Inductieve statistiek voor informatiewetenschappers

HENK VOORBIJ

1. Inleiding

Er zijn twee soorten statistiek: beschrijvende en inductieve (ook wel inferentiële genoemd). Het resultaat van een kwantitatief onderzoek is een spreadsheet met data in Excel of SPSS. Deze gegevens moeten we samenvatten in de vorm van tabellen, grafieken, centrummaten, spreidingsmaten of correlatiematen. Dit is het terrein van de beschrijvende statistiek. Het hoofdstuk over gebruikersonderzoek elders in dit Handboek geeft een beknopt overzicht (Voorbij, 2015). Bij inductieve statistiek is altijd een steekproef uit de populatie onderzocht. De centrale vraag is: welke uitspraak kunnen we nu doen over de populatie?

Ook voor informatiewetenschappers is het van belang enig inzicht te hebben in de achtergronden van de inductieve statistiek. Stel we verspreiden een vragenlijst onder een steekproef van wetenschappers om daarmee inzicht te krijgen in hun leesgedrag. Enkele, volstrekt fictieve, resultaten zijn:

1. Gemiddeld leest een wetenschapper 16,5 tijdschriften voor zijn werk.
2. 77,6% van de tijdschriften die wetenschappers lezen, is afkomstig van de eigen universiteitsbibliotheek.
3. Wetenschappers op STM gebied (Science, Technology and Medicine) lezen gemiddeld 19,2 tijdschriften; wetenschappers op andere gebieden lezen gemiddeld 14,8 tijdschriften. Een verschil dus van 4,4.
4. Er is een verband tussen het aantal tijdschriften dat men leest en het aantal publicaties dat men de afgelopen vijf jaar vervaardigd heeft (r=.480).

Deze resultaten zijn gebaseerd op de antwoorden van een beperkt aantal personen. Hoe kunnen we die nu doortrekken naar de gehele populatie? Met behulp van de inductieve statistiek kunnen we uitspraken doen waarbij we telkens twee slagen om de arm houden. Deze zien er, weer ingevuld met fictieve getallen, als volgt uit:

1. We zijn er 95% zeker van dat wetenschappers gemiddeld tussen 16,5 ± 3,4 tijdschriften lezen, dus ergens tussen 13,1 en 19,9.
2. We zijn er 95% zeker van dat wetenschappers gemiddeld tussen 77,6 ± 6,9% tijdschriften van de eigen universiteitsbibliotheek betrekken, dus ergens tussen 70,7% en 84,5%.
3. We zijn er 95% zeker van dat STM wetenschappers tussen 2,4 en 6,4 meer tijdschriften lezen dan wetenschappers op andere gebieden. Of, iets minder informatief: er is een significant verschil tussen de twee groepen in het aantal tijdschriften dat zij lezen (p=.036).
4. We zijn er 95% zeker van dat het verband tussen het aantal tijdschriften dat men leest en het aantal publicaties dat men de afgelopen vijf jaar heeft vervaardigd ergens tussen r=.401 en r=.552 ligt. Of, iets minder informatief: er is een significant verband tussen de twee variabelen (p=.028).

Dit soort uitspraken komen we veelvuldig tegen in de literatuur. Het doel van dit hoofdstuk is om de principes van de inductieve statistiek toe te lichten en daarbij het gebruik van formules tot een minimum te beperken. Enkele inleidende beschietingen komen aan de orde in paragraaf 2. Om het vervolg te begrijpen moeten we vertrouwd zijn met de begrippen standaarddeviatie, normaalverdeling en steekproevenverdeling. Paragraaf 3 is gewijd aan het schatten van het gemiddelde van de populatie op basis van het gemiddelde van de steekproef, paragraaf 4 doet ditzelfde voor percentages. Deze richten zich dus op het eerste en tweede punt van ons fictieve onderzoek. Paragraaf 5 gaat in op het toetsen van (een hypothese over) verschillen tussen groepen, paragraaf 6 op het toetsen van (een hypothese over) verbanden tussen twee variabelen. Daarmee raken we het derde en vierde punt van ons onderzoek. Het begrip significantie speelt daarbij een belangrijke rol. Paragraaf 7 handelt over het bepalen van de benodigde steekproefomvang. Dit lijkt een vreemde eend in de bijt. Er is echter een rechtstreekse samenhang met de paragrafen 3 en 4. Daarin worden formules gehanteerd die, als we daarmee spelen, ook gebruikt kunnen worden voor het bepalen van de steekproefomvang. Misschien is dit zelfs wel de

meest gestelde vraag van onderzoekers: 'Hoe groot moet mijn steekproef zijn?' Een slotwoord vormt paragraaf 8.

2. Enkele basisbegrippen

Om de principes van de inductieve statistiek te doorgronden moeten we vertrouwd zijn met enkele basisbegrippen. Deze paragraaf belicht de standaarddeviatie, normaalverdeling en steekproevenverdeling.

2.1 Standaarddeviatie

De standaarddeviatie is een maat voor de spreiding rond het gemiddelde. We nemen een passage over uit het hoofdstuk over gebruikersonderzoek.

Het gemiddelde zegt niet alles. De twee onderstaande reeksen hebben beide een gemiddelde van 14,0, maar verschillen sterk in spreiding
5, 7, 7, 8, 10, 10, 10, 12, 12, 14, 16, 17, 18, 20, 44
12, 12, 13, 13, 13, 14, 14, 14, 14, 14, 15, 15, 15, 16, 16

De meest gehanteerde spreidingsmaat is de standaarddeviatie (s). Deze wordt als volgt berekend:

1. Bepaal het gemiddelde m
2. Bereken het verschil van elke score x en het gemiddelde m
3. Kwadrateer elk verschil
4. Tel de gekwadrateerde scores op (Sum of Squares, ofwel SS)
5. Bereken de variantie (s^2) door SS te delen door het aantal respondenten (n) of het aantal respondenten min 1 (n – 1)
6. Bereken de standaarddeviatie (s) door de wortel te trekken uit de variantie

Reeks 1: m = 14,0			*Reeks 2: m = 14,0*		
$(5 - 14)^2$	=	81	$(12 - 14)^2$	=	4
$(7 - 14)^2$	=	49	$(12 - 14)^2$	=	4
$(7 - 14)^2$	=	49	$(13 - 14)^2$	=	1
$(8 - 14)^2$	=	36	$(13 - 14)^2$	=	1
$(10 - 14)^2$	=	16	$(13 - 14)^2$	=	1
$(10 - 14)^2$	=	16	$(14 - 14)^2$	=	0
$(10 - 14)^2$	=	16	$(14 - 14)^2$	=	0
$(12 - 14)^2$	=	4	$(14 - 14)^2$	=	0
$(12 - 14)^2$	=	4	$(14 - 14)^2$	=	0
$(14 - 14)^2$	=	0	$(14 - 14)^2$	=	0
$(16 - 14)^2$	=	4	$(15 - 14)^2$	=	1
$(17 - 14)^2$	=	9	$(15 - 14)^2$	=	1
$(18 - 14)^2$	=	16	$(15 - 14)^2$	=	1
$(20 - 14)^2$	=	36	$(16 - 14)^2$	=	4
$(44 - 14)^2$	=	900	$(16 - 14)^2$	=	4
		1236			22
a. Standaarddeviatie van volledige populatie Deel Sum of Squares door aantal cases → variantie s^2 = 1236 / 15 = 82,4 → standaarddeviatie s = $\sqrt{82,4}$ = 9,08			*a. Standaarddeviatie van volledige populatie* Deel Sum of Squares door aantal cases → variantie s^2 = 22 / 15 = 1,47 → standaarddeviatie s = $\sqrt{1,47}$ = 1,21		
b. Schatting van standaarddeviatie van populatie op basis van steekproef Deel Sum of Squares door aantal cases min 1 (n-1) → variantie s^2 = 1236 / 14 = 88,29 → standaarddeviatie s = $\sqrt{88,29}$ = 9,40			*b. Schatting van standaarddeviatie van populatie op basis van steekproef* Deel Sum of Squares door aantal cases min 1 (n-1) → variantie s^2 = 22 / 14 = 1,57 → standaarddeviatie s = $\sqrt{1,57}$ = 1,25		

Figuur 1. Berekening van de standaarddeviatie.

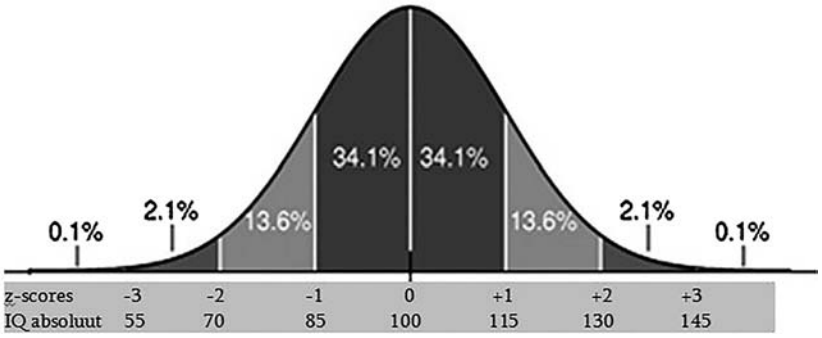
Inderdaad is de standaarddeviatie van de eerste reeks veel groter dan die van de tweede reeks. Standaarddeviaties krijgen pas betekenis door onderlinge vergelijking, de waarden zelf spreken niet echt tot de verbeelding. De hoogte van de standaarddeviatie bevindt zich altijd

ergens tussen de hoogste en laagste afwijking. In de eerste reeks dus tussen 0 en 30, in de tweede tussen 0 en 2.

Een complicatie is dat de standaarddeviatie op twee manieren berekend kan worden, afhankelijk van het doel. Als we de volledige populatie onderzocht hebben, delen we de Sum of Squares door n. Als we op basis van onze steekproefgegevens een uitspraak willen doen over de standaarddeviatie van de populatie, delen we de Sum of Squares door n – 1. SPSS kiest automatisch voor de tweede optie. In de praktijk maakt het echter weinig uit of we een getal nu delen door 400 of 399.

