voor de uitspraken op geaggregeerde niveau van de provincie).

* Dit betekent dat de 95 % betrouwbaarheidsintervallen voor uitspraken over percentages per regio hoogstens een breedte hebben van twee keer de gegeven nauwkeurigheid (zie Paragraaf 3.1).

4.2 Case 2, Slaapverstoringsonderzoek Schiphol

In het kader van het onderzoeksprogramma Gezondheidskundig Evaluatie Schiphol (GES) zal een slaapverstoringsonderzoek worden verricht rond Schiphol. De studie heeft als doel om informatie te verkrijgen omtrent

- De relatie tussen blootstelling aan vliegverkeersgeluid in de nacht enerzijds en slaapverstooring, gezondheid en dagelijks functioneren anderzijds. Tevens zal ook het effect van de periode in de nacht bestudeerd worden.
- Informatie die nodig is om de, door nachtelijk geluid veroorzaakte effecten op de bevolking in de omgeving van Amsterdam Schiphol Airport te kunnen schatten.

Om deze twee doelstellingen te verwezenlijken is eerst een pilotstudy uitgevoerd waarbij gegevens verzameld zijn om de onderzoeksopzet van de hoofdstudie te onderbouwen (zie Passchier-Vermeer et al [4]). Zo zijn bijvoorbeeld gegevens verzameld die nodig zijn om een poweranalyse uit te kunnen voeren voor de onderzoeksopzet van de hoofdstudie.

In de pilotstudie is een aantal verschillende relaties bestudeerd tussen verschillende expositievariabelen en effectparameters. De powerberekeningen in de studie zijn gebaseerd op twee typen van deze relaties. Dit zijn:

1. Relaties tussen parameters voor slaaperiode tijden (zie Passchier-Vermeer et al (1998), pagina 28):
 - “Equivalent sound level due to aircraft noise during sleep period time (Laeqslp) and perceived sleep quality”.
 - “Equivalent sound level due to aircraft noise during sleep period time (Laeqslp) and average value of mscore (MGEMOR) during sleep period time”.
 - “Number of aircraft noise events during sleep period time (nslp) and perceived sleep quality”.
 - “Number of aircraft noise events during sleep period time (nslp) and average value of mscore (MGEMOR) during sleep period time”.
2. Relatie voor de actometrische data verzameld op een continue schaal (zie Passchier-Vermeer et al (1998), pagina 29): “effects of aircraft events with $SEL(\text{indoors}) \geq 50 \text{ dB(A)}$ on mscore over an observation period of 3.0 minutes”.

Relatie type 1 is een dosis-effect relatie. De steekproef van de pilotstudie betrof een aantal metingen per nacht per proefpersoon, verspreidt over in totaal twee locaties in de gemeenten Rijsen-

hout en Spaarndam. Interesse gaat uit naar een dosis-effect relatie die algemeen is voor de populatie rond Schiphol. Omdat de relatie per persoon anders is, maar er wel sprake is van een soort gemeenschappelijk verband, is het aannemelijk om een multilevel model te veronderstellen waarbij zowel de intercept als de richtingscoëfficiënt van de dosis-effect relatie naast een “fixed part” ook een “random part” heeft. Dit is een geavanceerd model, waarbij het bijbehorende steekproefontwerp eigenlijk een tweetrapssteekproef is. Als we de stratificatie over de verschillende locaties even buiten beschouwing laten dan kun je de personen opvatten als primaire eenheden en de metingen per nacht per persoon zijn de secundaire eenheden. De eenheden binnen een tweetrapssteekproef komen overeen met de “levels” in een multilevel model. Voor de powerberekening hiervoor hebben we gebruik gemaakt van het programma PINT (gebaseerd op het artikel Snijders en Bosker (1993)). Dit programma berekent de optimale steekproefomvang voor multilevel modellen met twee levels, waarbij het model bestaat uit een lineair verband tussen de dosis en de response, dus als er sprake is van een intercept en een richtingscoëfficiënt.

Het programma PINT berekent, gegeven een van tevoren opgegeven budget, correlaties tussen de intercept en de richtingscoëfficiënt en de varianties binnen en tussen de levels, een optimale steekproefomvang en de bijbehorende power. Het budget is gebaseerd op een lineaire kostenfunctie, waarbij het totale budget van belang is en de verhouding tussen de kosten van het toevoegen van een extra persoon ten opzichte van het toevoegen van een extra nacht per persoon in de steekproef. Omdat het beschikbare budget een randvoorwaarde is bij de bepaling van de optimale steekproefomvang is het niet altijd mogelijk om een bepaald niveau van de power (standaard 80 %) te bereiken. Hoe hoger het budget, des te hoger zal de power zijn omdat er meer waarnemingen genomen kunnen worden.

