STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Studies naar samenhang

2. Uitbreiding

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Statistische studies naar samenhang

Deel 2. Uitbreiding

Deze tekst heeft dezelfde structuur als de vorige tekst die de basisbegrippen behandelde. Ook in deze tekst is elke studie opgevat als een "wetenschappelijk onderzoek".

De volgende 4 grote stappen komen bij elk onderzoek terug:

1. het stellen van de onderzoeksvraag
2. het verzamelen van de data
3. de analyse van de data
4. de interpretatie van de resultaten.

1. Biovoeding en pesticidenresidu's

De samenhang tussen geboortegewicht en zwangerschapsduur heb je vroeger bestudeerd. Ook bij de studie naar "inspanning en hartslag" heb je alle stappen van een onderzoek doorlopen. Zo leer je hoe een onderzoek echt werkt. Dat helpt ook om studies die door anderen zijn uitgevoerd te begrijpen en om er kritische vragen bij te stellen. Een deel van zo'n studie vind je hieronder. Bij deze studie is zowel de respons als de verklarende veranderlijke categorisch. De opmetingen komen dan terecht in een **kruistabel** (of **contingentietabel**).

De volgende studie werd uitgevoerd in 2000 in Californië.

Om de veiligheid van voedsel te bewaken neemt de eetwareninspectie regelmatig steekproeven en controleert of eetwaren geschikt zijn voor consumptie. Bij zo'n controle wordt er ondermeer gekeken naar de aanwezigheid van pesticidenresidu's. Er werd ook genoteerd of het over bioproducten ging of over producten die op de conventionele manier waren geteeld. Dat leverde het volgende resultaat.

		Aanwezigheid van pesticidenresidu's	
		Ja	*Neen*
Productiemethode	*Biologisch*	*29*	*98*
	Conventioneel	*19 485*	*7 086*

DISCUSSIEMOMENT 1.

Hoewel biologisch geteelde gewassen in principe pesticidenvrij zouden moeten zijn is dat toch niet altijd het geval. Maar als consument, die extra betaalt voor bioproducten, verwacht je toch dat pesticidenresidu's veel minder voorkomen in biologische producten dan in conventioneel geteelde.

Wat is hier de respons en wat is de verklarende veranderlijke? Van welk type zijn zij?
Wat betekenen de getallen 29 en 19485 uit de tabel? Kan je die zomaar vergelijken?
Is de eetwareninspectie alleen geïnteresseerd in de informatie uit die steekproef?

Gebruik je antwoord op de 5 vorige vragen om de onderzoeksvraag te formuleren.

1.1. De onderzoeksvraag

De verklarende veranderlijke is de manier van telen: biologisch of conventioneel. Dit is een categorische veranderlijke met 2 categorieën.
De respons is de aanwezigheid van pesticidenresidu's. De eetwareninspectie heeft hier enkel aangeduid of er in een product al dan niet pesticidenresidu's voorkomen. Er zijn geen hoeveelheden genoteerd. Daarom is ook de respons een categorische veranderlijke met 2 categorieën: "Ja" en "Neen".

Het getal 29 is het aantal bioproducten in deze steekproef die pesticidenresidu's bevatten. Er waren 19485 conventioneel geteelde producten met pesticidenresidu's. Je kan die twee getallen niet zomaar met elkaar vergelijken want in totaal waren er ook veel meer conventioneel geteelde producten in de steekproef dan bioproducten. Dus moet je per productiemethode naar verhoudingen kijken.

De eetwareninspectie is geïnteresseerd in pesticidenresidu's in eetwaren in Californië. Om dat te onderzoeken worden regelmatig steekproeven getrokken. Deze studie vond plaats in 2000.

Uit wat hierboven onderlijnd is kan je de volgende onderzoeksvraag afleiden: "Is in Californië in het jaar 2000 de proportie eetwaren met pesticidenresidu's kleiner bij bioproducten dan bij conventionele producten?".

1.2. Het verzamelen van de data

De data zijn al verzameld en samengevat in de bovenstaande tabel. Je mag onderstellen dat er in Californië op een goede manier steekproeven worden getrokken om eetwaren te controleren. Bij je conclusie mag je dus veralgemenen van steekproef naar populatie.

De verklarende veranderlijke zorgt ervoor dat je de populatie opsplitst in twee nieuwe populaties. Je wil immers de respons (aanwezigheid of afwezigheid van pesticidenresidu's) bij de populatie van alle conventioneel geteelde voedingsproducten in Californië vergelijken met de respons bij de populatie van alle voedingsproducten die daar biologisch geteeld worden. Je kan die twee populaties niet volledig opmeten en dus gebruik je een steekproef. De oorspronkelijke steekproef van de eetwareninspectie is getrokken uit de totale populatie van alle voedingsproducten. Splits die steekproef nu op volgens de productiemethode. Zo bekom je uit elk van de te bestuderen populaties een afzonderlijke steekproef.

1.3. De analyse van de data

De data zijn beschikbaar maar de analyse moet je zelf nog uitvoeren.

DISCUSSIEMOMENT 2.

Om te beginnen vervolledig je de tabel en vul je de randtotalen in.
Zorg ervoor dat je die tabel goed begrijpt.

		Aanwezigheid van pesticidenresidu's		
		Ja	Neen	Totaal
Productiemethode	Biologisch	29	98	
	Conventioneel	19 485	7 086	
	Totaal			

- Hoeveel gewassen zijn er in totaal gecontroleerd?
- Hoeveel gewassen zijn er met pesticidenresidu's?
- Hoeveel conventioneel geteelde gewassen zijn er met pesticidenresidu's?

De onderzoeksvraag gaat over een verschil in proporties bij twee populaties en dus moet jij per populatie naar de juiste steekproef kijken. Bovendien mag je de opgemeten frequenties niet zomaar met elkaar vergelijken. Je weet dat je per productiemethode naar verhoudingen (proporties) moet kijken.

Maak nu een nieuwe kruistabel waarin de proporties staan die je wel rechtstreeks met elkaar kan vergelijken.

De kruistabel met randtotalen ziet er als volgt uit.

		Aanwezigheid van pesticidenresidu's		
		Ja	Neen	Totaal
Productiemethode	Biologisch	29	98	**127**
	Conventioneel	19 485	7 086	**26 571**
	Totaal	**19 514**	**7 184**	**26 698**

In totaal zijn er 26 698 producten gecontroleerd en er zijn 19 514 gewassen met pesticidenresidu's. Dergelijke informatie lees je af in de randtotalen en in het algemeen totaal. "Binnenin" de kruistabel zie je hoeveel (= de frequentie) producten er zijn voor elke combinatie van de categorieën van de twee categorische veranderlijken. Een kruistabel is dus eigenlijk een tweedimensionale frequentietabel. Zo zijn er bijvoorbeeld 7 086 producten die tegelijkertijd aan de volgende 2 eigenschappen voldoen: zij zijn conventioneel geteeld en zij bevatten pesticidenresidu's.

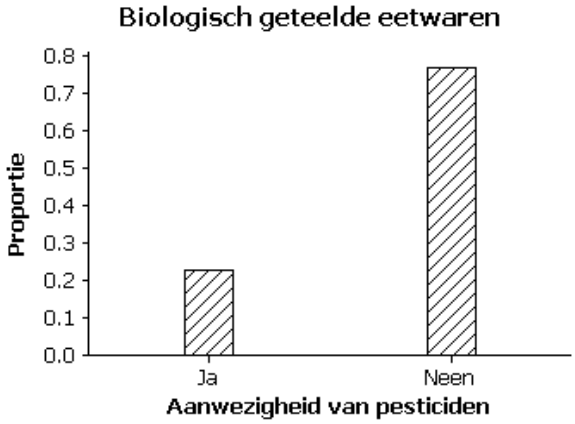
De onderzoeksvraag gaat over een verschil in proporties bij twee populaties. Jij beschikt niet over opmetingen van die volledige populaties, je hebt enkel steekproeven. Jij zal dus naar het verschil in proportie bij die twee steekproeven kijken. Bij een kruistabel komt dat neer op een studie van **conditionele** (of **voorwaardelijke**) proporties.