2.2. Normaalverdeling en standaardscores (z-scores)

Figuur 2 is een symmetrische verdeling. De top ligt precies in het midden, aan beide kanten is er een gelijkmatige uitwaaiering naar de uiteinden. Zo'n verdeling noemen we een normaalverdeling. We treffen dat aan bij het intelligentie quotiënt (IQ).



Figuur 2. Normaalverdeling.

Een normaalverdeling heeft bepaalde eigenschappen waar we in de inductieve statistiek veelvuldig gebruik van maken. Om maar met de deur in huis te vallen: 95% van de scores (waarnemingen, gevallen) bevindt zich in het middengebied met als begrenzing links het gemiddelde min 1,96 keer de standaarddeviatie en rechts het gemiddelde plus 1,96 keer de standaarddeviatie. Voor het IQ geldt een gemiddelde van 100 en een standaarddeviatie van 15. Deze getallen lijken te mooi om waar te zijn, maar IQ tests zijn met opzet zo ingericht dat ze

deze uitslagen geven. We kunnen dan ook zeggen dat 95% van de mensen een IQ heeft dat ligt tussen 100 ± (1,96 x 15), ofwel tussen 70,6 en 129,4. Slechts 2,5% heeft een IQ lager dan 70,6, slechts 2,5% heeft een IQ hoger dan 129,4. Van een willekeurig persoon kan ik nu met een zekerheid van 95% zeggen dat deze een IQ heeft ergens tussen 70,6 en 129,4.

De uitdrukking 'een score van het gemiddelde plus 1,96 keer de standaarddeviatie' is een mondvol. We spreken daarom liever van een standaardscore of z-score van 1,96, kortweg z = 1,96. In statistische formules die we in het vervolg zullen tegenkomen zien we vaak het element z. Vaak vult men hier de waarde 1,96 in, omdat dit de begrenzing is van het centrale 95% gebied.

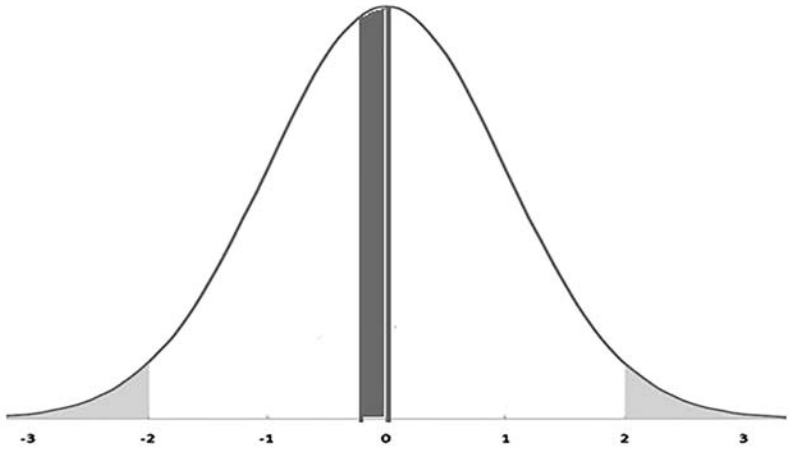
In figuur 2 zien we dat het gemiddelde een z-score heeft van 0. Naar rechts loopt dit op tot een z-score van rond de drie. Dat de z-scores aan de linkerkant een negatieve waarde hebben duidt erop dat ze onder het gemiddelde zijn. Ook nu loopt dit op tot een score van rond de min drie voor de meer extreme gevallen. Standaardscores maken het gemakkelijk om de relatieve positie van een individueel geval te plaatsen. Iemand met een z-score van – 0,58 bevindt zich onder het gemiddelde, maar is geenszins te betitelen als een extreem geval.

Absolute scores zijn gemakkelijk om te zetten in z-scores. Bereken het verschil tussen de score en het gemiddelde en deel dit vervolgens door de standaarddeviatie. Hieronder zien we de formule, toegepast op een tweetal gevallen.

Formule: $z = \frac{x-\mu}{\sigma}$ Voorbeeld $\frac{132-100}{15} = 2{,}13$ Voorbeeld $\frac{98-100}{15} = -0{,}13$

Iemand met een IQ van 132 heeft een z-score van 2,13, iemand met een IQ van 98 heeft een z-score van – 0,13. Via een tabel, afgebeeld in bijlage 1, kunnen we de relatieve positie van deze personen bepalen. We gaan als volgt te werk. De z-score is – 0,13. Op het snijvlak van de rij 0,1 en de kolom 0,03 vinden we de waarde 0.0517. Dat betekent dat 5,17% van de mensen een score heeft tussen het gemiddelde (z=0) en onze z-waarde van – 0,13. Figuur 3 illustreert dat dit een smalle reep is van het midden naar links (bij een z-waarde van + 0,13 zou dit een reep van het midden naar rechts

zijn). We kunnen nu zeggen dat 5,17% + 50% van de bevolking (het smalle reepje plus de gehele rechterkant) een hoger IQ heeft en 44,83% (het gehele gebied links van de smalle reep) een lager IQ. Het is van belang te beseffen dat de tabel het gebied weergeeft vanaf het midden tot de opgegeven z-waarde. De tabel benadert de waarde 0.500 en geeft dus maar een helft van de normaalverdeling weer.



Figuur 3: het gebied corresponderend met 0.0517.

Het gebied dat bij een z-score van 2,13 hoort, vinden we op het snijvlak van de rij 2,1 en de kolom 0,03. De uitkomst is 0,4834. Dat betekent dat als we 10.000 personen op een rijtje zetten, iemand met een IQ van 132 een positie inneemt van 5.000 + 4.834 = 9.834. Er zijn 166 mensen met een nog hoger IQ, en 9.833 mensen met een lager IQ.

Om tot deze uitspraken te komen, hebben we in feite het volgende traject afgelegd:

1. De absolute score is 98 (of 132)
2. Deze is omgezet naar een relatieve score of z-score van – 0,13 (of 2,13)
3. Via de tabel is de z-score omgezet in een gebied van 0.0517 (of 0.4834).
4. Met de wetenschap dat de tabel maar een helft van de normaalverdeling weergeeft, concluderen we dat 55,17% (of 1,66%) een hogere score heeft.

We zijn nu beter in staat om te begrijpen waarom de waarde 1,96 zo'n belangrijke rol speelt. Kijken we nog eens naar de tabel. Op het snijvlak van de rij 1,9 en de kolom 0,06, dus bij een z-score van 1,96, vinden we een gebied van .475. Dit is het gebied van het midden naar een van de uiteinden. We kunnen dus zeggen dat 47,5% van de gevallen een z-score heeft tussen 0 en + 1,96, en ook dat 47,5% van de gevallen een z-score heeft tussen 0 en – 1,96. Dus dat 95% van de gevallen een z-score heeft tussen – 1,96 en + 1,96.

In de inductieve statistiek trekken we dit door naar steekproeven. Het getal 1,96 is op de achtergrond altijd aanwezig als we uitspraken doen op 95%. We doen ook veel uitspraken op 99% niveau. Weer afgaande op de tabel zien we dat we dan te maken hebben met een z-score van 2,58. We vinden daar een gebied van .4951. Als we dat met twee vermenigvuldigen komen we uit op rond 99%.

Op het internet zijn ook diverse tools te vinden die z-scores omzetten in gebiedsscores. Tik een z-score in, het bijbehorende gebied komt er uitrollen. Een voorbeeld is https://www.easycalculation.com/statistics/p-value-for-z-score.php

Samenvattend de belangrijkste bevindingen:

1. Een z-score geeft aan hoe extreem een waarneming is. Een z-score dicht bij nul is heel gewoon, een z-score kleiner dan – 2 of groter dan +2 is al vrij extreem.
2. In de inductieve statistiek doen we gewoonlijk uitspraken met een betrouwbaarheid van 95% of 99%. Deze zijn gebaseerd op de eigenschappen van een normaalverdeling:
   - 95% van de gevallen is hooguit 1,96 maal de standaardafwijking verwijderd van gemiddelde. Dus 95% van de mensen heeft een IQ tussen 100 – (1,96 x 15) en 100 + (1,96 x 15), ofwel tussen 70,6 en 129,4.
   - 99% van de gevallen is hooguit 2,58 maal de standaardafwijking verwijderd van gemiddelde. Dus 99% van de mensen heeft een IQ tussen 100 – (2,58 x 15) en 100 + (2,58 x 15), ofwel tussen 61,3 en 138,7.

2.3. Steekproevenverdeling en standaardfout

Intelligentietests zijn met opzet zo geconstrueerd dat de uitslagen een zuivere normaalverdeling vormen. In de praktijk echter zullen we zelden of nooit zoiets aantreffen. Maar in theorie wel. We nemen een steekproef van 400 wetenschappers en vinden dat zij gemiddeld 16,5 tijdschriften lezen voor hun werk. Stel nu eens dat we een tweede steekproef van 400 personen uit dezelfde populatie zouden nemen. Mogelijk komen we dan uit op een gemiddelde van 15,3 tijdschriften. En een derde, vierde etcetera. De centrale gedachte in de inductieve statistiek is dat we een oneindig aantal steekproeven uit een populatie nemen, telkens van dezelfde omvang. Deze leveren allemaal een gemiddelde op: de eerste 16,5, de tweede 15,3 etcetera. Als we al die gemiddelden in een verdeling zouden uitzetten, zouden we een perfecte normaalverdeling zien. Dezelfde redenering geldt ook voor percentages.

In dit bijzondere geval hanteren we ook een bijzondere naam voor de normaalverdeling. We spreken van een steekproevenverdeling (sample distribution). De steekproevenverdeling is niet gebaseerd op de scores van individuele proefpersonen, maar op de eindresultaten in de vorm van gemiddelden en percentages. Daarom maakt het ook niet uit of de individuele scores ongeveer normaal verdeeld zijn of niet. Er is wel een andere voorwaarde. De steekproefomvang moet minimaal 30 zijn. Een oneindig aantal gemiddelden gebaseerd op een oneindig aantal steekproeven telkens met een omvang van 16 levert geen perfecte normaalverdeling op.