Bij het tweede type relatie is het doel om een verschil aan te tonen in *mscore* (het gemiddelde over drie minuten van de output van de actometer, die elke 15 seconde een waarde geeft) tussen de bijbehorende tijden over de nacht met geen of wel een “aircraft event” (er is sprake van een “aircraft event” als $SEL(\text{indoors}) \geq 50 \text{ dB(A)}$). Net als bij de vorige onderzochte relaties is er sprake van een tweetrapssteekproef: de primaire eenheden zijn de personen, de secundaire eenheden zijn de nachten waarbij gemeten is per persoon. Vervolgens zijn we uitgegaan van een random effects model, wat een speciaal soort multilevel model is. In feite kun je in deze situatie dezelfde formule gebruiken als voor het populatiegemiddelde gegeven is in Tabel 3, Paragraaf 3.2. Het enige verschil zit in de variantie die geldt voor dit geval. Die bestaat uit de variantie per persoon over de nachten (de binnenpersoons variantie) en de variantie tussen de verschillende personen (tussenpersoonsvariantie). We gaan hier nu niet dieper op in, maar de belangrijkste boodschap uit dit voorbeeld is het feit dat gegeven de structuur die in de gegevens zit (verschillende niveaus) kun je in principe niet de standaardmethoden toepassen. Het multilevel model is in zo'n geval een goed model om de gegevens te analyseren. Omdat dit type model vaker gebruikt wordt verschijnen er in de literatuur ook steeds meer artikelen over hoe een adequate powerberekening aangepakt kan worden voor verschillende situaties met zo'n hiërarchische structuur in de data.

4.3 Case 3, Dosis-effect relaties wegverkeer en geluidshinder in de EU

In opdracht van de Europese Commissie zal (naar verwachting) een studie verricht worden naar de dosis-effect relaties van wegverkeer en geluidshinder ervaren door omwonenden in de verschillende landen van de Europese Unie.

Wegverkeer is er in alle soorten en maten, zo ook de geluidshinder veroorzaakt door wegverkeer. Geluid veroorzaakt door verkeer op snelwegen is qua sterkte regelmatig en van een niet veel variërend niveau. Geluid veroorzaakt door verkeer in de buurt van lokale wegen met kruispunten, stoplichten e.d. daarentegen is afwisselend in niveau en vooral minder regelmatig. Hierbij heb je bijvoorbeeld te maken met optrekkende of remmende (vracht)auto's. De mate van hinder door wegverkeer bij de bevolking zal o.a. afhangen van de afstand tot de wegen. Daarnaast zijn er nog andere aspecten die een rol spelen. Zo zal in de winter, als alle ramen dicht zijn, men veel minder last hebben van geluid door wegverkeer, dan in de zomer als alle deuren open staan of het leven zich grotendeels buiten afspeelt. Dit zal ook variëren per land, omdat per land ook de gewoonten verschillen t.a.v. de mate waarin het dagelijkse leven zich binnen of buiten afspeelt. De isolatie en dus het dimmen van het geluid zal ook per woning en zeker ook per land verschillen.

Het voorlopige voorstel dat verwerkt is in de aanvraag voor het vijfde kaderprogramma is als volgt (in het Engels):

The sampling plan is based on the (expected) noise exposure degree, demographic factors like urbanization degree and geographical factors. For each country that participates in the study exposure-response relations for the different transportation noise sources will be studied. Therefore, for each country, sampling designs should be provided for all transportation noise sources.

In general the sampling strategy is as follows: Depending on the noise source, a basic geographical unit is chosen. Geographical units will be selected in the sample with weights proportional to the number of inhabitants. Next, for each selected unit, a map is needed on which the network of the particular traffic source is displayed. Based on this network the geographical units are divided into smaller areas and, proportional with the population density, areas are selected for the sample. The final step is the selection of households in the areas. The sampling scheme is visualized in figure 4.

The traffic sources are split into the following four categories:

1. Air traffic
2. Train traffic
 - A. Road traffic on highways
 - B. Road traffic on inter local roads

3. Other road traffic (in cities with more than 25.000 inhabitants)

Category four is the most important noise source since the largest number of people are exposed to it. Therefore the sample sizes for this category are larger than for the other categories in order to obtain more precise estimates for this category. Remark that people usually are not exposed to all or some of the categories simultaneously.

Countries, incorporated in the sampling design are not chosen on basis of randomization but on the practical reason of cooperation in the study. Extrapolation to other countries of the EU is possible if necessary information on important variables is available and under the assumption that there is no other systematical difference present that is not taken into account by the explaining variables.