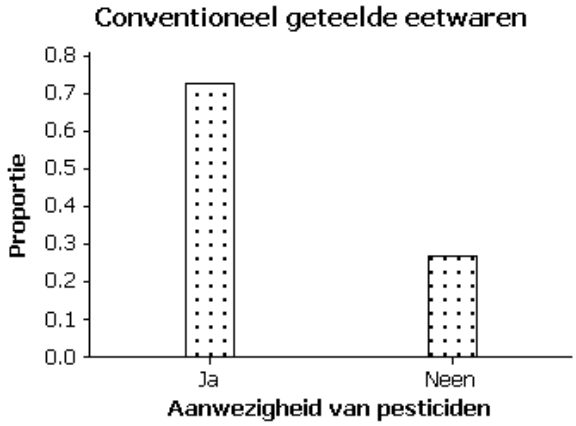
Conditioneel op het feit dat de geteste producten biologisch geteeld zijn kan je kijken naar de proportie bioproducten met pesticidenresidu's bij alle opgemeten bioproducten. Dat betekent dat je in de kruistabel alleen maar kijkt naar de eerste rij. Die beschouw je als een steekproef van grootte $n = 127$ uit de populatie van alle biologisch geteelde gewassen in Californië. In die steekproef zijn er 29 producten met pesticidenresidu's. Dat is een proportie van $29/127 \cong 0.23$.

Op eenzelfde manier kan je conditioneel op "conventioneel geteeld" werken en ook daar alles in proporties uitdrukken. Zo krijg je een tabel met conditionele proporties. Bemerk dat je conditioneert op de verschillende waarden die de verklarende veranderlijke aanneemt.

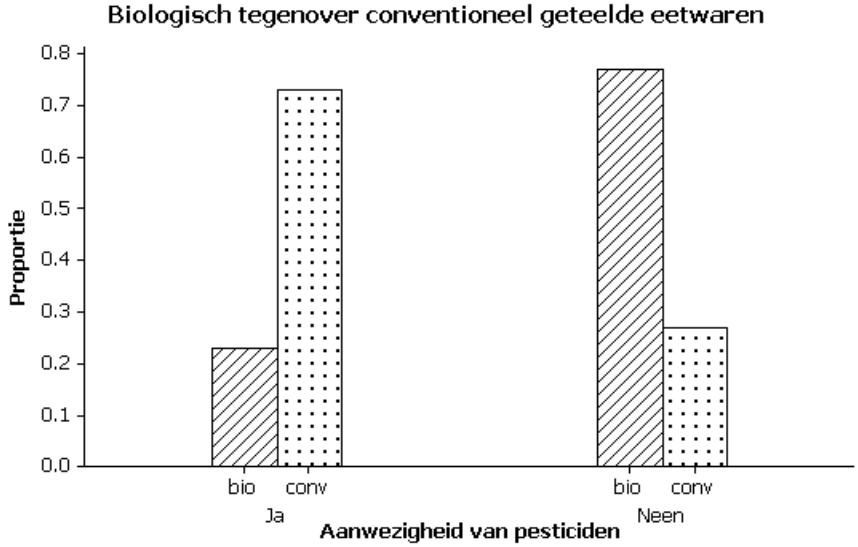
		Aanwezigheid van pesticidenresidu's		
		Ja	Neen	Totaal
Productiemethode	Biologisch	29	98	**127**
		0.23	0.77	**1.00**
	Conventioneel	19 485	7 086	**26 571**
		0.73	0.27	**1.00**

Het al dan niet aanwezig zijn van pesticidenresidu's (en in welke proportie) kan je per productiemethode grafisch uitzetten in een staafdiagram. In deze studie heeft de respons maar twee waarden en dus krijg je slechts 2 staafjes in je grafiek. Als je een categorische respons hebt met meerdere categorieën dan krijg je een staafdiagram met meerdere staafjes.





Om de twee productiemethoden gemakkelijk te vergelijken is een staafdiagram met subtypes hier een aangewezen figuur.



1.4. Interpretatie van de resultaten

DISCUSSIEMOMENT 3.

Onderstel eens dat er geen verschil zou zijn tussen de populaties en dat er bij beide populaties 70 % producten zouden zijn die pesticidenresidu's bevatten. Hoe ziet een kruistabel met conditionele proporties er dan uit voor die populaties? En welke figuur krijg je dan voor het bijhorende staafdiagram met subtypes? Is er hier iets wat je aandacht trekt?

Als je nu uit die ene populatie een steekproef van grootte $n = 127$ trekt en uit de andere een steekproef van grootte $n = 26\,571$ denk je dan dat jij in die toevallige steekproeven ook telkens exact 70 % producten zal hebben die pesticidenresidu's bevatten? Waarom? Verwacht je anderzijds heel grote verschillen tussen de gevonden proporties in de steekproeven als ze allebei getrokken zijn uit populaties die exact hetzelfde percent producten bevatten met pesticidenresidu's? Waarom?

Kijk nu wat er in die echte steekproef in Californië is gevonden. Wat denk je nu over de populatie van alle biologisch geteelde producten in Californië tegenover de populatie van alle conventioneel geteelde producten? Motiveer je antwoord en onderstel dat de steekproef in Californië op een goede statistische manier getrokken is.

Heb jij hier bewijzen dat biologische landbouw een vermindering aan pesticidenresidu's veroorzaakt? Leg uit.

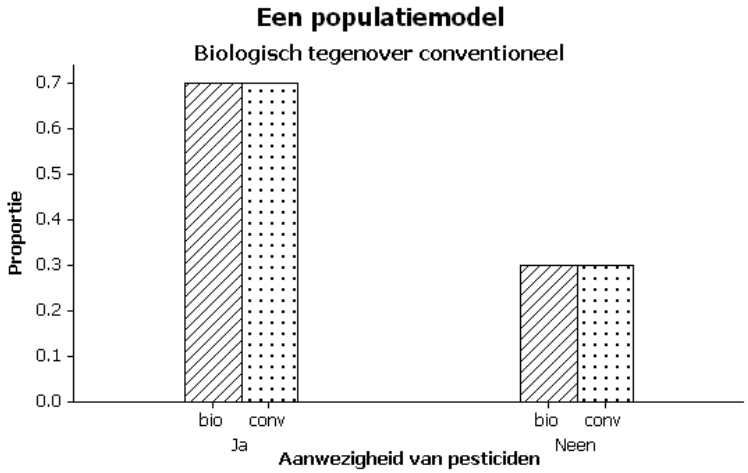
Informatie uit grafieken en kengetallen

Is er wel een verband tussen de teeltmethode en pesticidenresidu's?

Denk eerst even aan wat er gebeurt wanneer er geen verband is en de aanwezigheid van pesticidenresidu's onafhankelijk is van de productiemethode. Op het niveau van de populatie betekent dit dat er zowel bij de biologisch geteelde als bij de conventionele producten exact dezelfde proportie producten met pesticidenresidu's voorkomt. Onderstel eens dat die proportie 0.70 zou zijn. De theoretische tabel, die een model is voor twee populaties met dezelfde proportie, ziet er dan als volgt uit.

		Aanwezigheid van pesticidenresidu's	
		Ja	Neen
Productiemethode	Biologisch	0.70	0.30
	Conventioneel	0.70	0.30

Dat populatiemodel kan je ook grafisch voorstellen



Als er bij de populaties geen verschil is dan zijn, voor elke responscategorie, de staafjes voor biologisch geteelde producten even groot als voor conventioneel geteelde. Bemerk dat die staafjes de proportie in de populatie weerspiegelen (en niet in de steekproef die je daar later zal uit trekken).

Als beide populaties er zouden uitzien zoals aangegeven en je zou uit elke populatie een steekproef trekken dan verwacht je niet dat er in beide steekproeven exact 70 % van de producten pesticidenresidu's bevat. Daar zal wel wat verschil op zitten, gewoon door het lukraak trekken. Maar heel grote verschillen verwacht je eigenlijk niet wanneer beide populaties waaruit je trekt echt dezelfde proportie pesticidenresidu's bevatten.