De gehele inductieve statistiek is gebaseerd op het concept steekproevenverdeling. Bij elke steekproevenverdeling hoort een gemiddelde (voortbordurend op het voorbeeld gebaseerd op de waarden 16,5, 15,3 en verder tot in het oneindige) en dus ook een standaarddeviatie (te berekenen volgens de eerder beschreven richtlijnen). De standaarddeviatie bij een steekproevenverdeling heeft ook een aparte naam: standaardfout of standard error (SE). Uiteraard zullen we die nooit kennen, maar het is wel mogelijk een schatting te maken.

Ten slotte nog een belangrijk axioma: men gaat ervan uit dat het gemiddelde van de steekproevenverdeling gelijk is aan het gemiddelde van de echte, volledige populatie. Stel dat de steekproevenverdeling een overall gemiddelde oplevert van 15,8 (al zullen we dat in werkelijkheid

natuurlijk nooit kunnen vaststellen). Dit zou ook het gemiddelde van de populatie moeten zijn.

Samenvattend de belangrijkste bevindingen

> Ook al nemen de daadwerkelijke resultaten van een onderzoek zelden of nooit de vorm aan van een normaalverdeling, toch zijn de eigenschappen van de normaalverdeling toepasbaar in de inductieve statistiek. Een voorwaarde is dat de steekproefomvang minimaal 30 bedraagt.

3. Schatten van gemiddelden

We doen een onderzoek met een steekproef van 400 personen (n=400). We verzamelen gegevens over diverse variabelen. Een daarvan levert een gemiddelde op van 40 (m=40). Als we een gemiddelde kunnen uitrekenen, kunnen we ook de spreiding uitrekenen in de vorm van een standaarddeviatie. Deze bedraag 18 (s=18).
We kunnen niet zeggen dat het gemiddelde van de populatie ook 40 bedraagt. Een andere steekproef van 400 personen zou waarschijnlijk een ander gemiddelde opleveren.
We kunnen wel zeggen: we zijn er voor 95% (of 99%) zeker van dat het gemiddelde ligt tussen 40 plus of min een bepaalde waarde. De uitspraak bevat een *betrouwbaarheidsniveau* (95%, 99%) en een *betrouwbaarheidsinterval* (foutmarge).

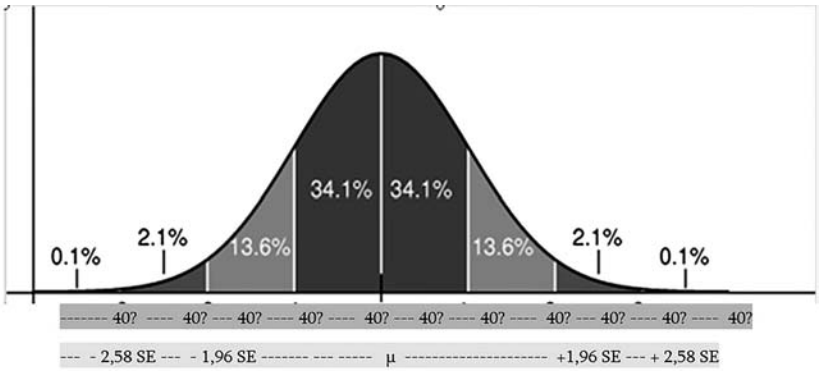
We maken nu gebruik van het principe van de steekproevenverdeling. Stel dat we een oneindig aantal steekproeven zouden nemen. Als we de gemiddelden uitzetten in een verdeling, zien we een normaalverdeling. Ons gemiddelde neemt daarin een bepaalde positie in. Het kan zijn dat ons gemiddelde dicht bij het midden zit, ergens aan de linkerkant of ver aan de rechterkant. We zullen het nooit weten. Figuur 4 illustreert dat nog eens. Wat we wel weten is dat 95% van de steekproeven een gemiddelde oplevert dat niet al te ver van het midden is verwijderd. ‘Niet al te ver’ kunnen we ook exact omschrijven: hooguit 1,96 keer de standaarddeviatie. Met standaarddeviatie bedoelen we niet de standaard-

deviatie van onze eigen gegevens, maar van de steekproevenverdeling. Zoals eerder aangegeven, heeft deze bijzondere standaarddeviatie een bijzondere naam: standaardfout (standard error, SE). Deze kennen we natuurlijk niet, maar we kunnen toch een schatting maken op basis van gegevens die we wel kennen, namelijk de standaarddeviatie van ons eigen onderzoek, en de steekproefomvang:

$$\text{Formule: } SE = \frac{s}{\sqrt{n}}. \qquad \text{Voorbeeld: } SE = \frac{18}{\sqrt{400}} = 0{,}9$$

We vervolgen onze redenering:

1) 95% van de steekproeven heeft een gemiddelde dat hooguit 1,96 x 0,9 = 1,76 lager of hoger is dan het gemiddelde van de steekproevenverdeling.
2) We zijn er dus voor 95% zeker van dat *ons* gemiddelde hooguit 1,76 afligt van het gemiddelde van de steekproevenverdeling.
3) Het gemiddelde van de steekproevenverdeling komt overeen met het gemiddelde van de populatie. We kunnen nu dus zeggen dat ons gemiddelde hooguit 1,76 afligt van het 'echte' gemiddelde, het gemiddelde van de *populatie*.
4) We kunnen dit ook omdraaien: het gemiddelde van de populatie ligt hooguit 1,76 af van ons gemiddelde. Als Den Haag twintig kilometer afligt van Leiden, ligt Leiden ook twintig kilometer af van Den Haag.
5) Ons gemiddelde is 40, het gemiddelde van de populatie ligt daar hooguit 1,76 vanaf en bevindt zich dus ergens tussen 38,24 en 41,76. Althans, daar zijn we voor 95% zeker van.
6) We vergeten het in onze feestvreugde gauw, maar er is 5% kans dat we er naast zitten.
7) We kunnen strengere eisen stellen. We zijn er voor 99% zeker van dat het gemiddelde van de populatie ligt tussen 40 ± (2,58 x 0,9), dus tussen 37,68 en 43,32.



Figuur 4. Positie van steekproefresultaat op steekproevenverdeling.

Het kader hieronder geeft de stappen in kort bestek weer.

Benodigde ingrediënten: omvang steekproef, gemiddelde steekproef, standaarddeviatie steekproef. Stel: n = 400, m = 40, s = 18

1. Bereken de geschatte standaardfout van de steekproevenverdeling:

$$SE = \frac{s}{\sqrt{n}} = \frac{18}{\sqrt{400}} = 0{,}9$$

2. We willen een uitspraak doen met een betrouwbaarheid van 95%.
   1. Vermenigvuldig de standaardfout met 1,96. De marge wordt dus 1,96 x 0,9 = 1,76.
   2. De kans is dus 95% dat het populatiegemiddelde μ ligt tussen 40 ± 1,76, ofwel tussen 38,24 en 41,76.
3. Om uitspraken te doen die gelden met een betrouwbaarheid van 99% vervangen we 1,96 in de formule door 2,58. De formule in zijn algemeenheid luidt:

$$\mu = m \pm z. \frac{s}{\sqrt{n}}$$

4. Merk op dat het interval kleiner wordt indien s kleiner is en / of n groter is. Dit is logisch.

Het is altijd goed om naar de ingrediënten van een formule te kijken. We leren hieruit dat de foutmarge afhangt van het betrouwbaarheidsniveau (95 of 99%), de standaarddeviatie en de steekproefomvang. De standaarddeviatie hebben we niet in eigen hand. De steekproefomvang wel: een grotere steekproef wordt beloond met een kleinere foutmarge. In paragraaf 7 maken we gebruik van deze formule om de benodigde steekproefomvang te bepalen. We redeneren dan de andere kant op. Niet: hoe groot is de foutmarge bij een bepaalde steekproefomvang, maar: hoe groot moet de steekproef zijn als we een van tevoren bepaalde foutmarge willen halen.

Er valt nog iets op aan de formule: de omvang van de populatie (N) maakt er geen deel van uit. Maakt het dan, voor het bepalen van de foutmarge, niet uit of we een steekproef trekken uit een populatie van 4.000, 10.000, 50.000 of 500.000 eenheden? Maar zeer ten dele. Bij kleinere populaties heeft het nog wel enig effect. Eigenlijk hebben we dat veronachtzaamd door de formule in te korten. Voluit luidt deze:

$$\mu = m \pm z.\sqrt{\frac{s^2}{n} - \frac{s^2}{N}}$$

Toegepast op ons voorbeeld met m=40, n=400 en s=18, geeft dat de volgende resultaten. Het geringe effect van de populatieomvang komt hier duidelijk naar voren.

Omvang populatie	**Foutmarge**
4.000	1,96 x 0,85 = 1,67
10.000	1,96 x 0,88 = 1,73
50.000	1,96 x 0,90 = 1,76
500.000	1,96 x 0,90 = 1,76

4. Schatten van percentages

We kunnen hier dezelfde redenering volgen, met één essentieel verschil: een percentage is geen gemiddelde, er hoort geen standaarddeviatie bij.

Er is daarom een andere formule om de standaardfout van de steekproevenverdeling te schatten:

$$SE = \sqrt{\frac{pq}{n}}$$

Daarbij zijn p en q de in ons onderzoek gevonden percentages. Wanneer wetenschappers 77,6% van de tijdschriften die ze lezen betrekken via hun eigen universiteitsbibliotheek, geldt p=77,6 en q=22,4. Om het percentage van de populatie (π) te bepalen, moeten we verder rekening houden met de steekproefomvang n. Het kader geeft een voorbeeld.

1. Een steekproef van 400 cases geeft als resultaat 60% ja en 40% nee. Wat kunnen we nu zeggen over het percentage van de populatie?
2. We zijn er voor 95% zeker van dat ook voor de populatie een percentage van 60 geldt, met een afwijking van 1,96 maal $\sqrt{\frac{pq}{n}}$, in dit geval 1,96. $\sqrt{\frac{60x40}{400}} = 4,80$
3. Met 95% zekerheid kunnen we dus zeggen dat het percentage van de populatie ligt tussen 55,2 en 64,8%.
4. Bij een betrouwbaarheidsniveau van 99% vervangen we 1,96 door 2,58. We zijn er voor 99% zeker van dat het percentage van de populatie ligt tussen 53,7 en 66,3%. In zijn algemeenheid luidt de formule:

$$\pi = p \pm z. \sqrt{\frac{pq}{n}}$$

5. Voluit luidt de formule $\Pi = p \pm z.\sqrt{\frac{pq}{n-1} - \frac{pq}{N-1}}$ Bij kleine populaties geeft dit een iets kleinere foutmarge.