Above the general sampling scheme is discussed, now we describe the details for each traffic source, starting with air traffic.

For air traffic noise we restrict ourselves to airports with large traffic (more than 6000 kg) and to those having a ground control. For example, for the Netherlands this means that only Schiphol Airport, Rotterdam Airport, Maastricht Airport, Groningen Airport and Twente Airport are considered. Based on the number of air movements airports are separated into two categories: the busy airports, that is, airports with more than 100.000 air movements a year, and the other airports, thus with less than 100.000 air movements a year.

Table 10: Airports in the Netherlands and the number of movements with large traffic.

Airport	Schiphol	Rotterdam	Maastricht	Groningen	Twente
Large traffic	368.000	29.904	16.550	12.294	?

From each category an airport is chosen, which is the basic geographical unit. For the other traffic categories the units are selected based on population sizes, but for airports there is no use for doing so, since in general airports are in the neighborhood of highly urbanized areas. The area around the selected airports is divided into smaller regions, which are sampled proportional to the population density and the noise exposure degree. Finally in these regions households are selected.

For train traffic the same strategy of sampling is followed, but at a lower level for the basic geographical unit (to save costs and time). Each country is split into natural regions, which are taken as the basic geographical units. For example, Spain consists of Catalonia, Andalusia, Castile la Mancha, Galicia etcetera. The Netherlands, e.g., is divided into regions called Friesland, Groningen, Drente, and Utrecht etcetera. A few regions are selected with probability proportional to the number of inhabitants of the region (regions with large populations are more important). In Table 11 figures are given for the Netherlands.

In each selected region a map is made of the network of railways. The railway net is divided into pieces of one kilometer. From the one-kilometer pieces a number are selected, proportional with the population density around the rail. Next, geographical areas around the selected railpieces, are defined proportionally to the traffic intensity on the rail (based on the number of rails, night traffic intensity, amount of goods traffic, number of trains averaged over a day). In these areas households are selected and noise exposure is measured.

Tabel 11 Regions of the Netherlands with population size and density (1 january 1998).

Region	Size of population (x 1000)	Population density (per km ²)
Groningen	558,0	238
Friesland	618,1	184
Drenthe	464,7	175
Overijssel	1063,5	319
Flevoland	293,3	206
Gelderland	1895,7	380
Utrecht	1088,6	798
Noord-Holland	2486,1	935
Zuid-Holland	3359,0	1172
Zeeland	369,9	207
Noord-Brabant	2319,3	470
Limburg	1137,9	525
The Netherlands	15654,2	462

For the road traffic on (inter) national roads the sampling plans are identical to train traffic. The maps are made on the same level of geographical basic unit.

The last source of traffic is road traffic, not on (inter) national roads, but in cities. Here the basic geographical units will be cities with more than 25.000 inhabitants. Cities will be selected proportional with the number of inhabitants. From the selected cities a map is made of the traffic. Sub-regions of the city are selected based stratified to the degree of noise exposure.

The noise exposure will be divided into 5 classes, ranging with steps of 5 dBA from 50 dBA to 75 dBA. We will stratify the sample according to these noise classes. For a particular noise class, the precision of the estimates depends on the sample size. Depending on population sizes, significance levels and precision we have to choose the sample sizes. Table 3 gives an indication of the precision in relation with the significance level and the population size (these figures are based on the worst case, so the precision is in practice better)

Tabel 12 Minimal sample sizes for given significance level and precision..

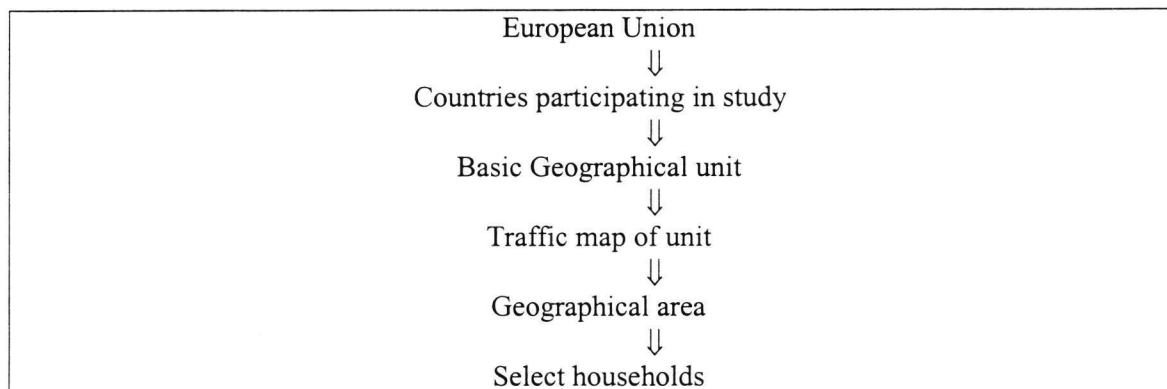
Population size	Significance level = 0.1 Precision = 0.1	Significance level = 0.05 Precision = 0.05
500	69	219
1000	74	293
5000	78	370
10000	78	390
50000	79	402
100000	79	405