In de steekproeven van Californië heb je grote verschillen gevonden in de conditionele proporties met pesticidenresidu's (0.23 tegenover 0.73). Je hebt dus een sterke samenhang in je steekproef gezien. Bovendien zijn de steekproeven niet klein en zijn ze op een goede statistische manier getrokken. Daarom kan je vermoeden dat er ook bij de totale populaties een samenhang is tussen de aanwezigheid van pesticidenresidu's en de manier waarop gewassen geteeld worden. Je vermoeden dat bij biologisch geteelde gewassen de proportie

pesticidenresidu's kleiner is dan bij conventioneel geteelde kan je bevestigen met methoden uit de verklarende statistiek, maar dit wordt hier niet behandeld.

Informatie uit het ontwerp van de studie

Merk op dat dit onderzoek een observatiestudie is. Als je hier samenhang ontdekt dan kan je alleen maar zeggen dat er een associatie is vastgesteld tussen de aanwezigheid van pesticidenresidu's en de productiemethode. Je kan niet zeggen dat biologische landbouw de waargenomen vermindering in pesticidenresidu's veroorzaakt. Dit betekent nog niet dat dit toch wel waar kan zijn. Het betekent alleen dat het ontwerp van onderzoek dat hier gebruikt is je niet toelaat om dergelijke sterke uitspraak te doen.

2. Geslacht en discriminatie

De volgende studie komt uit het domein van de psychologie en werd gepubliceerd in het wetenschappelijk tijdschrift "Journal of Applied Psychology". De studie werd uitgevoerd in de Verenigde Staten en dateert al van 1972 maar het thema is nog uitermate actueel. Je kan het woord "vrouw" eventueel vervangen door "moslim", "zwarte", "allochtoon", enz. De studie was veel uitgebreider en we geven hier een verkorte en aangepaste versie.

Een grote bank in Amerika heeft in elk van de 50 staten heel wat filialen. De verantwoordelijke directeur is daar bijna altijd een man. Er zijn er zo meer dan duizend. Uit die mannen werd een lukrake steekproef van grootte $n = 48$ *getrokken. Aan die 48 bankdirecteuren werd een "dossier van een bankbediende" bezorgd met de vraag of die bankbediende goed genoeg was om in aanmerking te komen voor promotie. Het "dossier van de bankbediende" was in alle 48 gevallen identiek met één verschil: bij 24 dossiers stond er dat het om een man ging en bij de andere 24 dossiers ging het om een vrouw. De dossiers werden op een lukrake manier onder die bankdirecteuren verdeeld. De 48 bankdirecteuren wisten niets van elkaar en handelden helemaal autonoom. Het resultaat was als volgt.*

		Aanbevolen voor promotie	
		Ja	*Neen*
Geslacht	*Man*	*21*	*3*
	Vrouw	*14*	*10*

DISCUSSIEMOMENT 4.

Gebruik de informatie die je over deze studie gekregen hebt om de stappen in dit onderzoek te bespreken en verder te vervolledigen. Ga systematisch te werk en overleg in je groep wat je allemaal zal doen. Denk aan de 4 grote stappen bij een wetenschappelijk onderzoek en inspireer je op het onderzoek over pesticidenresidu's voor de te gebruiken statistische technieken.
Let bij je conclusie op het fundamentele onderscheid tussen een observatiestudie en een experiment.
Vergelijk ten slotte jouw onderzoek met de onderstaande tekst.

2.1. De onderzoeksvraag

Als je het ontwerp van dit onderzoek goed leest, dan zie je dat je hier te maken hebt met een experiment waarbij de deelnemers (mannelijke bankdirecteuren) door een lukrake steekproef (EAS) uit een grotere populatie bij jou zijn terechtgekomen. Dat betekent dat, als je een samenhang ontdekt, er hier sprake is van oorzaak en gevolg voor heel de populatie waaruit die steekproef getrokken is.
Als je ervan uitgaat dat zo goed als alle directeuren van die bank mannen zijn, dan kan je de onderzoeksvraag formuleren als: "Is het waar dat die bank de vrouwelijke bankbedienden bij promotie discrimineert ten opzichte van de mannelijke?".

2.2. Het verzamelen van de data

De data zijn al verzameld en samengevat in de bovenstaande tabel.

Het lukraak toekennen van dossiers aan bankdirecteuren kan je ook als volgt bekijken. Je beslist vooraf dat je de mannelijke dossiers aan de eerste groep zal geven en de vrouwelijke dossiers aan de tweede. Die twee groepen maak je nu door randomizatie. Je kan bijvoorbeeld elk van die directeuren een volgnummer geven van 1 tot 48. Daarna gebruik je het programma TREKZNDR om uit de eerste 48 getallen er lukraak een groepje van 24 te trekken. Aan dat groepje geef je de mannelijke dossiers en aan het resterende groepje de vrouwelijke.

2.3. De analyse van de data

Zowel de respons als de verklarende veranderlijke zijn hier categorisch met elk twee categorieën. Je kan dus te werk gaan zoals bij de studie over pesticidenresidu's.

Vervolledig de tabel en vul de randtotalen in. Zorg ervoor dat je aan elk getal in de tabel de juiste interpretatie kan geven.

		Aanbevolen voor promotie		
		Ja	Neen	Totaal
Geslacht	Man	21	3	**24**
	Vrouw	14	10	**24**
	Totaal	**35**	**13**	**48**

In totaal zijn er 48 dossiers. De kruistabel toont hoeveel (= de frequentie) dossiers er zijn voor elke combinatie van de categorieën van de twee categorische veranderlijken. Zo zie je bijvoorbeeld dat er 21 dossiers zijn die tegelijkertijd aan de volgende 2 eigenschappen voldoen: het gaat over een man en de promotie wordt aanbevolen.

De onderzoeksvraag gaat over een verschil in promotie voor mannen en vrouwen. In deze studie is het totaal aantal mannelijke dossiers gelijk aan het totaal aantal vrouwelijke en je zou dus de "aantallen" aanbevolen promoties kunnen vergelijken. Bij heel wat studies (zoals bij

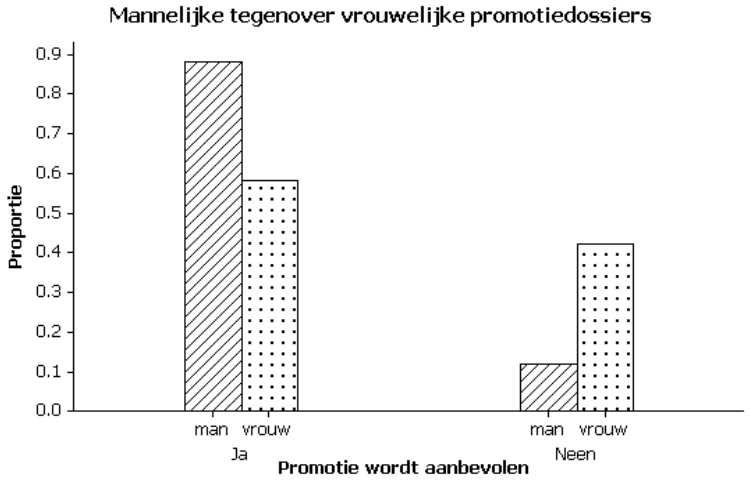
biovoeding en pesticidenresidu's) zijn de totale aantallen van de subgroepen niet gelijk. Daar moet je per subgroep de proporties vergelijken. Doe dat hier ook zo.

Proporties per subgroep bestuderen betekent dat je werkt met **conditionele** (of **voorwaardelijke**) proporties. Conditioneel op het feit dat de dossiers over mannen gaan, kan je kijken naar de proportie dossiers die daar zijn aanbevolen voor promotie. Dat is $21/24 \cong 0.88$. Bij de vrouwelijke dossiers is de conditionele proportie van dossiers die voor promotie zijn aanbevolen gelijk aan $14/24 \cong 0.58$.

De kruistabel met conditionele proporties ziet er als volgt uit.

		Aanbevolen voor promotie		
		Ja	Neen	Totaal
Geslacht	Man	21	3	**24**
		0.88	0.12	**1.00**
	Vrouw	14	10	**24**
		0.58	0.42	**1.00**

Grafisch kan dit resultaat voorgesteld worden door een staafdiagram met subtypes.