5. Schatten en toetsen van verschillen

5.1. Schatten van het betrouwbaarheidsinterval

In paragraaf 3 is een schatting gemaakt van het gemiddelde dat geldt voor de populatie, gebaseerd op het gemiddelde dat de steekproef opleverde.

In deze paragraaf gaat de aandacht uit naar het verschil tussen de gemiddelden van twee groepen. Een mooi voorbeeld is beschreven door de University of Strathclyde.[1] Een steekproef wees uit dat mannen gemiddeld 48,73 punten halen op een toets, vrouwen gemiddeld 47,55. Een verschil dus van 1,18. Om dit door te trekken naar de populatie moeten we weer een uitspraak doen als: we zijn er voor 95% zeker van dat het verschil tussen de twee groepen in de populatie ook 1,18 is, plus of min een bepaalde marge.

Daartoe stellen we ons weer een steekproevenverdeling voor, nu van verschillen in gemiddelden. De uiteinden van het 95% gebied kunnen we weer berekenen door de standaardfout (standaarddeviatie van de theoretische steekproevenverdeling) te vermenigvuldigen met 1,96. Die standaardfout kennen we uiteraard niet, maar kunnen we ook nu schatten op basis van gegevens die wel bekend zijn, met de formule:

$$\mathrm{SE} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Toepassing van de formule levert een SE op van 0,96. De marge, zowel naar boven als naar beneden, is dan 1,96 x 0,96 = 1,88. We concluderen, met een betrouwbaarheid van 95%, dat het verschil tussen de twee groepen in de gehele populatie ligt tussen – 0,70 en + 3,07. Het verschil loopt van een negatieve waarde naar een positieve waarde. Dit betekent dat een andere steekproef zou kunnen uitwijzen dat vrouwen beter scoren dan mannen. Met SPSS of een ander statistisch pakket kunnen we ons veel werk besparen. De toets die we nodig hebben is de T-toets.

Figuur 5 toont het resultaat. In het eerste gedeelte zien we de gegevens van de steekproef: aantal personen, gemiddelde scores, standaarddeviaties. In het tweede gedeelte zien we onder meer het verschil in scores (+ 1,18) en het 95% betrouwbaarheidsinterval. Er zijn zelfs twee rijen met een licht afwijkend interval. Dat we in dit geval moeten afgaan op de onderste rij, wordt straks duidelijk (paragraaf 5.2, punt 3).

Group Statistics

	sex of participant	N	Mean	Std. Deviation	Std. Error Mean
pipscience	male	221	48.7330	11.10676	.74712
	female	218	47.5505	8.87755	.60126

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
pipscience	Equal variances assumed	10.180	.002	1.231	437	.219	1.18257	.96046	-.70513	3.07027
	Equal variances not assumed			1.233	419.047	.218	1.18257	.95901	-.70251	3.06765

Figuur 5. Voorbeeld resultaat t-toets.

5.2. Toetsen van de nulhypothese en significantieniveau

Onderzoekers zijn echter eerder geneigd om af te gaan op het significantieniveau. Dat bedraagt hier .218. Het begrip significantie komt veelvuldig voor in de onderzoeksliteratuur, het is daarom goed daar iets langer bij stil te blijven staan. De essentie is samengevat in zes punten. De punten 3, 4 en 5 gaan in op details van de berekening en kunnen desgewenst worden overgeslagen.

1. Nulhypothese en alternatieve hypothese

Wetenschappers zijn behoudend. Het is moeilijk te bewijzen dat een bewering waar is, je kunt wel aantonen dat een bewering niet waar is. En wetenschappers hechten veel belang aan verschillen tussen groepen of verbanden tussen variabelen. Deze helpen om de werkelijkheid beter te begrijpen. Tezamen leiden deze twee uitspraken ertoe dat wetenschappers een nulhypothese opstellen, die luidt: er is geen verschil, of geen verband, tussen ... Bijvoorbeeld: er is geen verschil tussen de scores van mannen en vrouwen. Het omgekeerde, er is wel een verschil of verband, noemen we de alternatieve hypothese of ook wel researchhypothese.

De nulhypothese betekent:

- Niet dat het verschil tussen de twee groepen in de steekproef nul is. Dit resultaat zou wel zeer toevallig zijn en kunnen we bovendien niet zonder meer doortrekken naar de populatie.
- Niet dat het verschil tussen de twee groepen in de populatie nul is. Ook dit zou wel zeer toevallig zijn. Bovendien zullen we het exacte verschil nooit weten, we kunnen hooguit een uitspraak doen in termen van betrouwbaarheidsniveau en betrouwbaarheidsinterval.
- Wel dat het verschil tussen de twee groepen in de steekproef zo dicht bij nul ligt dat het goed voorstelbaar is dat een volgende steekproef een verschil van nul zou kunnen opleveren.

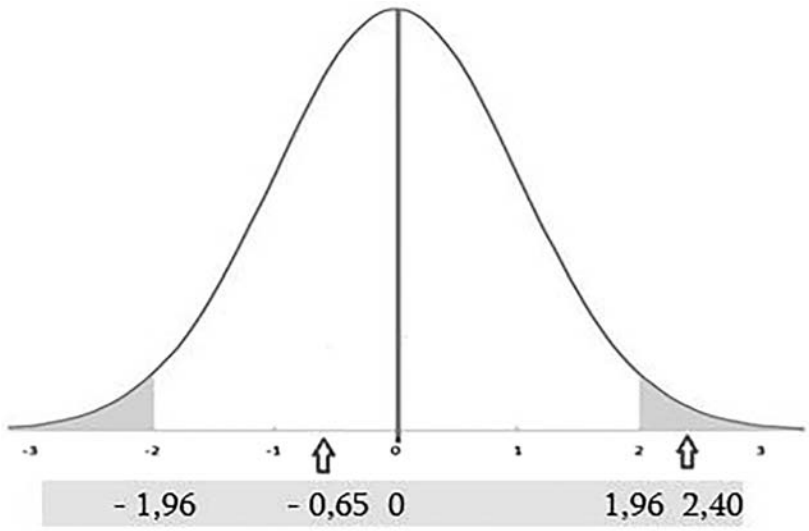
Het rekenwerk is erop gericht na te gaan of we de nulhypothese kunnen verwerpen. Daarvan is sprake als voor hooguit vijf op de honderd steekproeven geldt dat die een verschil van nul zouden kunnen opleveren.

2. Significantie

Het begrip significantie berust in wezen op hetzelfde principe als het begrip betrouwbaarheid, maar de aandacht gaat nu uit naar iets anders. De redenering rond het betrouwbaarheidsinterval is als volgt. In onze steekproef vinden we een verschil van 1,18. Een tweede steekproef zal hoogstwaarschijnlijk een ander verschil opleveren. Hoe groot dat verschil zal zijn is uiteraard onbekend, maar we weten wel met een zekerheid van 95% dat dit ligt tussen – 0,70 en + 3,07.

Bij significantie draait het om de vraag of we de nulhypothese ('er is geen verschil tussen de twee groepen in de populatie') kunnen verwerpen. De redenering verloopt als volgt:

1. Het uitgangspunt is dat het verschil nul is. Teken een normaalverdeling met nul in het midden. De twee staarten zijn duidelijk aangegeven.



Figuur 6. Toetsen van de nulhypothese.

2. Het daadwerkelijke onderzoek levert een verschil op dat zeer waarschijnlijk anders is dan nul. Waar het om gaat, is of het gevonden verschil in de buurt ligt van het veronderstelde verschil van nul. Met 'in de buurt' bedoelen we in het centrale 95% gebied rondom nul. Om dit te bepalen zetten we het absolute verschil om in een z-score.
3. Stel z = – 0,65: groep 1 scoort iets lager dan groep 2. Deze waarde valt in het centrale 95% gebied. De nulhypothese wordt niet verworpen. Het gevonden verschil in de steekproef is best verenigbaar met een verondersteld verschil van nul in de populatie.
4. Stel z = + 2,40: groep 1 scoort beduidend hoger dan groep 2. Deze waarde valt in de rechterstaart en overschrijdt daarmee de grens. Het

verschil in de steekproef is zo groot dat we de veronderstelling dat het verschil in de populatie nul bedraagt moeten verwerpen.

Er is pas sprake van significantie als het gevonden verschil zich bevindt in de linker of rechter 2,5% staart. Anders gezegd: er is sprake van significantie als $p < .05$. Letterlijk staat hier dat de kans ('probability') dat de nulhypothese waar is, kleiner is dan 5%. De p-waarde is dus eigenlijk de kans dat we een vergissing maken en de nulhypothese ten onrechte verwerpen. Hoe kleiner p, des te kleiner de kans op een dergelijke vergissing. Maar een verkeerde conclusie is nooit uitgesloten. Het kan zijn dat de steekproef toevallig bestaat uit objecten of personen die niet representatief zijn. We doen alles volgens de regelen der kunst, maar het toeval kan ons parten spelen.
De berekening van het significantieniveau verloopt via de volgende stappen (zie ook figuur 5):

1. Het verschil tussen de gemiddelde scores van mannen en vrouwen is 1,18.
2. De standaardfout bedraagt 0,96.
3. Vertaal het absolute verschil in een standaardscore (z-score).

$$z = \frac{(m_1 - m_2) - 0}{SE} \qquad \text{Vb } \frac{1{,}18 - 0}{0{,}96} = 1{,}23$$

4. Vertaal de z-score in een gebiedsscore met behulp van de tabel in de bijlage. Het bijbehorende gebied is .3907. Er resteert dus, zowel naar links als naar rechts, een gebied van .1093, in totaal .2186. Dat we hier beide uiteinden meerekenen en niet, zoals misschien verwacht, alleen het rechter uiteinde, komt ter sprake in punt 4. Dat we eigenlijk buiten ons boekje gaan en niet te maken hebben met z-scores maar met t-scores, wordt uitgelegd in punt 5.
5. De conclusie is p=.218. Deze waarde is hoger dan .05, het verschil is dus niet significant. De kans dat we de nulhypothese ten onrechte verwerpen is veel te groot. Eigenlijk zagen we dit al aan het betrouwbaarheidsinterval: dit loopt van een negatieve naar een positieve waarde en omvat dus de waarde nul.