Based on this table and practical reasons we will choose sample sizes according to Figure 5. In traffic category 1 and 4 we will have 100 respondents for each noise class. This means that for air traffic we sample 500 respondents for each airport, giving 1000 respondents in a country. For noise from traffic in cities we will have in total 2000 respondents. The precision will be between 0.05 and 0.1. If however not much difference exists between the regions the precision will be higher. For category 2 en 3 the minimal precision is something less than 0.1, but since we have more regions and sub-regions which have some phenomena in common, the expected precision will be smaller than 0.1. For these categories we include 1500 respondents each. In total these numbers give 7500 respondents in a country.

Inevitable we have to deal with non-response. The larger the (expected) non-response, the more intensive a non-response survey should be, in order to obtain information on the non-respondents. Such investigation gives evidence whether and how the results of the sample can be generalized to the whole population. The non-response survey should be designed such that

- Information is obtained of the reason of people not cooperating with the study.
- The opinion of the non-respondent towards the project.
- An indication of the noise annoyance (may be at a far less precise scale as for the respondents).

Indications of the relevant covariates of the non-respondents are obtained (or on population level the distribution is known, so can be compared with the respondents).



Figuur 5: The sampling plan scheme.

N countries	→	2 regions	→	2 sub-regions
↓		↓		↓
2 airports		3 road-parts 3 railway-parts		2 road-parts
↓		↓		↓
500 households, 100 in each dBA class		250 Households, 50 in each dBA class		500 Households, 100 in each dBA class

Figuur 6: The sampling plan in total for each country.

5 Inventarisatie literatuur en software over steekproefontwerp

Over steekproefontwerp en survey onderzoek is veel geschreven in boeken en artikelen. Daarnaast is er ook de nodige software ontwikkeld voor analyses, steekproefomvang berekeningen (powercalculations) en voor het (her)wegen van de steekproef. In dit hoofdstuk zullen we een overzicht geven van de belangrijkste steekproefontwerp literatuur. Ook zullen we een aantal nuttige internetlinks behandelen, maar daarbij moet wel in de gaten gehouden worden dat er niets zo veranderlijk is als het internet dus dat deze links op den duur niet meer bruikbaar kunnen zijn. Daarnaast moet je bij berekeningen van software van het internet oppassen of de berekeningen wel correct zijn. Je kan er namelijk niet vanuit gaan dat de software goed gecontroleerd is op eventuele fouten. Tevens weet je ook niet, als de software niet goed gedocumenteerd is, op welke ideeën, aannamen de berekeningen gebaseerd zijn. Natuurlijk zullen de overzichten van boeken, artikelen, software en internetlinks zeker niet volledig zijn.

5.1 Boeken

Hieronder volgt een lijstje, in volgorde van verschijning, met beschrijving van een aantal boeken over survey sampling, en steekproeftechnieken. De twee Nederlandse boeken zijn wellicht de meest toegankelijke boeken. Daarnaast zijn de boeken van de schrijvers Kish en Cochran traditioneel en heel uitgebreid. Het boek van Thompson is het meest modern. Op het boek van Kish na zijn de hieronder beschreven boeken momenteel aanwezig bij de vakgroep Statistiek.

(1965) **Kish, L., *Survey Sampling***, Wiley New York (ISBN 0-471-48900-X of de Classics Edition: ISBN 0-471-10949-5).

Trefwoorden: Simple random sampling, Stratified sampling, Proportionate sampling, Systematic sampling, Cluster sampling, Economic design of surveys, Area sampling, Multistage Sampling, Sampling from imperfect frames, Nonresponse

Beschrijving: In dit standaardwerk worden, voor traditionele steekproefontwerpen, de bijbehorende schatters, varianties, betrouwbaarheidsintervallen, optimale steekproefomvang behandeld. Er wordt vrij veel aandacht besteed aan het economisch ontwerpen van surveys met behulp van verschillende kostenfuncties. Het boek is volledig maar ook zeer technisch.