2.4. Interpretatie van de resultaten

Informatie uit grafieken en kengetallen

Is er wel discriminatie?

Om dat te onderzoeken kijk je even naar wat je zou verwachten als er geen discriminatie is.
In totaal zijn er 35 van de 48 dossiers voorgesteld voor promotie. Dat is $35/48 \cong 73\,\%$ van alle dossiers. Zonder discriminatie verwacht je dan dat er (ongeveer) 73 % van de mannen en ook (ongeveer) 73 % van de vrouwen promotie krijgen. In deze studie echter werden 88 % van de mannen en 58 % van de vrouwen voor promotie voorgesteld. Dat is toch wel een groot verschil. Dit laat vermoeden (en het wordt in het facultatief deeltje hieronder statistisch aangetoond) dat er bij die 48 bankdirecteuren een verband is tussen het geslacht van de

kandidaat en het voorstel tot promotie. Aangezien er een goede steekproef getrokken is kan deze conclusie veralgemeend worden tot de hele populatie van alle mannelijke bankdirecteuren van die bank.

Informatie uit het ontwerp van de studie

Deze studie is opgezet als een experiment. De 48 deelnemers werden op een goede manier over twee groepen gerandomiseerd en de onderzoeker bepaalde de behandeling (een mannelijk of een vrouwelijk dossier ter beoordeling voorgelegd krijgen). Als er dus een verschil in respons is dan is dat te wijten aan de behandeling. Als vrouwelijke dossiers in een beduidend mindere mate worden aanbevolen voor promotie dan is dat te wijten aan het feit dat er op die dossiers staat dat het om een vrouw gaat. Dat is een bewijs van discriminatie.

2.4.1. Facultatief deeltje: een statistisch bewijs van discriminatie

Als er helemaal geen discriminatie is, dan verwacht je dat er ongeveer 17 à 18 mannen voor promotie worden aanbevolen. Dat is 73 % van de mannen en dat is ook gelijk aan het percent van de hele groep (mannen + vrouwen) dat een promotie krijgt. En dan krijgt ook 73 % van de vrouwen een promotie. Maar bij een klein verschil waarbij in plaats van 17 of 18 er 19 mannen promotie krijgen kan je nog niet over duidelijke discriminatie van vrouwen spreken. Vanaf wanneer vind je dat er echt te veel mannen promotie krijgen? Vanaf wanneer wordt er volgens jou echt gediscrimineerd?

In de statistiek redeneer je nu als volgt.

Onderstel eens dat er echt geen discriminatie is en dat dossiers voor promotie worden aanbevolen zonder dat men naar het geslacht gekeken heeft. Dan heeft elk dossier dezelfde kans op promotie want behalve het geslacht zijn zij identiek. Als elk dossier dezelfde kans heeft, en als je uit die groep van 48 er lukraak 35 selecteert voor promotie, wat is dan de kans dat, door puur toeval en zonder dat het geslacht een rol speelt, er 20 of 21 of… mannen tussen zitten?

Om een idee te krijgen over de kans om, door puur toeval, 20 of 21 of… mannen te vinden kan je het volgende spel spelen. Neem een boek kaarten (dat zijn er 52) en verwijder de vier azen. Nu heb je 48 kaarten waarvan er 24 zwart zijn (noem dat de mannen) en 24 rood (noem dat de vrouwen). Schud de kaarten heel goed en deel dan 35 kaarten op tafel (dat zijn de 35 promoties). Draai de kaarten om en tel hoeveel zwarte kaarten (mannen) er zijn. Dat zijn er deze keer bijvoorbeeld 16.

Gebeurt het veel of weinig dat je 16 zwarte kaarten hebt als je, na goed schudden, er lukraak 35 van de 48 deelt? Wat is bij dit spel de kans op 16 zwarte kaarten? Dat kan je (benaderend) te weten komen als je dit spel heel veel keren speelt (in theorie een oneindig aantal keren) en kijkt naar de verhouding van het aantal keren dat je 16 zwarte kaarten hebt ten opzichte van het totaal aantal keren dat je speelt (dat is dus de relatieve frequentie van de uitkomst “16 zwarte kaarten”).

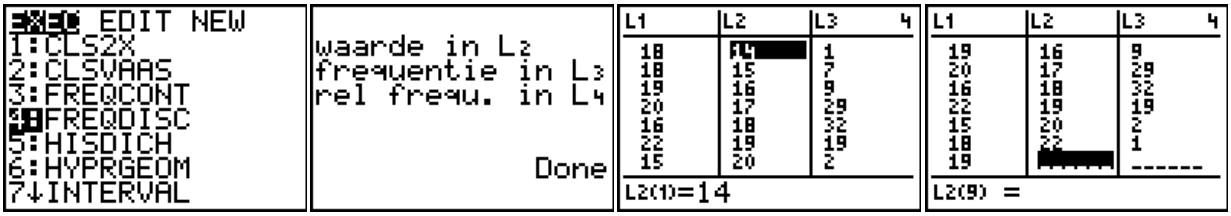
De kans op 16 (of 17 of…) zwarte kaarten wordt gegeven door de hypergeometrische verdeling. Misschien heb je die bestudeerd in de wiskundeles. Maar in ieder geval kan je die kansen benaderen door een eenvoudige simulatie met je GRM. Je gebruikt daarvoor het programma HYPRGEOM. Hoe dat werkt lees je hieronder.

Denk aan een vaas met een totaal van T kaartjes (in dit voorbeeld is T=48). Op S "succes"-kaartjes staat een 1 (bij jou is S=24) en op de overige T–S "mislukking"-kaartjes staat een 0 (hier is T–S=24). De benaming "succes" en "mislukking" heeft niets te maken met goed of slecht. Het zijn klassieke woorden in de statistiek die verwijzen naar wat je aan het tellen bent in situaties waarbij er slechts twee mogelijkheden zijn. Als je bij de promotie van mannen en vrouwen het aantal mannen telt dan noem je "man" = "succes".
Je trekt nu zonder terugleggen lukraak 35 kaartjes uit die vaas en noteert hoeveel successen je in je steekproef hebt (hoeveel kaartjes met een 1 = hoeveel mannen). Werk samen zodat je "35 kaartjes trekken uit die vaas" 1000 keer kan herhalen. Zo vind je een goede benadering van de kansen.

Verzamel de resultaten van 10 leerlingen die elk 100 keer een steekproef van 35 kaartjes uit die vaas trekken. Doe dat als volgt.

Tik [PRGM] loop naar HYPRGEOM en tik 2 keer [ENTER]. Zeg eerst wat er in die vaas zit, namelijk een totaal van T=48 kaartjes waaronder er S=24 successen zijn. Daarna geef je aan dat je uit die vaas n=35 kaartjes wil trekken zonder terugleggen. Tenslotte zeg je dat je dit allemaal 100 keer wil herhalen. Hou er rekening mee dat je GRM nu ongeveer 2 minuten nodig heeft om de opdracht uit te voeren. Dan krijg je de boodschap dat het aantal successen voor elk van die 100 herhalingen in de lijst [L1] staat.

Nu moet je tellen welke succesaantallen er allemaal voorkomen en hoe dikwijls. Daarvoor gebruik je FREQDISC om een FREQuentietabel op te stellen voor DISCrete gegevens. Tik [PRGM] loop naar FREQDISC en tik 2 keer [ENTER].



De gevonden succesaantallen staan in [L2] en de bijhorende frequenties in [L3]. Let op bij het overschrijven want een ontbrekend succesaantal wordt niet aangegeven. In dit voorbeeld is het bij die 100 herhalingen 2 keer gebeurd dat er op 20 van de 35 kaartjes een 1 stond. Ook kwam het 1 keer op de 100 voor dat er op 22 van de 35 kaartjes een 1 stond. Maar dat er op 21 van de 35 kaartjes een 1 stond kwam bij deze 100 herhalingen niet voor. Bij "21 successen" moet je hier dus een frequentie gelijk aan nul zetten.

Schrijf je resultaten nu over in de onderstaande tabel samen met de resultaten van je medeleerlingen. Vervolledig daarna de (benaderende) kansuitspraken. Is er statistisch bewijs van discriminatie? Waarom?