Het is gebruikelijk om in de rapportage de exacte p-waarde te vermelden. Vaak voegt men daaraan toe welke grens overschreden is:
$p < .05$ Bijvoorbeeld p=.036. Ook wel aangegeven met *

$p < .01$. Bijvoorbeeld $p=.006$. Ook wel aangegeven met **
$p < .001$. Waarden als $p=.000482$ worden afgekort tot $p=.000$. Ook wel aangegeven met ***

Als we de nulhypothese kunnen verwerpen, geeft dat steun aan de alternatieve hypothese. Of we blij moeten zijn met deze uitkomst, hangt van ons doel af. Meestal is dat wel het geval. Bij een experiment hopen we bijvoorbeeld dat er een significant verschil is tussen de scores van een experimentele en controlegroep.

We keren even terug naar de paragrafen 3 en 4 over schattingen. Daarin zijn betrouwbaarheidsintervallen berekend waar we 95% (of 99%) zeker van zijn. Er is dus een kans van 5% (of 1%) dat we een verkeerde uitspraak doen. Dat geldt ook voor het toetsen van de nulhypothese. Er zijn twee fouten mogelijk:

- Type 1 fout. De nulhypothese wordt ten onrechte verworpen. Onze steekproef levert een verschil op dat in een van de staarten valt. Dit is echter gebaseerd op toeval, op een atypische samenstelling van de steekproef. Een paar keer op de honderd kan dit voorkomen. Wij hadden net die pech.
- Type 2 fout. De nulhypothese wordt ten onrechte aangenomen. Onze steekproef levert een verschil op dat valt in het centrale 95% gebied rond nul. Bij een tweede, derde, vierde ... steekproef zou het verschil echter in een van de staarten vallen. Ook nu weer is dit het gevolg van een atypische samenstelling van de steekproef.

Uiteraard weten we voor onze steekproef niet of er sprake is van een dergelijke fout. Maar het is goed te beseffen dat we te maken hebben met kansberekening en geen absoluut zekere uitspraken kunnen doen.

3. Levene's toets

Figuur 5 bevat twee rijen met resultaten. De bovenste is geldig als de standaarddeviaties van de scores van mannen en vrouwen niet te veel van elkaar verschillen. De onderste is altijd geldig, ook wanneer er wel een groot verschil is tussen die standaarddeviaties. Voor deze twee situaties geldt een afzonderlijke formule om de standaardfout te berekenen. Levene's toets test de nulhypothese: 'er is geen verschil tussen de standaarddeviaties'. De uitslag is $p=.002$. Deze waarde is significant. Hier

betekent dat, dat er wel een groot verschil is tussen beide standaarddeviaties. In dat geval is de onderste resultatenrij van toepassing. De verschillen zijn echter marginaal.

4. Tweezijdige of eenzijdige toets
Bij het significantieniveau zien we de aanduiding '2-tailed'. Dit wijst op een tweezijdige toets. De nulhypothese luidt: er is geen verschil tussen de scores van mannen en vrouwen. De alternatieve hypothese luidt: er is wel een verschil. Hierbij hebben we op voorhand niets gezegd over de richting van het verschil. Daarvan zou sprake zijn bij een nulhypothese als: mannen scoren niet hoger dan vrouwen (of omgekeerd). De alternatieve hypothese luidt dan: mannen scoren wel hoger dan vrouwen (of omgekeerd). In het eerste geval verrichten we een tweezijdige toets, in het tweede een eenzijdige toets.

Over het algemeen zijn onderzoekers vrij terughoudend: als er niet een heel goede reden is om een verschil of verband in een bepaalde richting te veronderstellen, doen we dat niet. In sommige gevallen is er misschien wel reden om een verschil of verband in een bepaalde richting te verwachten. Dit doet zich vaak voor bij experimenten: we verwachten dat proefpersonen die een cursus hebben gevolgd over meer vaardigheden beschikken dan de proefpersonen die niet een cursus hebben gevolgd. De nulhypothese luidt dan: 'de experimentele groep behaalt geen hogere score op een vaardighedentest dan de controlegroep'. De alternatieve hypothese is: 'de experimentele groep behaalt wel een hogere score dan de controlegroep'.

Met een eenzijdige toets kun je eerder de nulhypothese verwerpen dan met een tweezijdige toets. We kunnen de nulhypothese verwerpen als het resultaat uit ons onderzoek niet valt in de linker- of rechterstaart van 2,5%, maar in een enkele staart van 5%. Vertaald in z-scores: bij een tweezijdige toets moeten we de z-waarde van 1,96 overschrijden, bij een eenzijdige toets is het al voldoende om de z-score van 1,65 te overschrijden, om de nulhypothese op .05 niveau te verwerpen.

5. z-scores of t-scores?
Voor het gemak zijn we tot nu toe uitgegaan van z-scores. Eigenlijk is dat onjuist en moeten we spreken van t-scores. De toets heet niet voor niets t-toets (voluit Student's t-toets; Student was het pseudoniem van

William Gosset, een chemicus in dienst van de Guinness brouwerij in Dublin). In paragraaf 2 is de z-score geïntroduceerd: verminder een individuele score met de gemiddelde score en deel dit door de standaarddeviatie. Bijvoorbeeld (118 – 100) / 15 = 1,20. Ook nu hebben we een score (1,18) verminderd met een verondersteld gemiddelde (0). We deelden dit echter niet door de standaarddeviatie, maar door de geschatte standaardfout (SE). Daarom is er nu geen sprake van z-scores, maar van t-scores.

Bij grote steekproeven is het effect van t-scores nagenoeg hetzelfde als dat van z-scores. We hebben gezien dat 95% van de gevallen zich bevindt tussen de z-scores van – 1,96 en + 1,96. Bij een steekproef van 1000 eenheden geldt eveneens dat 95% van de gevallen zich bevindt tussen de t-scores van – 1,96 en + 1,96. Bij kleine steekproeven is er wel een verschil. Bij een steekproef van 60 cases ligt het 95% gebied nog steeds tussen de z-scores van – 1,96 en + 1,96, maar voor t-scores gelden nu ruimere grenzen: tussen – 2,00 en + 2,00. Dat maakt het iets lastiger om de nulhypothese te verwerpen.

Dat in figuur 5 rekening is gehouden met de steekproefomvang blijkt uit de vermelding df (degrees of freedom, aantal vrijheidsgraden). In de bovenste rij is de df gemakkelijk te herleiden: n – 2, dus (221 + 218) – 2 = 437. Dit is een redelijk grote steekproef, we zouden met z-waarden dus nagenoeg hetzelfde resultaat bereiken als met t-waarden. De t-toets houdt uiteraard uitsluitend rekening met t-waarden.

6. Interpretatie van significantie

Hoewel significantie een veel gebruikt begrip is, valt er wel het een en ander op af te dingen:

1. Significantie is minder informatief dan het betrouwbaarheidsinterval. Uit p=.036 valt niet op te maken tussen welke grenzen het verschil zich bevindt.
2. Een verschil, of verband is pas significant bij een p-waarde van .05 of lager. Bij een p-waarde van .051 is er geen sprake van een significant verschil, bij een p-waarde van .049 wel.
3. De ingrediënten voor de berekening van de mate van significantie zijn: het daadwerkelijke verschil dat uit de steekproef rolt, de steekproefomvang van beide groepen en de standaarddeviaties van beide groepen. De steekproefomvang hebben we in eigen hand. Als we deze maar groot genoeg maken, zullen we bijna altijd een significant resultaat behalen.

4. Hierop voortbordurend: als de steekproefomvang maar groot genoeg is, is het verschil al gauw significant, ook al is de omvang van het verschil maar klein. Een statistisch significant verschil is daarom, zeker bij een grote steekproef, niet noodzakelijk een groot of praktisch relevant verschil. Een verschil kan zeer klein zijn en toch significant.
5. Er is altijd wel een verschil tussen twee groepen of een verband tussen twee variabelen. Het zou zeer toevallig zijn als de gemiddelde scores van mannen en vrouwen, tot achter de komma, gelijk zijn. En, zoals zojuist vermeld, bij een grote steekproefomvang is er bijna vanzelf sprake van significantie. Krijgen we de handen dan nog wel op elkaar voor de conclusie: 'Er is een significant verschil tussen de gemiddelde scores van mannen en vrouwen?' Onderstaande tabel zet een aantal uitspraken op een rijtje, gebaseerd op een voorbeeld met weer geheel fictieve waarden. Deze zijn naar onderen toe steeds minder informatief.

Uitspraak	**Toepasbaarheid**
Vrouwen lenen op jaarbasis gemiddeld 12,8 boeken meer dan mannen uit de Openbare Bibliotheek.	Onderzoek onder gehele populatie.
Zeer waarschijnlijk (95%, 99%) lenen vrouwen tussen 10,0 en 15,6 boeken meer dan mannen.	Onderzoek onder steekproef. Resultaat in de vorm van betrouwbaarheidsintervallen.
Vrouwen lenen significant meer boeken dan mannen (p=.024).	Onderzoek onder steekproef. Resultaat in de vorm van significantieniveau (p-waarde). Eenzijdige toets.
Er is een significant verschil in het aantal boekuitleningen tussen mannen en vrouwen (p=.048).	Onderzoek onder steekproef. Resultaat in de vorm van significantieniveau (p-waarde). Tweezijdige toets.

Waarom dan toch de voorkeur voor significantie? Dit heeft te maken met het eenduidige karakter van p-waarden. Wetenschappers kunnen een vermelding als p=.008 direct plaatsen. Waarschijnlijk ook verwachten

auteurs dat uitgevers eerder geneigd zijn een artikel te publiceren wanneer er sprake is van significante verbanden of verschillen.