(1977) **Cochran W.G., *Sampling Techniques***, Wiley New York (ISBN 0-471-16240-X).

Trefwoorden: Simple random sampling, Sampling proportions and percentages, Sample size calculation, Stratified random sampling, Allocation problems, Ratio estimators, Regression estimators, Systematic sampling, Cluster sampling, Subsampling, Double sampling.

Beschrijving: Dit boek lijkt qua onderwerpen erg op het boek van Kish. Binnen de statistiek zelf wordt dit boek als een standaardwerk gezien. De leesbaarheid van dit boek is iets beter dan het boek van Kish maar toch is het nog erg technisch met veel formules.

(1985) **Bethlehem J.G.,** *Theorie en praktijk van het steekproefonderzoek*, CBS 98, hoofdafdeling statistische methoden (ISBN 9-067-86023-9).

Trefwoorden: Elementaire kansrekening, Enkelvoudige steekproefontwerpen zoals EAS, systematische steekproefontwerp, enkelvoudige steekproefontwerp met ongelijke kansen, Meervoudige steekproefontwerpen zoals het gestratificeerde steekproefontwerp, clustersteekproefontwerp, tweetrapssteekproefontwerp en het tweedimensionale steekproefontwerp, schatters, kwaliteit van de uitkomsten, rapportage.

Beschrijving: Deze syllabus is geschreven als ondersteuning voor een cursus steekproeftheorie van het CBS. Er wordt eerst een inleiding in de benodigde kansrekening gegeven. Vervolgens passeren de standaard steekproefontwerpen de revue. Bij elk steekproefontwerp worden de Horvitz-Thompson schatters voor het gemiddelde, populatietotaal en populatieproportie gegeven, met de bijbehorende variantie en steekproefomvang berekening. Voor de diverse ontwerpen worden handige schema's gegeven.

(1988) **J. Cohen,** *Statistical power analysis for the behavioral Sciences*, (2nd ed.), Hillsdale, NJ: Erlbaum, (ISBN).

Trefwoorden: Concepts of power analysis, t-test for means, difference between correlation coefficients, difference between proportions, Chi-Square test, power analysis for analysis of variance, multiple regression analysis, multivariate methods.

Beschrijving: Power berekeningen voor standaard toetsen zoals de t-toets, verschillen in proporties, verschillen tussen correlatie coëfficiënten, de Chi-Square test, Variantie analyses, Multiële regressie analyses, Multivariate methoden. Het is een praktisch boek met veel handige formules en tabellen.

(1990) **Lemeshow, S., Hosmer Jr, D.W., Klar, J.,** *Adequacy of Sample Size in Health Studies*, Wiley, New York (ISBN 0-471-92517-9).

Trefwoorden: The one and two-sample problem, Sample size for case-control studies, Sample size determination for cohort studies, Lot quality assurance sampling, the Incidence rate, Sample size for continuous response variables, Sample size for sample surveys.

Beschrijving: Dit boek is geschreven vanuit een gezondheidskundig en epidemiologisch onderzoeksoogpunt. Het verschil met de voorgaande beschreven boeken is dat het dus niet op survey-achtig onderzoek gericht is maar meer op het gebied van case-control stu-

dies, interventie studies, cohort studies etcetera. Het boek is zeer praktisch gericht met een aantal makkelijke en nuttige tabellen.

(1992) **Muilwijk, J., Snijders, T.A.B., Moors, J.J.A., *Kanssteekproeven***, Stenfert Kroese Uitgevers Leiden/Antwerpen (ISBN 90-207-2221-2).

Trefwoorden: Enkelvoudig aselechte steekproeven en directe schatters, Gelede steekproeven (gestratificeerde steekproeven), Trekking met ongelijke kansen, Trossteekproeven (geclusterde steekproeven), Systematische steekproeven, Getrapte steekproeven, Matrix steekproeven, Niet-steekproeffouten en de gang van zaken bij steekproefonderzoek.

Beschrijving: Ook dit boek gaat uit van de traditionele steekproefontwerpen en geeft daarvoor de bijbehorende schatters, varianties, steekproefomvangen e.d. Er wordt ook vrij uitvoerig aandacht besteedt aan de praktische kanten en problemen bij steekproefonderzoek. Het boek is, ondanks de formules, goed leesbaar en één van de weinige in het Nederlands geschreven boeken op dit gebied. Helaas is dit boek wel 'out of print' en daarom in de boekhandels niet meer verkrijgbaar.