	Aantal successen											
	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
leerling 1												
leerling 2												
leerling 3												
leerling 4												
leerling 5												
leerling 6												
leerling 7												
leerling 8												
leerling 9												
leerling 10												
TOTAAL												
relatieve frequentie												

De relatieve frequenties zijn benaderingen voor kansen en we zullen voor de verdere redenering over kansen spreken.

Als er bij promotie een voorkeur voor mannen is dan is het duidelijk dat er meer zullen zijn dan de verwachte 17 à 18. Maar als de bank nu echt blijft volhouden dat er helemaal geen discriminatie is, hoe groot is dan de kans dat het aantal mannen gewoon door toeval “extreem groot” is? In de statistiek kijk je daarbij naar extreme “gebieden” en in dit voorbeeld is dat naar het rechtse gebied waar de grote aantallen successen (= mannen) zitten.

Als er alleen toeval speelt en niets anders, dan zegt bovenstaande simulatie (benaderend):

- de kans om bij die 35 promoties bij toeval 24 mannen te hebben ≅ ……
- de kans om bij die 35 promoties bij toeval minstens 23 mannen te hebben ≅ ……
- de kans om bij die 35 promoties bij toeval minstens 22 mannen te hebben ≅ ……
- de kans om bij die 35 promoties bij toeval minstens 21 mannen te hebben ≅ ……

In de statistiek is het gebruikelijk om te zeggen dat een gebeurtenis "zeldzaam" is wanneer die gebeurtenis hoogstens 5 keer voorkomt bij 100 herhalingen. Wat heb je hier nu gezien?

Waarschuwing.

Deze studie heeft echt plaatsgevonden omdat men vanuit toegepaste psychologie wou nagaan of het geslacht een (bewuste of onbewuste) rol speelt bij de beoordeling van mensen. Maar in deze studie zit een onnatuurlijk element. Alle dossiers waren hier, op het geslacht na, identiek. In werkelijkheid ziet voor elke bankbediende zijn of haar dossier er anders uit. Als je in een analoge situatie een bedrijf van discriminatie wil beschuldigen dan moet je ook kunnen aantonen (bijvoorbeeld met onafhankelijke experts) dat de voorgelegde dossiers evenwaardig zijn.

3. Zelfevaluatie

Hieronder staan enkele studies die door anderen werden uitgevoerd. Lees ze aandachtig, bespreek ze in groep en beantwoord de vragen die na elke studie staan.

3.1. Binge drinken en geslacht

Eén van de definities van binge drinken die je op het web kan vinden luidt als volgt: "Meer dan tien glazen alcohol na elkaar voor mannen en meer dan zeven glazen voor vrouwen, en dat in zo kort mogelijke tijd, zelfs met de chronometer erbij. Er bestaat zelfs een naam voor: binge drinken. Vrij vertaald: zuipen tot je erbij neervalt."

Bij een studie in 2001 van de Harvard universiteit werd aan 10 904 lukraak geselecteerde hogeschoolstudenten gevraagd naar hun drinkgewoonte. In het bijzonder wou men weten of zij al dan niet aan "binge drinken" deden. Bij die studie werd ondermeer ook het geslacht van de respondent genoteerd.

		Binge drinker		
		Ja	*Neen*	*Totaal*
Geslacht	*Jongen*	*1908*	*2017*	***3925***
	Meisje	*2854*	*4125*	***6979***
	Totaal	***4762***	***6142***	***10 904***

- Is er bij de beschrijving van deze studie informatie over de manier waarop de deelnemers in die studie zijn terechtgekomen? Welke gevolgen heeft dat?

- Wat is in deze studie de respons en wat is de verklarende veranderlijke?

- Is dit een observatiestudie? Motiveer je antwoord.

- Is deze studie retrospectief, cross-sectioneel of prospectief? Verklaar.

- Hoeveel jongens zeggen dat zij binge drinker zijn en hoeveel meisjes zeggen dat in deze studie?

- Kan je de twee getallen uit vorige vraag gebruiken om een antwoord te geven op: “Is de proportie binge drinkers bij Amerikaanse hogeschoolstudenten dezelfde bij de jongens als bij de meisjes?”. Verklaar.

- Maak een contingentietabel met de conditionele proporties van de ondervraagde studenten die al dan niet binge drinker zijn, conditioneel op het geslacht.

		Binge drinker		
		Ja	Neen	Totaal
Geslacht	Jongen			
	Meisje			

- Gebruik de tabel die je zopas gemaakt hebt om iets te zeggen over een eventuele samenhang tussen geslacht en binge drinken. Over wie gaat je uitspraak dan?

3.2. De autogordel en dodelijke ongevallen

Voorafgaande opdracht.

Bij het vergelijken van twee groepen werk je met “proportie per groep” = ”conditionele proportie”. Tot nu toe heb je gekeken naar het verschil van die proporties. Als dat verschil groot is dan vermoed je een samenhang tussen de verklarende veranderlijke en de respons.

Kijken naar een verschil is niet altijd de beste manier om dingen te vergelijken. Veel heeft te maken met grootteorde. Je doet dat spontaan voor dingen in het dagelijkse leven. Als je zopas een LCD televisie gekocht hebt voor 1199 euro en je komt daarna een winkel tegen waar hetzelfde toestel 1198 euro kost dan zeg je dat er eigenlijk geen prijsverschil op die TV zit. Maar als je voor een snoepje € 1.99 betaalt en je ziet daarna dat het slechts € 0.99 kost op een

andere plaats dan voel je je bedrogen omdat jij “meer dan het dubbele” betaald hebt. In beide voorbeelden gaat het over een verschil van 1 euro. Maar in het tweede geval gaat het over een klein bedrag en daar kijk je niet naar het verschil maar daar kijk je automatisch naar een verhouding: 1.99 is meer dan 2 keer 0.99.

Als je bij kruistabellen te maken hebt met “kleine” conditionele proporties dan is het dikwijls beter om niet naar het verschil te kijken maar naar de verhouding. Zo’n verhouding wordt “relatief risico” genoemd. Je kan hierover iets meer leren in de tekst “Kruistabellen: achtergrondinformatie” op http://www.uhasselt.be/lesmateriaal-statistiek . De bladzijde uit die tekst die je hier nodig hebt is hieronder overgenomen. Lees die aandachtig zodat je ze goed begrijpt.

Relatief risico.

In een experimentele fase worden geneesmiddelen uitgetest op proefdieren, ondermeer om de schadelijke neveneffecten te onderzoeken. Als de proportie proefdieren die neveneffecten vertoont 0.49 is voor geneesmiddel A en 0.48 voor geneesmiddel B dan lijkt dat goed in elkaars buurt te liggen. Maar wanneer deze geneesmiddelen uiteindelijk op de markt gebracht worden en de neveneffecten bij mensen treden op met een proportie van 0.011 (elf op duizend) in het ene geval en 0.001 (één op duizend) in het andere, dan lijkt dit een belangrijke vermindering voor geneesmiddel B. Als je echter het verschil in proporties uitrekent, dan is dat hetzelfde, zowel bij de proefdieren als bij de mensen. Dat verschil is 0.01.

Het kan dus verstandig zijn om, naast het verschil in proporties, ook andere maatstaven te hanteren om samenhang te bestuderen.

Een andere maat voor de samenhang in een 2x2 tabel is het relatief risico. Hierbij vergelijk je het risico dat je loopt bij het ene geneesmiddel met het risico dat je loopt bij het andere. Bij een concrete dataset betekent dit dat je de verhouding uitrekent van twee conditionele proporties.

$$relatief\ \ risico = \frac{proportie\ \ met\ \ neveneffecten\ bij\ \ geneesmiddel\ \ A}{proportie\ \ \ met\ \ neveneffecten\ bij\ \ geneesmiddel\ \ B}$$

Als je naar de definitie kijkt dan zie je dat het relatief risico een getal is dat berekend wordt als een verhouding van twee proporties. De uitkomst kan dus gelijk welk positief getal zijn. Als het relatief risico groter is dan één, dan is het risico groter bij geneesmiddel A dan bij B. Bij een relatief risico kleiner dan één is het juist andersom. En een relatief risico dat gelijk is aan één wijst erop dat er geen samenhang is tussen het al dan niet krijgen van neveneffecten en het soort gebruikte geneesmiddel (A of B).