5.3. Overzicht van toetsen

Om de achtergronden te belichten, zijn we uitgegaan van het meest simpele geval: twee groepen, die we vergelijken op een variabele op rationiveau. De t-toets is hiervoor van toepassing. Voor andere situaties gelden andere toetsen, die hier alleen genoemd worden.

Toets	**Toepasbaarheid**
T-test	Vergelijking tussen gemiddelde scores van twee groepen (bijvoorbeeld mannen en vrouwen) op een variabele op rationiveau (bijvoorbeeld aantal boekleningen)
Paired T-test	Vergelijking van telkens twee scores van een zelfde proefpersoon of case. Vaak gebruikt bij voor- en nameting bij experimenten en longitudinaal onderzoek. Bijvoorbeeld vergelijking van zuurgraad gemeten bij 400 boeken en zuurgraad van diezelfde boeken twintig jaar later.
Mann-Whitney U-test = Wilcoxon rank-sum test = Mann-Whitney-Wilcoxon test	Variant van de T-test wanneer de variabele niet een ratio, maar een ordinaal karakter heeft (bijvoorbeeld schaalwaarden). Ook meer geschikt dan de T-toets bij hele kleine steekproeven.
Wilcoxon signed-ranks test / sign test (tekentoets)	Variant van de paired T-toets, toegepast wanneer de scores niet betrekking hebben op een variabele op rationiveau, maar ordinaal niveau.

F-test = variantieanalyse = ANOVA (Analysis of Variance)	Vergelijking tussen gemiddelde scores van drie of meer groepen (bijvoorbeeld studenten op alfa, bèta, gammagebied) op een variabele op rationiveau (bijvoorbeeld aantal artikeldownloads)
Kruskal-Wallis test	Variant van de F-toets, toegepast wanneer de scores niet betrekking hebben op een variabele op rationiveau, maar ordinaal niveau.
Chi kwadraat (χ^2)	Verschil tussen percentages van twee of meer groepen. Zie verder paragraaf 6.

6. Toetsen van verbanden

Er zijn verschillende methoden om de significantie van het verband tussen twee variabelen te bepalen, afhankelijk van het meetniveau. Deze paragraaf gaat uitgebreid in op de Pearson correlatiecoëfficiënt en de chi kwadraat en besluit met een kort overzicht van toetsen.

6.1. Pearson correlatiecoëfficiënt

De Pearson correlatiecoëfficiënt (r) is van toepassing als we beschikken over twee variabelen op rationiveau. Figuur 7 toont een voorbeeld van de output van SPSS. Er is een vrij sterk negatief verband tussen het netto inkomen en het aantal supermarktbezoeken (r = –.615). Dit is plausibel: voor hogere inkomens is het minder noodzakelijk op koopjes te jagen en dus vaak naar een supermarkt te gaan. De significantie bedraagt p=.000 (dit is een afkorting van zoiets als .000375), de correlatie is dus significant op .001 niveau (en zeker, zoals aangegeven, op .01 niveau). Als we 100.000 steekproeven zouden nemen, zouden we bij meer dan 99.900 de nulhypothese 'er is geen verband tussen inkomen en aantal supermarktbezoeken' kunnen verwerpen. SPSS toont de resultaten voor elk denkbaar paar: A-A, A-B, B-A, B-B. Eén daarvan (A-B dan wel B-A) is voldoende.

Correlations

		winkelbezoeken	netto inkomen
winkelbezoeken	Pearson Correlation	1	-,615**
	Sig. (2-tailed)		,000
	N	80	76
netto inkomen	Pearson Correlation	-,615**	1
	Sig. (2-tailed)	,000	
	N	76	76

**. Correlation is significant at the 0.01 level (2-tailed).

Figuur 7. Sterkte en significantie van Pearson correlatiecoëfficiënt (SPSS output).

Bij de T-toets, waarmee we de significantie van het verschil in gemiddelden kunnen berekenen, zagen we tevens nog een betrouwbaarheidsinterval. Dit ontbreekt hier geheel, al zou het ook met SPSS via een moeizame omweg te berekenen zijn. Op internet zijn echter diverse tools te vinden. Een daarvan is http://www.how2stats.net/2011/09/confidence-intervals-for-correlations.html. Figuur 8 toont het beginscherm. Vul de sterkte van het verband en de steekproefomvang in, de tool toont dan het betrouwbaarheidsinterval op 90, 95 en 99% niveau. Uit het voorbeeld blijkt dat, anders dan we tot nu toe gezien hebben, bij correlaties de marge naar beneden (0,506 – 0,350 = 0,156) niet exact hetzelfde is als de marge naar boven (0,635 – 0,506 = 0,129). Dat heeft te maken met het ordinale karakter van de correlatiecoëfficiënt: je kunt wel zeggen dat een r-waarde van 0,40 hoger is dan een r-waarde van 0,20, maar niet dat een verband van r=0,40 twee keer zo sterk is als een verband van r=0,20. Wie iets meer van de achtergronden wil begrijpen en een handmatige berekening wil uitvoeren, kan terecht bij http://davidmlane.com/hyperstat/B8544.html.

Confidence Intervals for Corre... Opslaa

B3 | *fx* | 0.506

	A	B	C
1			
2		**Enter Data**	
3	Correlation	0.51	
4	Sample Size	107	
5			
6			
7		**Results**	
8		Lower	Upper
9	**90% CI**	0.377	0.616
10			
11	**95% CI**	0.350	0.635
12			
13	**99% CI**	0.296	0.670

Figuur 8. Tool voor de berekening van het betrouwbaarheidsinterval van een correlatiecoëfficiënt.

6.2. Chi kwadraat

De (Pearson) chi kwadraat toets is van toepassing als we beschikken over twee variabelen op nominaal niveau. In het voorbeeld proberen we te onderzoeken of er een significant verband is tussen geslacht en wel/niet gebruik van een bepaalde voorziening. Figuur 9 geeft de output van SPSS weer. Het eerste gedeelte toont de resultaten van het onderzoek. Wanneer beide metingen op nominaal niveau zijn, kan het resultaat altijd worden weergegeven in een kruistabel. De verdere berekeningen zijn gebaseerd op absolute aantallen, niet op percentages. Daarom zijn in de kruistabel ook alleen absolute aantallen opgevoerd.

Het tweede gedeelte bevat het wellicht meest interessante gegeven: het significantieniveau. Het verband is, met een p-waarde van .264, niet significant. Deze waarde ligt ver boven de kritische grens van .05.

De chi kwadraat toets maakt alleen uitspraken mogelijk over de significantie, niet over de sterkte van het verband. Daartoe dienen de op chi kwadraat

gebaseerde maten Phi en Cramer's V. Het derde gedeelte van de output laat zien dat het verband zeer zwak is. Phi en Cramer's V geven een sterkte aan van .057. Dit is ver verwijderd van de maximale sterkte 1,00 of – 1,00. We kunnen dit ook handmatig berekenen: deel de chi kwadraat waarde (1,247) door de steekproefomvang en trek daaruit de wortel:

$$\text{Phi} = \sqrt{\frac{\text{chi kwadraat}}{n}} \qquad \text{Vb: } \sqrt{\frac{1.247}{380}} = 0{,}057$$

Phi gebruiken we bij 2 x 2 tabellen, Cramer's V is ook van toepassing op bijvoorbeeld een 5 x 3 tabel.

gebruik * geslacht Crosstabulation

Count

		geslacht		
		man	vrouw	Total
gebruik	ja	120	100	220
	nee	78	82	160
Total		198	182	380

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,247[a]	1	,264		
Continuity Correction[b]	1,025	1	,311		
Likelihood Ratio	1,247	1	,264		
Fisher's Exact Test				,299	,156
Linear-by-Linear Association	1,243	1	,265		
N of Valid Cases	380				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 76,63.

b. Computed only for a 2x2 table

Symmetric Measures

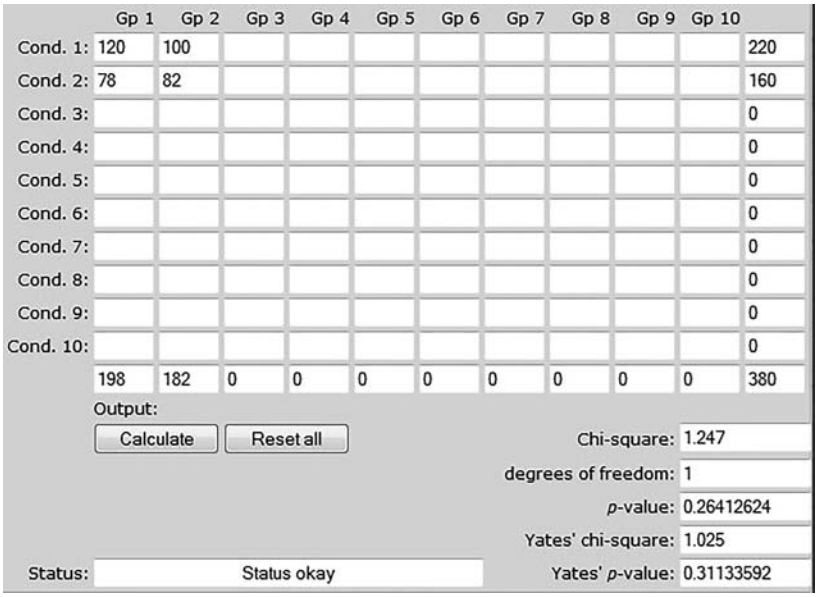
		Value	Approx. Sig.
Nominal by Nominal	Phi	,057	,264
	Cramer's V	,057	,264
N of Valid Cases		380	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

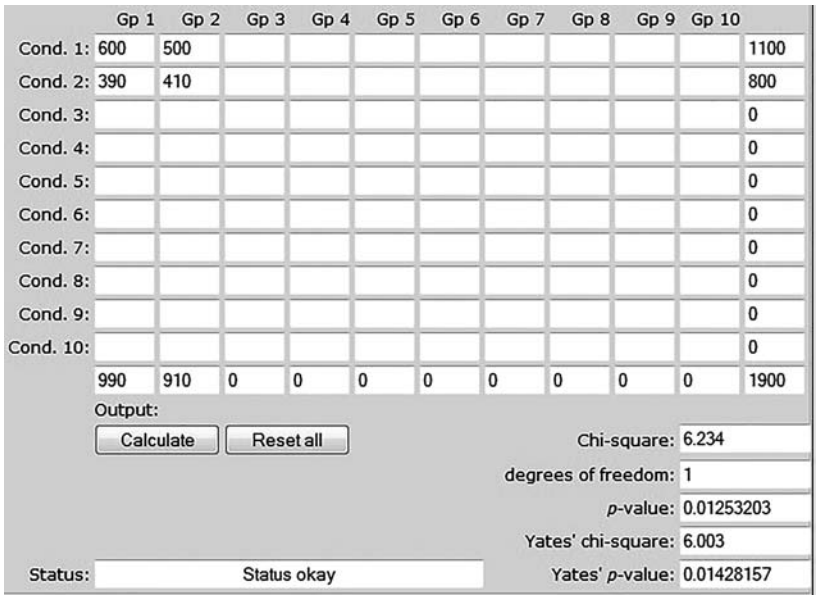
Figuur 9. Chi kwadraat toets (SPSS output).