(1992) **Thompson, S.K., *Sampling***, Wiley New York (ISBN 0-471-54045-5)

Trefwoorden: Simple random sampling, Confidence intervals, Sample size calculation, Estimating proportions, ratios en subpopulation means, unequal probability sampling, Auxiliary data, Ration estimation, Regression estimation, Sufficient statistic in sampling, Stratified sampling, Cluster and systematic sampling, multistage Designs, Double sampling, Network sampling, Detectability methods for elusive populations (e.g. Capture-Recapture sampling), Spatial sampling, Adaptive sampling.

Beschrijving: In dit boek komen, naast de traditionele methoden ook de wat modernere steekproefontwerpen zoals network sampling, spatial sampling en adaptive sampling aan de orde. Tegenwoordig is dit boek een standaardwerk in de statistiek.

(1994) **Diggle, P., Liang K.Y., Zeger, S.L., *Analysis of Longitudinal data***, Clarendon Press (ISBN: 0-198-52284-3)

Trefwoorden: (2^e hoofdstuk) Design and Sample size considerations for longitudinal studies.

Beschrijving: Dit boek gaat over analyse van longitudinale data. In het tweede hoofdstuk worden enkele praktische tips en formules gegeven voor de berekening van steekproefomvangen voor longitudinale studies. Het betreft hier wel de meest eenvoudige datastructuren. Geavanceerdere modellen zijn te vinden in meer recent gepubliceerde artikelen (zie Paragraaf 5.2).

(1995) Lehtonen, R., Pahkinen, E.J., *Practical Methods for Design and Analysis of Complex Surveys*, Wiley New York, (ISBN 0-471-93934-X)

Trefwoorden: Basic sampling techniques, Further use of auxiliary information (stratified sampling and cluster sampling), Handling missing data, Linearization and sample re-use in variance estimation, Covariance-matrix estimation of ratio estimators, Analysis of one-way and two-way tables, Multivariate survey analysis.

Beschrijving: Naast de bekende ontwerpen gaat dit boek in op survey analysetechnieken. Het bestaat voor de helft uit casestudies van sociaal en gezondheidsonderzoek. Commercieel verkrijgbare survey analyse softwarepakketten voor o.a. herwegen en poststratificatie, zoals OSIRIS IV, PC CARP en SUDAAN, worden besproken

5.2 Artikelen

De artikelen die hier genoemd zijn gaan over de wat ingewikkeldere situaties die tot nu toe nog nauwelijks beschreven zijn in de boeken. Het gaat dan met name over meertraps-en geclusterde steekproeven. De samenvattingen zijn in het engels, net zoals de artikelen zelf.

(1981) Donner, A., Birkett, N., and Buck, C. *Randomization by cluster: sample size requirements and analysis*, American journal of Epidemiology, Vol. 114, No. 6, pp 906-914.

Summary: This article discusses the statistical aspects of cluster randomization, considering, first, the comparison of means, and second, the comparison of proportions. In the situation of k clusters of n individuals the formula for the variance of the sample means (both control and experimental group) implies that the correct sample size for a clustered sample can be obtained by inflating the usual estimate of the required number of individuals in each group by the "inflation factor" $1+(n-1)\rho$. Again, ρ is the intra-cluster correlation. For the situation of comparing an experimental and a control group the number of individuals one should take is given by

$$n = \frac{2(Z_{\alpha} - Z_{\beta})^2 \sigma^2 (1 + (n-1)\rho)}{\delta^2},$$

where Z_{α} is the z-score of α and Z_{β} is the z-score of β . In the denominator stands δ , the meaningful difference, specified in advance by the investigator.

(1993) Snijders, T.A.B., Bosker, R.J., *Standard Errors and Sample sizes for two-level research*, Journal of Educational Statistics, Vol. 18, No. 3, pp. 237 – 259

Summary: In this article the hierarchical linear model approach to a two-level design is considered, some variables at the lower level having fixed and others having random

regression coefficients. An approximation is derived to the covariance matrix of the estimators of the fixed regression coefficients (for variables at the lower and the higher level) under the assumption that the sample sizes at either level are large enough. This covariance matrix is expressed as a function of parameters occurring in the model. If a researcher planner can make a reasonable guess as to these parameters, this approximation can be used as a guide to the choice of sample sizes at either level.

The formulas of this article are implemented in the program PINT (see on the internet-site <http://stat.gamma.rug.nl/snijders/#progPINT>), which calculates the optimal sample size. One of the parameters of input is the cost to add an extra school or student. The output consists of standard errors that can be expected from the chosen design, given a value of the power. The effect size can be obtained by the formula: $(\text{effect size} / \text{standard error}) \approx (Z_\alpha + Z_\beta)$.