Bij steekproefresultaten moet je natuurlijk altijd rekening houden met variabiliteit, context, manier van opmeten, enz. Een relatief risico kan je goede aanwijzingen geven over samenhang, maar definitieve conclusies kan je pas trekken met methoden van de verklarende statistiek. In het voorbeeld over de schadelijke neveneffecten van geneesmiddelen zie je dat, bij de uitgevoerde studies, het relatief risico gelijk was aan 11 bij mensen terwijl het slechts 1.02 was bij proefdieren.

De autogordel.

In de volgende tabel zie je gegevens over auto-ongevallen en het al dan niet dragen van een gordel. De gegevens zijn afkomstig van het "Department of Highway Safety and Motor Vehicles" van Florida. Ze werden daar recent gedurende twee jaar opgemeten in de stad Jacksonville.

		Ongeval met dodelijke afloop		
		Ja	*Neen*	*Totaal*
Droeg een gordel	*Ja*	*51*	*41 237*	***41 288***
	Neen	*160*	*16 253*	***16 413***

- Wat is de respons en wat is de verklarende veranderlijke? Van welk type zijn ze?

- Welk soort studie is dit? Waarom? Zou je er ook een studie van het andere soort kunnen van maken? Hoe zou je dat doen en wat onderzoek je dan precies? Is dat ethisch verantwoord?

- Maak een contingentietabel met conditionele proporties.

- Is er een sterke samenhang in deze studie te bespeuren als je naar een verschil in proporties kijkt? Hoe komt dat?

- Werk nu met het relatief risico en kies daarbij de teller en noemer op die manier dat je later een uitspraak kan doen die door iedereen eenvoudig te begrijpen is. Bemerk je nu een sterke samenhang in deze studie? Welke conclusie kan je hier formuleren?

3.3. Aspirine en hartinfarct

In de jaren 80 werd aan de universiteit van Harvard een studie verricht waarbij 22 071 mannelijke artsen betrokken waren. Zij werden lukraak in 2 groepen verdeeld. De artsen moesten gedurende 5 jaar om de twee dagen een pilletje innemen. Ofwel kreeg een arts 5 jaar lang aspirine te slikken ofwel een placebo, maar hij wist het niet. Gedurende die 5 jaar werden de deelnemers aan de studie opgevolgd om te kijken wie er in die periode een hartinfarct kreeg. Die opvolging gebeurde door specialisten die niet wisten welke soort pil de deelnemer genomen had.
De resultaten waren als volgt.

		Kreeg hartinfarct		*Totaal*
		Ja	*Neen*	
Medicatie	*Placebo*	*239*	*10 795*	*11 034*
	Aspirine	*139*	*10 898*	*11 037*

Een video over deze studie kan je bekijken op:
http://www.learner.org/channel/courses/learningmath/video/data/wmp/dat_06_ch4.html

- Is er bij de beschrijving van deze studie informatie over de manier waarop de deelnemers in die studie zijn terechtgekomen? Geef aan wat je moet onderstellen om tot een algemene conclusie te komen. Hoever reikt die conclusie dan?

- Zoek in de onderstaande tabel alle woorden die op deze studie van toepassing zijn. Zeg ook wat zij hier concreet betekenen. Geef voldoende uitleg en verantwoording zodat je tot een volledige bespreking van dit onderzoek komt.

randomizatie	placebo	verstrengelende factor
observatiestudie	associatie	lukrake steekproef
retrospectief		
cross-sectioneel	respons	oorzakelijk verband
prospectief		
	dubbelblind	populatie
controlegroep		
	experiment	verklarende veranderlijke

3.4. Thee met melk

De volgende studie staat bekend als "Fisher's tea tasting experiment". Fisher was een beroemde geneticus en statisticus. Als echte Engelsman was hij dol op de "afternoon tea". Zijn collega Muriel Bristol nam altijd melk in haar thee. Zij vond dat zo'n kopje lekkerder was als je eerst de melk en dan de thee inschonk in plaats van eerst de thee en dan de melk. Fisher wou uitproberen of zij daar echt wel een onderscheid kon tussen maken. Hij deed het volgende.

Er werden 8 identieke kopjes op een rij gezet, genummerd van 1 tot 8. Op een lukrake manier werden 4 getallen uit die 8 getallen getrokken. Bij de kopjes die met de getrokken nummers overeenstemden werd eerst melk ingeschonken en bij de andere eerst thee. Pas dan mocht Muriel Bristol binnenkomen en kopje na kopje proeven. Er was haar vooraf gezegd dat er bij 4 kopjes eerst melk was ingeschonken en bij de andere 4 eerst thee. Het resultaat van dit onderzoek was als volgt.

		Muriel zegt dat er eerst is ingeschonken:		
		Melk	*Thee*	*Totaal*
In feite is er eerst ingeschonken:	*Melk*	*3*	*1*	***4***
	Thee	*1*	*3*	***4***
	Totaal	***4***	***4***	***8***

- Wat is de respons en wat is de verklarende veranderlijke? Van welk type zijn ze?

- Wat is hier de onderzoeksvraag? Let op de juiste formulering

- Welk soort onderzoek is dit? Was er randomizatie bij het ontwerp van deze studie? Als er samenhang is, welke conclusie kan je dan trekken? Kan je veralgemenen?

- Maak nu eens een “ideale” tabel waarbij er helemaal geen samenhang zou zijn tussen wat Muriel zegt en wat er echt met die kopjes gebeurd is. Kijk goed naar die tabel en bedenk dat, door louter toevallige schommelingen, er enige afwijkingen van de “ideale” tabel kunnen zijn wanneer je een studie uitvoert. De kleinst mogelijke “toevallige afwijking” brengt je al bij het resultaat van Muriel. Het aantal elementen in deze studie (2 keer 4 kopjes) is zeer klein. Wat leer je hieruit?

3.4.1. Facultatief deeltje: statistisch bewijs bij een zeer kleine steekproef

Probeer nu voor het "Tea tasting experiment" een statistisch bewijs te geven op basis van een simulatie. Denk daarom over dit experiment als volgt.

In feite zijn er 4 kopjes waarin "eerst melk" werd geschonken. Noem dat succes en stop 4 kaartjes met een 1 erop in een vaas. Er zijn ook 4 kopjes waarin "eerst thee" werd geschonken. Noem dat mislukking en stop ook 4 kaartjes met een 0 erop in die vaas. Iemand die helemaal het verschil niet kent tussen "eerst melk" of "eerst thee" moet gokken. Dat komt erop neer dat hij "eerst melk" zegt tegen 4 lukraak getrokken kaartjes uit die vaas. Wat is nu de kans dat er tussen die 4 getrokken kaartjes "ten minste" 3 kaartjes met een 1 (= eerst melk geschonken) zitten?

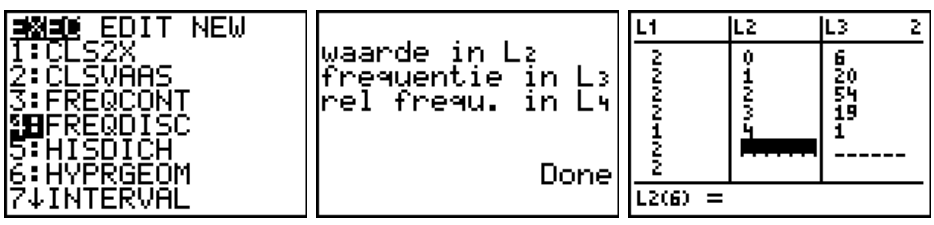
In de statistiek kijk je naar "zeldzame uitkomsten" in "extreme **gebieden**" en dus heb je de kans nodig om "**minstens** 3 van de vier gokken" juist te hebben. Komt dat meer dan 5 keer op 100 voor, puur door toeval en zonder er iets van te kennen? Of is dat zeldzaam en komt het bij puur gokwerk hoogstens 5 keer op 100 voor? Zoek dat nu uit met een simulatie.

Je kan de kansen benaderen met het programma HYPRGEOM. Hoe dat werkt heb je vroeger in deze tekst geleerd.