Er zijn op internet allerlei tools om chi kwadraat te berekenen. Een daarvan is http://www.quantpsy.org/chisq/chisq.htm. Figuur 10 laat zien dat we dezelfde uitkomst krijgen als bij SPSS: p=.264. Met de chi kwadraat waarde van 1.247 kunnen we desgewenst weer phi uitrekenen.



Figuur 10. Chi kwadraat toets (gratis tool).

In paragraaf 5 is reeds vermeld dat het significantieniveau mede afhankelijk is van de steekproefomvang. Dit kunnen we mooi uittesten met de tool. We vermenigvuldigen de waarden in de vier cellen telkens met 5. Er is nog steeds sprake van een 60,6% : 54,9% verhouding en van een zwak verband (phi = .057, niet weergegeven in de figuur). Maar het verband is nu wel significant. De waarde p=.0012 ligt ver beneden de kritische grens p=.05 en benadert nagenoeg de p=.001 grens. De conclusie is dat bij een grote steekproefomvang een significant verband niet zoveel zegt.



Figuur 11. Effect van grotere steekproefomvang op significantieniveau.

6.3. Overzicht van toetsen

	Variabele 1	*Variabele 2*	*Toets*	*Meting sterkte*	*Meting significantie*
1	Ratio	Ratio	Pearson correlatie-coëfficiënt	Ja	Ja
2	Ratio	Ordinaal	* Spearman's rho * Kendall's tau	Ja	Ja
3	Ordinaal	Ordinaal	* Spearman's rho * Kendall's tau	Ja	Ja
4	Nominaal	Nominaal	Chi kwadraat	Nee	Ja
5	Nominaal	Ratio	T-toets	Nvt	Ja
6	Nominaal	Ordinaal	Mann-Whitney U test	Nvt	Ja

Er zijn zes combinaties mogelijk, afhankelijk van het meetniveau van de variabelen. Situatie 1 en 4 zijn hierboven toegelicht. Als een of beide variabele(n) een nominaal karakter heeft/hebben, komen de begrippen

verschil en verband op hetzelfde neer, zoals hieronder geïllustreerd. Daarom kiezen we in de situaties 5 en 6 voor toetsen die al eerder genoemd zijn in paragraaf 5 over het toetsen van verschillen.

Ad 4, nominaal – nominaal
- Is er een verschil tussen het percentage mannen en vrouwen dat gebruik maakt van een bepaalde voorziening?
- Is er een verband tussen geslacht en het gebruik van een bepaalde voorziening?

Ad 5, nominaal – ratio
- Is er een verschil tussen de uitleenfrequentie van mannen en vrouwen?
- Is er een verband tussen geslacht en uitleenfrequentie?

Ad 6, nominaal – ordinaal
- Is er een verschil tussen de mate van tevredenheid van mannen en vrouwen?
- Is er een verband tussen geslacht en mate van tevredenheid?

7. Bepalen van steekproefomvang

7.1. Steekproefomvang bij onderzoek gericht op bepalen van percentages

In paragraaf 4 kwam de volgende problematiek aan de orde. Een steekproef van 400 cases geeft als resultaat 60% ja en 40% nee. Wat kunnen we nu zeggen over het percentage van de populatie? Met andere woorden: hoe berekenen we de foutmarge, ofwel het betrouwbaarheidsinterval? Laten we deze foutmarge e (van 'error') noemen. De formule luidt dan:

$$e = z.\sqrt{\frac{pq}{n}} \qquad \text{Vb } e = 1{,}96.\sqrt{\frac{60x40}{400}} = 4{,}80$$

Met 95% zekerheid kunnen we dus zeggen dat het percentage van de populatie ligt tussen 55,2 en 64,8%. In deze situatie hadden we al op voorhand de steekproefomvang bepaald en stelden we achteraf de

foutmarge vast. We kunnen dit ook omdraaien. Hoe groot moet de steekproefomvang zijn als we van tevoren bepaalde eisen stellen aan de foutmarge? Met een beetje algebraïsch geknutsel kunnen we de formule herschikken tot:

$$n = z^2 . \frac{pq}{e^2}$$

Nu is de steekproefomvang n de onbekende. We moeten echter wel de andere waarden invullen.

- z stellen we op 1,96 of 2,58, al naar gelang het gewenste betrouwbaarheidsniveau (95%, 99%).
- e wordt vaak gesteld op 5(%). Een marge van 5% is, als er geen mensenlevens op het spel staan, vaak toereikend.
- p en q zijn de percentages die uit het onderzoek rollen. Hier wordt ons gevraagd die nu al te geven. Dat lijkt een onmogelijke vraag, maar uiteindelijk is altijd wel een antwoord mogelijk.
  1. Op basis van eerdere ervaringen of vergelijkbaar onderzoek kunnen we een inschatting maken.
  2. Begin alvast met de uitvoering van het onderzoek en stel na enkele tientallen cases vast hoe de percentageverdeling is. Dit is vaak goed te doen als de steekproef bestaat uit objecten, zoals titelrecords of gedigitaliseerde pagina's. Op basis van de tussentijdse uitslag kunnen we de uiteindelijke steekproefomvang bepalen. Bij een onderzoek met mensen is het vaak praktisch om alle uitnodigingen in één keer te versturen. De optie om alvast op beperkte schaal met het onderzoek te beginnen is in dit geval niet altijd goed uitvoerbaar.
  3. We stellen p en q beide op 50. Volgens de formule moeten we p en q vermenigvuldigen. De grootste waarde krijgen we bij 50 x 50 (= 2500). Het product van 60 x 40 is 2400, het product van 90 x 10 is 900. In de formule staat de waarde pq in de teller. Hoe hoger pq, des te hoger dus de steekproefomvang n. Met 50-50 spelen we op safe, zorgen we ervoor dat er niet een te kleine steekproefomvang uit de bus rolt.

Uitgaande van deze waarden, komt de benodigde steekproefomvang uit op 384.

$$n = z^2 . \frac{pq}{e^2} \qquad \text{Vb } n = 1{,}96^2 . \frac{50x50}{5^2} = 3{,}84 \text{ x } 100 = 384$$

In de praktijk berust veel onderzoek op een steekproefomvang van 400 à 500. Dat wil zeggen: 400 à 500 responses. De al dan niet bewuste gedachte hierachter is bovenstaande redenering.

We hebben p en q op 50 gesteld, maar zeer waarschijnlijk zullen deze in werkelijkheid daarvan afwijken. Stel we hebben 400 cases onderzocht en komen uit op percentages van 70 en 30%. We hadden dus, achteraf gezien, een iets kleinere steekproef kunnen nemen. Dat weten we alleen maar achteraf. Maar nu worden we beloond doordat de foutmarge iets kleiner uitvalt dan de vereiste 5%. Onderstaand kader vat de berekening samen.

1. De benodigde steekproefomvang (bij een betrouwbaarheidspercentage van 95%, een marge van 5% en een veronderstelde 50-50 verdeling), is
   n = 3,84 x 2500 / 25 = 3,84 x 100 = 384
2. We nemen nu een steekproef van 384 cases en vinden een verdeling van 70% - 30% (ja-nee)
3. De marge e zal iets kleiner uitvallen dan onze eis van 5%.

$$e = z . \sqrt{\frac{pq}{n}} \qquad = 1.96 \sqrt{\frac{70x30}{384}} = 4{,}58$$

4. Conclusie: we zijn er voor 95% zeker van dat 'ja' geldt voor 70 ± 4,58% van de populatie.

7.2. Afwijkende situaties

1. Strengere eisen

Strengere eisen (een betrouwbaarheidsniveau van 99% en een kleinere foutmarge) leiden tot een grotere benodigde steekproefomvang. Figuur 12

laat enkele waarden zien. Daarbij is telkens uitgegaan van een 50-50 verdeling. Met name voor een kleinere foutmarge betalen we een hoge prijs.

	±5%	±4%	±3%	±2%	±1%
95%	384	600	1.067	2.401	9.604
99%	664	1.041	1.843	4.147	16.590

Figuur 12. Benodigde steekproefomvang, uitgaande van een 50-50 verdeling, bij verschillende betrouwbaarheidsniveaus en foutmarges.