(1995) **M.Mok** *Sample size requirements for 2-level designs in educational research* <http://www.ioe.ac.uk/multilevel/workpap.html>

Summary: This study is concerned with the effect of sample size on the efficiency of estimates for 2-level survey research designs. One component of sample design is the number of level-2 units and the number of level-1 units to be included in the sample to produce accurate information. For example, in the study of school culture, the researcher needs to decide how many schools, and how students from each school, have to be included in the study. Results are given from an extensive simulation study. Models with an intercept and slope are considered.

Main conclusions: If resources are available for a sample size (in total) of n , comprising J schools with I students from each school then less bias and more efficiency would be expected from sample designs involving more schools (large J), and fewer students per school (small I) than from sample designs involving fewer schools (small J), and more students per school (that is with large I). Converting the actual sample size into effective sample size according to the formula $n_{\text{eff}} = n / [1 + (n_{\text{clus}} - 1)\rho]$ removed at least for the fixed effects the clustering effect in estimating the fixed components of the slope and the intercept. Here ρ is the intra-cluster correlation.

(1998) **Cohen, M.P.**, *Determining Sample Sizes for Surveys with data analyzed by Hierarchical Linear models*, Journal of Official Statistics, Vol. 14, No. 3 pp 267-275.

Summary: Cohen considers random slope models. He calculates the desired sample sizes for estimating school-level and student-level variances. Next he considers the situation of regression coefficients. All his calculations are based on the simple cost-function $C = m * C_s + m * n * C_k$, where C_s expresses the cost to include an extra school (m schools) into the sample, C_k (n students per school) is the cost to include an extra student to the sample. The author states that traditional samples designs may do very

well in enabling us to estimate the regression coefficients. In analyses where it is important to estimate also the variance components, the school-level variance in particular, one must instead sample more students per school and fewer schools.

Finally he comments that hierarchical models become more widely used by researchers analyzing survey data, so the need grows for survey design statisticians to understand the implications of such use for good survey design. All current articles scarcely scratched the surface and therefore there are opportunities abound for further research on this topic.

5.3 Software

Binnen TNO Preventie en Gezondheid zijn een aantal statistische pakketten aanwezig die ook de mogelijkheid bieden om steekproeven achteraf te wegen met eenvoudige weegprocedures. Dit zijn SPSS, Stata, SAS en S-PLUS. Indien er binnen SPSS gewogen wordt kan het gebeuren dat de correlaties verkeerd geschat worden. Hiervoor dient dus opgepast worden.

Naast de eenvoudige weegprocedures die aanwezig zijn in de standaard pakketten is er ook het programma Bascula. Bascula is een programma dat ontwikkeld is bij het CBS. Met behulp van dit programma kunnen eenvoudige wegingen voor bijvoorbeeld poststratificatie uitgevoerd worden. Ook ingewikkeldere methoden als lineair wegen en multiplicatief wegen zijn beschikbaar. Deze laatste twee worden gebruikt als er te weinig informatie is van de populatie om poststratificatie toe te passen of als er bij poststratificatie lege of kleine steekproefcellen zouden ontstaan.

Voor het programma moet de data aangeleverd worden in ASCII format of in het Blaise formaat. Blaise is een systeem voor computer-geassisteerd surveyonderzoek, dat ook ontwikkeld is bij het CBS (Referentie: Bascula, for weighting sample survey data; Reference Manual; CBS; ISBN 903 571 603 5).

Voor power berekeningen (steekproefomvang berekeningen) is het pakket nQuery advisor beschikbaar binnen TNO-PG. Dit programma kan voor diverse standaard analyses de benodigde steekproefomvang berekenen. Bij alle verschillende situaties kun je, gegeven een steekproefomvang, en een betrouwbaarheid de bijbehorende power berekenen, of bij een gegeven power en betrouwbaarheid de benodigde steekproefomvang.

5.4 Nuttige internetlinks

<http://members.aol.com/johnp71/javastat.html> Deze pagina bevat een groeiende lijst van platform onafhankelijke statistische software packages (o.a. een aantal java-applets) waarmee standaard statistische analyses kunnen worden uitgevoerd, zoals poweranalyses. Vanuit deze pagina

kan men verder klikken naar online statistische boeken, tutorials, downloadable software en gerelateerde bronnen.

<http://yates.coph.usf.edu/research/psmg/Power/schoolpower.html> Calculations for School and Classroom Based Randomized Designs. Deze webpagina berekent de statistische power voor drie-niveau modellen (bijvoorbeeld leerlingen binnen klassen binnen scholen) volgens "Principles for Designing Randomized Trials in Mental Health" by Brown CH and Liao J (1999), geaccepteerd voor publicatie in the American Journal of Community Psychology. Deze pagina is dus met name interessant voor het geval van clustersteekproeven en/of drietrapssteekproeven.