Voor de huidige studie denk je aan een vaas met een totaal van T=8 kaartjes. Op S "succes"-kaartjes staat een 1 (hier is S=4) en op de overige T–S "mislukking"-kaartjes staat een 0 (hier is T–S=4).

Je trekt nu zonder terugleggen lukraak 4 kaartjes uit die vaas en noteert hoeveel successen je in je steekproef hebt (hoeveel kaartjes met een 1). Doe dat als volgt.

Tik [PRGM] en loop naar HYPRGEOM en tik 2 keer [ENTER]. Zeg eerst wat er in die vaas zit, namelijk een totaal van T=8 kaartjes waaronder er S=4 successen zijn. Daarna geef je aan dat je uit die vaas n=4 kaartjes wil trekken zonder terugleggen. Tenslotte zeg je dat je dit allemaal 100 keer wil herhalen. Hou er rekening mee dat je GRM nu ongeveer 1 minuut nodig heeft om de opdracht uit te voeren. Dan krijg je de boodschap dat het aantal successen voor elk van die 100 herhalingen in de lijst [L1] staat. Nu moet je tellen welke succesaantallen er allemaal voorkomen en hoe dikwijls. Daarvoor gebruik je FREQDISC om een FREQuentietabel op te stellen voor DISCrete gegevens. Tik [PRGM] en



loop naar FREQDISC en tik 2 keer [ENTER]. Schrijf je resultaten nu over in de onderstaande tabel. Herhaal dit met 5 leerlingen elk 4 maal. Zo heb je een simulatie met 2000 herhalingen.

	Aantal successen				
	0	**1**	**2**	**3**	**4**
leerling 1a					
leerling 1b					
leerling 1c					
leerling 1d					
leerling 2a					
leerling 2b					
leerling 2c					
leerling 2d					
leerling 3a					
leerling 3b					

	Aantal successen				
	0	**1**	**2**	**3**	**4**
leerling 3c					
leerling 3d					
leerling 4a					
leerling 4b					
leerling 4c					
leerling 4d					
leerling 5a					
leerling 5b					
leerling 5c					
leerling 5d					
TOTAAL					
relatieve frequentie					

Bij een “statistisch” bewijs mag je niet alleen maar kijken naar proporties. Ook de grootte van de steekproef speelt een belangrijke rol.

3.5. Pepsi of Coke

3.5.1. Facultatief deeltje: statistisch bewijs en steekproefgrootte

Dit project is een vervolg op het facultatief deeltje van "Thee met melk".

De grootte van de steekproef is enorm belangrijk. Je kan dat met een extreem voorbeeld gemakkelijk inzien.

Jij beweert dat je helderziende bent en dat je weet op welke zijde een muntstuk zal vallen.

- Jij zegt: "De eerste keer kruis en de tweede keer munt". Ik gooi twee keer en inderdaad, het muntstuk valt de eerste keer op kruis en de tweede keer op munt. In dit experiment waren al je antwoorden juist.
- We spelen hetzelfde spel maar gooien nu 10 keer. En jij hebt 10 keer de juiste zijde voorspeld. In dit experiment waren al je antwoorden terug juist.

Bij beide experimenten is het percent juiste antwoorden hetzelfde: 100 %. Dus zou je denken dat de conclusie in beide gevallen dezelfde moet zijn. Dit is niet waar. De grootte van de steekproef speelt ook een rol, zoals hieronder uitgelegd.

Waarom is het eerste experiment geen bewijs van je helderziendheid? Omdat er, door puur toeval, al 25 % kans is om (K,M) = eerst kruis en dan munt te hebben. Er zijn immers maar 4 (evenwaardige) mogelijkheden: (K,K) (K,M) (M,K) en (M,M). Een kans van 25 % is veel groter dan de klassieke drempel van 5 % die in de statistiek aangenomen wordt om te zeggen dat het geen puur toeval meer is maar dat er een andere oorzaak is.

Waarom is het tweede experiment wel een statistisch bewijs van je helderziendheid? Omdat 10 successen bij 10 herhalingen slechts voorvalt met kans 0.001 (1 keer op duizend!) wanneer jij er niets van kent en er alleen puur toeval in het spel is. Dat dit nu juist bij jou gebeurt, dat is dus zeer onwaarschijnlijk. De andere verklaring is nu veel logischer: jij bent echt helderziende.

In statistiek mag je niet alleen "in verhoudingen" denken. De grootte van de steekproef speelt een cruciale rol. Om dit te ervaren ga je het "Tea tasting experiment" nabootsen waarbij je niets verandert behalve dat je alles vermenigvuldigt met 3. Je kan dit zelf spelen in je klas als er een leerling is die beweert dat hij bij het proeven weet wat Pepsi Cola en wat Coca Cola (Coke) is.

Neem 24 bekertjes en plak daar nummers op, van 1 tot 24. Als je nu in de eerste 12 Pepsi giet en in de rest Coke, dan is er een kans dat de deelnemer dat raadt. Na enkele bekertjes geproefd te hebben zegt hij dan systematisch bijvoorbeeld Pepsi tegen de lagere nummers en Coke tegen de hogere. Om dat te vermijden ga je als volgt te werk.

Zet alle bekertjes op een rij, in volgorde van 1 tot 24. Neem je GRM en roep het programma TREKZNDR op. Zeg dat de grootte van de totale populatie gelijk is aan 24 (T=24) en dat je een steekproef van grootte $n=12$ wil trekken. Op die manier krijg je 12 lukrake getallen in de lijst [L1] . In de bekertjes die overeenstemmen met die getallen giet je een beetje Coke. In de overige 12 bekertjes giet je wat Pepsi. Noteer voor jezelf wat bij welk nummer hoort maar laat de bekertjes op hun plaats staan, in volgorde.

Pas nu mag de deelnemer de klas binnenkomen (of anders zorg je er in ieder geval voor dat hij van die voorbereiding niets gezien heeft). Je zegt tegen de deelnemer dat er 12 bekers zijn met Pepsi en 12 met Coke en je geeft hem 12 stickers met het woord Pepsi op en 12 stickers met het woord Coke. Dan proeft hij, in volgorde, van elk bekertje en legt één sticker per bekertje (een Pepsi-sticker of een Coke-sticker). Terwijl hij bezig is mag niemand reageren. Je mag dus tijdens het proeven niet zeggen of het juist of fout is. Op het einde heeft de deelnemer 12 bekertjes geïdentificeerd die volgens hem Pepsi bevatten. In de andere 12 zit er dus volgens hem Coke.

Maak nu een tabel om het resultaat van deze proef samen te vatten.

		De deelnemer zegt: Pepsi	Coke	Totaal
In feite is het:	Pepsi			**12**
	Coke			**12**
	Totaal	**12**	**12**	**24**

DISCUSSIEMOMENT 5.

Je kan vooraf, met behulp van een simulatie, zeggen hoe groot het getal in de eerste cel **minstens** moet zijn om een "statistisch bewijs" te hebben dat die leerling de cola's echt kan identificeren. Leg kort uit hoe dat werkt door de onderstaande vragen te beantwoorden.

- Leg uit hoe je de simulatie zal doen.
  - Welk programma zal je daarvoor gebruiken?
  - Welke getallen moet jij ingeven bij dit programma?
  - Geef een korte verklaring bij elk van die in te geven getallen.
  - Wanneer dit programma ten einde is, welke getallen staan er dan in welke lijst?
  - Welk programma ga je daarna gebruiken? Waarom? Wat ga je dan noteren?

- Hoe zal je dan de bekomen resultaten gebruiken om te zeggen hoe groot het getal in de eerste cel **minstens** moet zijn om een "statistisch bewijs" te hebben? Hoeveel bekertjes die met Pepsi gevuld zijn moet hij **minstens** kunnen herkennen als Pepsi (want dat is het getal dat in de eerste cel zal terechtkomen)?