2. Kleine populatieomvang

Vaak wordt gedacht dat de omvang van de steekproef (n) in een bepaalde verhouding moet staan tot de omvang van de populatie (N). Dat dat niet juist is, blijkt alleen al uit het feit dat de populatieomvang geen deel uitmaakt van de formule die we tot nu toe gebruikten. Bij kleine aantallen speelt de omvang van de populatie nog wel een rol. In volledige vorm luidt de formule:

$$n = \frac{z^2 \cdot pq}{e^2 + z^2 \cdot \frac{pq}{N}}$$

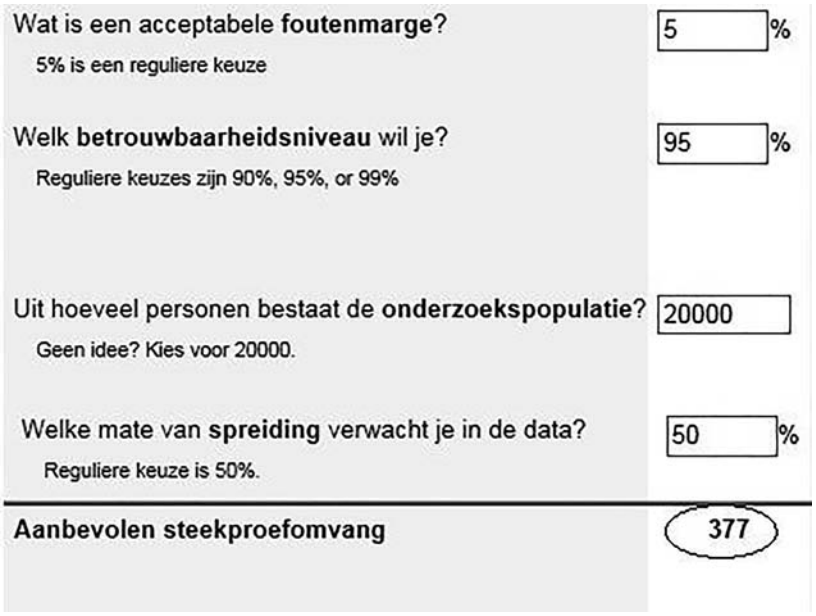
In figuur 13 zijn enkele waarden ingevuld. Deze maken duidelijk dat de benodigde steekproefomvang zeker niet evenredig stijgt met de populatieomvang. De waarde in de onderste rij kennen we al van figuur 12.

Omvang populatie	±5%	±4%	±3%	±2%	±1%
100	80	86	92	96	99
200	132	150	169	185	196
300	169	200	235	267	291
400	196	240	291	343	384
500	218	273	341	414	476
1000	278	375	517	706	906
1500	306	429	624	924	1298
5000	357	536	880	1622	3288
10000	370	567	964	1936	4898
∞	384	600	1067	2401	9604

Figuur 13. Benodigde steekproefomvang bij kleine populaties, betrouwbaarheidsniveau 95%.

Op internet zijn diverse tools die handmatige berekening overbodig maken:
http://www.steekproefcalculator.com/steekproefcalculator.htm
http://www.journalinks.be/steekproef/
http://www.corpos.nl/producten/Steekproef/steekproefcalculator.html

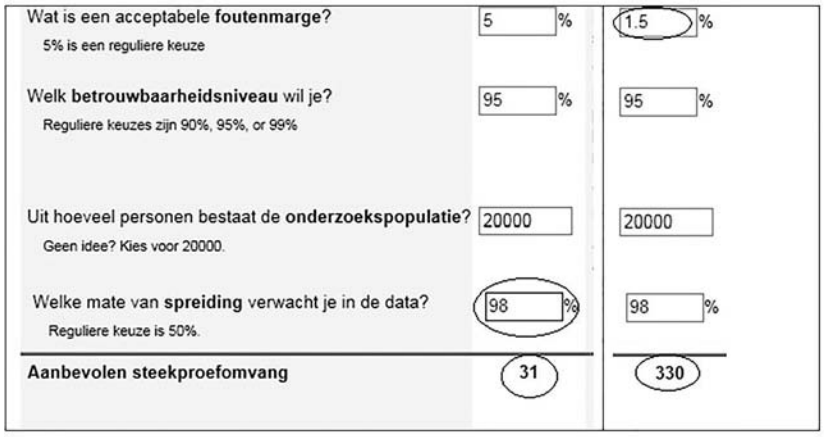
Deze vragen je een betrouwbaarheidsniveau, percentageverdeling en foutmarge in te vullen, met als resultaat de benodigde steekproefomvang. Figuur 14 geeft hiervan een indruk.



Figuur 14. Voorbeeld steekproefcalculator.

3. Scheve verdeling
Tot nu toe zijn we uitgegaan van een 50-50 verdeling. Dat is niet altijd van toepassing. We willen bijvoorbeeld aan de hand van een steekproef meten hoeveel procent van een verzameling gedigitaliseerde pagina's wel of niet voldoet aan de kwaliteitseisen. We gaan dan eerder uit van een 90-10, of misschien wel 98-2 verhouding. Op het eerste gezicht heeft dat een bijzonder gunstig effect: de benodigde steekproefomvang zakt naar 31. Dit

lijkt te mooi om waar te zijn en dat is het ook. We vergeten dat een foutmarge van 5% in dit geval relatief erg groot is. Als we dat terugbrengen tot een meer acceptabele marge van bijvoorbeeld 1,5%, zien we de steekproefomvang alweer stijgen naar 330.



Figuur 15. Effect van scheve percentageverhouding (98-2).

Tegen extreem lage percentages is eigenlijk geen kruid gewassen. Stel we hebben een database met een miljoen records en weten dat er één record tussen zit van een boek in het Armeens (0,0001%). Zouden we deze er met een steekproef uit kunnen halen? Ook al nemen we een steekproef van 900.000 titels, dan nog is het zeer wel mogelijk dat die ene Armeense titel daar geen deel van uitmaakt. Er is maar één oplossing: onderzoek de gehele populatie.

4. Uitspraken over subgroepen

Tot nu toe zijn we er van uitgegaan dat we uitspraken willen doen over de gehele populatie. Het is niet mogelijk om met eenzelfde mate van nauwkeurigheid uitspraken te doen over subgroepen: geslacht, studierichting, leeftijdscategorie etcetera. We zouden dan de steekproefomvang per categorie moeten vaststellen. Dit wordt al gauw een dure zaak. We zouden dan 384 mannen en 384 vrouwen nodig hebben of 1920 personen verdeeld over vijf leeftijdscategorieën. Er moeten wel heel goede redenen zijn om hiertoe over te gaan.

5. Response
Objecten zoals titelrecords en gedigitaliseerde pagina's kunnen hun medewerking aan het onderzoek niet ontzeggen. Onderzoek onder mensen heeft vaak te leiden onder een lage response. Als we, afgaande op eerdere ervaringen, een respons verwachten van 25%, moet de steekproefomvang vier keer zo groot zijn. Als de calculator 384 aangeeft, kunnen we dat beter ophogen naar circa 1600.

7.3. Steekproefomvang bij onderzoek gericht op bepalen van gemiddelden
Wat geldt voor percentages, geldt ook voor gemiddelden. We zijn eerder uitgegaan van een formule, waarmee na afloop van het onderzoek de foutmarge berekend is, op basis van het gewenste betrouwbaarheidsniveau (95 of 99%), de standaarddeviatie en de steekproefomvang. We kunnen dat ook omdraaien, met als resultaat een formule om voorafgaand aan het onderzoek de benodigde steekproefomvang te berekenen.

$$e = z\sqrt{\frac{s^2}{n}} \qquad \rightarrow n = \frac{z^2 \cdot s^2}{e^2}$$

Een complicatie is dat we s (standaarddeviatie van steekproef) niet weten. We moeten het onderzoek immers nog doen. Het is ook niet gemakkelijk om te bepalen aan welke eisen de foutmarge e moet voldoen. Dit is nu geen percentage, maar een absoluut getal. We kunnen hier dus niet zomaar de waarde 5 invullen. Stel dat het gemiddelde 250.000 is (bijvoorbeeld de gemiddelde prijs van een koopwoning in euro's), dan zouden we voor e 1250 (5%) kunnen invullen. We zouden dus op voorhand ook een inschatting van het gemiddelde moeten maken. Eigenlijk is er maar één goede oplossing. Begin alvast met het onderzoek en bepaal na enkele tientallen cases het gemiddelde en de standaarddeviatie op dat moment. Vul vervolgens de formule in en je ziet hoe ver je de steekproef moet doortrekken. Bij een onderzoek gericht op objecten is dat heel goed mogelijk, bij een onderzoek gericht op mensen kan dat op praktische bezwaren stuiten.

Het is dus een stuk lastiger om de benodigde steekproefomvang te bepalen bij onderzoek dat beoogt om gemiddelden te berekenen. Dat

steekproefcalculators altijd uitgaan van percentages en niet van gemiddelden, is dan ook niet toevallig.

8. Slot

Niet iedereen heeft affiniteit met statistiek. Beschrijvende statistiek is meestal geen struikelblok, inductieve statistiek gaat een stap verder. Wie onderzoek doet of de onderzoeksliteratuur goed wil begrijpen, moet op zijn minst enigszins vertrouwd zijn met basale statistische principes. Wat betekent significantie nu eigenlijk en hoe moet je p=.188 lezen? Hoeveel waarde moeten we hechten aan significante verschillen of verbanden, in de wetenschap dat de steekproef heel groot was? Wat is een z-waarde ofwel standaardscore? Waarom hebben we vooral te maken met z-scores van 1,96 of 2,58? Wat houdt het theoretische concept van een steekproevenverdeling in? Het belang van de steekproevenverdeling is dat de eigenschappen van de normaalverdeling daarop toepasbaar zijn: wat zijn dit dan voor eigenschappen? Hoe komt een steekproefcalculator tot zijn resultaten?

Deze bijdrage hoopt het inzicht in de achtergronden van de inductieve statistiek te bevorderen. Het doel is niet om in de schoenen van wiskundigen te stappen en te achterhalen waarom formules zijn zoals ze zijn of waarom de gebieden binnen een normaalverdeling zijn zoals ze zijn. Dat is ons vak niet en kunnen we gevoeglijk aan anderen overlaten. Het doel is evenmin om enkele kookboekrecepten voor te schotelen en simpelweg te verwijzen naar toetsen zoals de t-toets of chi kwadraat, of om te wijzen op het bestaan van een steekproefcalculator. Dat zou juist weer te bescheiden zijn. Deze bijdrage bewandelt een tussenweg. Voor wie zich verder wil bekwamen in deze materie is een overvloed aan handboeken en internetsites beschikbaar.

Literatuur

Voorbij, Henk. De techniek van gebruikersonderzoek. *Handboek Informatiewetenschap voor bibliotheek en archief*. Red. G.M. van Trier et al. Vakmedianet bv, Alphen aan den Rijn, juli 2015, I 515.
www.iwabase.nl

Bijlage 1. Vertaling *z*-scores in gebiedsscores

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4855	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

Noten

1. http://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/unit6/calculatingt-testsonspss/