<http://www.stat.ucla.edu/calculators/powercalc> Hierop is een power calculator gegeven voor diverse standaard situaties. Als uitgangspunt kan men kiezen voor een vaste power om vervolgens de steekproefomvang te berekenen of een vaste steekproefomvang en dan wordt de power berekend.

<http://sustain.forestry.ubc.ca/cacb/power/> Op deze pagina is een lijst gegeven met allerlei programmatuur plus beschrijving ervan voor het berekenen van steekproefomvangen en de power.

<http://www.bio.ri.ccf.org/power.html> Beschrijving en download mogelijkheid van de gratis SAS Module/Macro, genaamd UnifyPow, voor steekproefomvang analyse.

<http://stat.gamma.rug.nl/snijders/#progPINT> Dit programma berekent benaderingen van standaardafwijkingen voor geschatte fixed effect parameters in een hiërarchisch lineair model (een zogenaamd *multilevel model*) met twee niveaus. De formules die gebruikt worden komen uit Snijders en Bosker (1993).

Referenties

- Bethlehem JG. *Theorie en praktijk van het steekproefonderzoek*. Voorburg (etc.): CBS 98, hoofdafdeling statistische methoden, 1985.
- Bethlehem JG, Kersten HMP. *On the treatment of nonresponse in sample surveys*. J Official Statistics 1984;1:287-300.
- Brick JM, Kalton G. *Handling missing data in survey research*. Stat Methods Med Res 1996;5:215-38.
- Buuren S van, Oudshoorn CGM. *Flexible multivariate imputation by MICE*. Leiden: TNO PG, 1999. Publ.nr. 99.054
- Cochran WG. *Sampling Techniques*. New York:Wiley, 1977.
- Cohen J. *Statistical power analysis for the behavioral Sciences*, (2nd ed.). Hillsdale, NJ: Erlbaum, 1988.
- Cohen, M.P., *Determining Sample Sizes for Surveys with data analyzed by Hierarchical Linear models*, J Official Statistics 1998;14(3):267-75.
- Diggle P, Liang KY, Zeger SL. *Analysis of Longitudinal data*, Clarendon Press, 1994.
- Dongen JEF van, Vos H. *Peiling Milieuhinder Utrecht 1997*. Leiden: TNO PG, 1997. Publ.nr. 98.026.
- Donner A, Birkett N, Buck C. *Randomization by cluster: sample size requirements and analysis*, Am J Epidemiol 1981;114(6): 906-14.
- Greenland S, Finkle WD. *A critical look at methods for handling missing covariates in epidemiologic regression analysis*. Am J Epidemiol 1995;124(12):1255-64.
- Kish L. *Survey Sampling*. New York: Wiley, 1965.
- Leeuw ED de, Hox JJ. *Nonresponse in surveys: een overzicht*. Statistica Neerlandica 1998;57:31-53.
- Lehtonen R., Pahkinen EJ. *Practical Methods for Design and Analysis of Complex Surveys*. New York, Wiley, 1995.
- Lemeshow S, Hosmer DW jr, Klar J. *Adequacy of Sample Size in Health Studies*. New York: Wiley, 1990.
- Little RJA, Rubin DB, *Statistical analysis with missing data*. New York: Wiley, 1987
- Mok, M., *Sample size requirements for 2-level designs in educational research*, 1995.
<http://www.ioe.ac.uk/multilevel/workpap.html>
- Muilwijk J, Snijders TAB, Moors JJA. *Kanssteekproeven*. Leiden (etc.): Stenfert Kroese Uitgevers 1992.

Oudshoorn CGM, Buuren S van, Rijckevorsel JLA van. *Flexible multiple imputation by chained equations of the AVO-95 survey*. Leiden: TNO PG, 1999. Publ.nr. 99.045.

Passchier-Vermeer W, Vos H, Gils K van, Miedema HME, Eoo F de, Verhoeff EJ. *Aircraft noise and sleep disturbance Pilotstudy*, 1999.

Reijneveld SA, Stronks K. *The impact of response bias on estimates of health care utilisation in a metropolitan area: a study using administrative data*. Geaccepteerd door Int. J. Epidemiology

Rice JA. *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks/Cole, 1988.

Snijders, T.A.B., Bosker, R.J., *Standard Errors and Sample sizes for two-level research*, J Educational Statistics 1993;18(3): 237–59

Thompson SK. *Sampling*. New York: Wiley New York, 1992.