Werk met het programma HYPRGEOM voor de simulatie en denk daarbij aan kaartjes trekken uit een vaas. In die vaas zitten in totaal 24 kaartjes (totale populatie T=24) waarvan er 12 Pepsi kaartjes zijn. Dat zijn de succeskaartjes (aantal successen in de populatie S=12). Uit die vaas trek je, zonder terugleggen, 12 kaartjes (steekproef n=12). Het trekken van zo'n steekproef wil je veel keren herhalen (K=100). Als het programma ten einde is zegt het dat "het aantal successen" die bij die honderd herhalingen werden gevonden in de lijst [L1] staan. Om nu te weten hoeveel keren er 12 of 11 of …. successen waren bij die 100 herhalingen gebruik je FREQDISC. De gevonden succesaantallen staan in [L2] en de bijhorende frequenties in [L3]. Let op bij het overschrijven want een ontbrekend succesaantal wordt niet aangegeven.

Je bent enkel geïnteresseerd in de grotere getallen voor de eerste cel. Dus volstaat een tabel waarin je alleen maar noteert hoeveel keren er 12 of 11 of 10 of 9 of 8 of 7 successen waren. Werk terug samen zodat je 2000 herhalingen hebt. Doe daarna de (benaderende) kansuitspraken over 12 keer succes hebben, of over minstens 11 successen, of minstens 10 successen, enz. De kansen die je met de simulatie bepaalt, zijn gebaseerd op wat er gebeurt als alleen het toeval speelt. Je GRM kent niets van Pepsi of Coke.

	7	**8**	**9**	**10**	**11**	**12**
leerling 1_{ab}						
leerling 1_{cd}						
leerling 2_{ab}						
leerling 2_{cd}						
leerling 3_{ab}						
leerling 3_{cd}						
leerling 4_{ab}						
leerling 4_{cd}						
leerling 5_{ab}						
leerling 5_{cd}						

	7	**8**	**9**	**10**	**11**	**12**
TOTAAL						
relatieve frequentie						

- Gebruik je gevonden (benaderende) kansen om een statistisch besluit te trekken over de prestaties van Muriel wanneer zij niet 3 keer op 4 juist zou geweest zijn (zoals hierboven) maar 9 keer op 12 (zoals hieronder).

		Muriel zegt dat er eerst is ingeschonken: Melk	Thee	Totaal
In feite is er eerst ingeschonken:	Melk	9	3	**12**
	Thee	3	9	**12**
	Totaal	**12**	**12**	**24**

Conclusie

In beide voorbeelden heeft Muriel 75 % van de kopjes juist geïdentificeerd maar toch zie je dat 9 van de 12 kopjes juist hebben een veel sterkere prestatie is dan er 3 van de 4 juist hebben.

4. Facultatief deeltje: roken en gezondheid

4.1. Is roken gezond?

Heel veel studies zijn op zoek gegaan naar de mogelijke nadelen van roken.

Een studie in Engeland in de jaren 1970 kwam tot een vreemde conclusie. Aan 1314 vrouwen werd toen gevraagd of zij rookten. Twintig jaar later werd nagegaan welke vrouwen er nog in leven waren. Gedurende die 20 jaar waren er 31 % van de niet-rokers gestorven en slechts 24 % van de rokers.

DISCUSSIEMOMENT 6.

Als je alleen naar deze cijfers kijkt dan gaan er (in percentage per groep) minder dood van de rokers dan van de niet-rokers.

- Mag je hieruit de conclusie trekken dat roken niet ongezond is?
- Welk soort studie is dit?
- Waar vond die studie plaats? In welk jaar? Bij wie?
- Kan je de context van die studie gebruiken om een belangrijke verstrengelende factor op het spoor te komen?

Vergelijk jouw antwoorden met de onderstaande tekst.

In Engeland werd er in de jaren 1970 vooral gerookt door jongere vrouwen en het was hoogst ongebruikelijk in die tijd dat oudere vrouwen rookten.

De leeftijd is een factor die samenhangt met de verklarende veranderlijke, namelijk met het rookgedrag van die vrouwen.

De leeftijd hangt ook samen met de respons, namelijk met dood gaan. In een oudere leeftijdsgroep sterven er meer dan in een jongere.

Je hebt hier een extra veranderlijke ontdekt, namelijk leeftijd, die zowel met de verklarende veranderlijke als met de respons te maken heeft. Dat is een verstrengelende factor.

Een analyse van dezelfde studie, maar per leeftijdsgroep, geeft volgend resultaat.

	Percent overlijdens (binnen de volgende 20 jaar) per leeftijdsgroep (bij het begin van de studie)			
	18–34	35–54	55–64	65+
Roker	2.8 %	17.2 %	44.3 %	85.7 %
Niet-roker	2.7 %	9.5 %	33.1 %	85.5 %

Je ziet bijvoorbeeld dat van de vrouwen die in 1970 tussen de 35 en 54 jaar waren en die rookten er 17.2 % binnen de volgende 20 jaar overleden. Voor hen die niet rookten was dat binnen die leeftijdsgroep 9.5 %. In elke leeftijdsgroep sterft een groter percent rokers dan niet-rokers.

4.2. Roken is ongezond

Een observatiestudie kan geen oorzakelijk verband *bewijzen*. Dat betekent niet dat er geen oorzakelijk verband is. Maar als je dat met statistiek wil aantonen dan heb je strikt genomen een studie nodig die ontworpen is als een experiment. Anders kan een tegenstander beweren dat een verstrengelende factor (zoals luchtverontreiniging, voedingspatroon, enz.) de oorzaak van die longkanker is en niet het roken. Deze houding is 50 jaar lang door de tabaksindustrie aangenomen.

Om toch goede argumenten te kunnen ontwikkelen, kan je als volgt te werk gaan. Als men beweert dat luchtverontreiniging een verstrengelende factor is, vorm dan twee groepen: een groep die in verontreinigd industriegebied woont en een groep die in de gezonde buitenlucht woont. Kijk bij elke groep afzonderlijk naar rokers en niet-rokers. Als dan bij elke groep de rokers meer longkanker krijgen dan de niet-rokers dan geeft je dat een extra aanwijzing. Je hebt dan een associatie (geen oorzakelijkheid want ook dit is een observatiestudie) tussen roken en longkanker, zowel in verontreinigde gebieden als in niet-verontreinigde gebieden. En zo kan je verder gaan, voor andere factoren die eventueel een verstrengelende invloed zouden kunnen uitoefenen.

In de voorbije jaren zijn er enorm veel onderzoeken uitgevoerd naar de invloed van roken op de gezondheid.

Eén van die vele studies naar de invloed van roken was een prospectieve studie die 50 jaar duurde. Zij startte in 1951 en eindigde in 2001. In hun eindrapport schrijven de onderzoekers dat bij rokers die niet stoppen met roken de levensduur gemiddeld 10 jaar korter is. Zij zeggen ook dat de helft van wie vanaf zijn jeugd rookt sterft aan een ziekte die door roken is veroorzaakt. Een ander recent onderzoek schat dat 30 % van *alle* kankers toe te schrijven is aan roken.

Een variëteit van studies leveren samen een massa argumenten tegen het roken:

- bij dieren kan men wel werken met studies die ontworpen zijn als experimenten en dergelijke experimenten hebben bij die dieren een *oorzakelijk* verband aangetoond: roken *veroorzaakt* daar longkanker.

- in heel wat landen zijn vrouwen in vergelijking met mannen meer beginnen roken en in die landen is het krijgen van longkanker bij vrouwen in vergelijking met mannen gestegen.

- een groot aantal observatiestudies, zowel retrospectief als prospectief, hebben een associatie gevonden. Die studies werden uitgevoerd bij verschillende menselijke populaties (verschillende landen, rassen, …) en allemaal vonden zij een associatie, zelfs nadat de onderzoekers hadden rekening gehouden met alle eventuele verstrengelende factoren die werden gesuggereerd.

- als meer en meer studies worden ondernomen die rekening houden met een variëteit van verstrengelende factoren dan wordt de kans heel klein dat er nog ergens een ongekende verstrengelende factor zou zijn (en niet het roken) die het krijgen van longkanker zou verklaren.

Als je al die resultaten samenlegt dan kan een zinnig mens er niet meer onderuit: roken is schadelijk voor de gezondheid.

Ook de tabaksindustrie, na een fel verzet dat meer dan 50 jaar duurde, is uiteindelijk gezwicht voor de combinatie van al die evidentie. Zij geeft nu zelf toe dat roken schadelijk is.