5. Statistische gevolgtrekking: schatten

Punt en intervalschattingen

Hoe gebruik je steekproef data voor het schatten van populatie parameters?

Met kwantitatieve variabelen schat je het populatie gemiddelde (bijv. hoeveel geld er gemiddeld is besteed aan medicijnen in 2011). Met categoriale variabelen schat je populatie proporties voor de categorieën (bijv. wie er wel of geen zorgverzekering heeft in 2011).

Er zijn twee typen parameterschattingen:

- Puntschatting (een getal dat de beste schatting is).
- Interval schatting (een interval rond een puntschatting, waarvan je denkt dat de populatie parameter er in valt).

Er is een verschil tussen een estimator en een estimate point (schatting), namelijk dat een estimator de manier is waarop je de schatting maakt (het schatten op zich), en een estimate point het getal dat eruit komt. Zo is je steekproef een estimator voor je populatie parameter, en is 0.73 (bijvoorbeeld) een estimate point voor je populatie gemiddelde. Dit betekent dat een bepaalde categorie op 73% wordt geschat. De korte naam van 'estimate point' is 'estimate'.

Puntschatting van parameters

Een goede schatting heeft een steekproefdistributie die 1) gecentreerd is rond de parameter, en 2) een zo klein mogelijke standaardfout heeft.

1: gecentreerd rond de parameter:

Een schatting is ‘niet vertekend’ (unbiased) wanneer de steekproefdistributie gecentreerd is rond de parameter. Helemaal natuurlijk wanneer het steekproefgemiddelde ook daadwerkelijk de populatieparameter is. Dus dan is □ (steekproefgemiddelde) gelijk aan μ (populatiegemiddelde). □ is dan een goede estimator voor μ.

Een schatting kan ook ‘vertekend’ (biased) zijn en dan is het steekproefgemiddelde geen goede schatting voor het populatiegemiddelde. Meestal zit het steekproefgemiddelde er dan onder, want de extremen in de steekproef kunnen nooit meer zijn dan die uit de populatie, alleen maar minder. Dus de distributie en variatie in de steekproef is dan kleiner, waardoor de steekproefvariatie de populatievariatie onderschat.

2: Een zo klein mogelijke standaardfout:

Een schatting is ‘efficiënt’ wanneer de standaardfout kleiner is dan die van andere schattingen (dit is de standaardfout van de estimators gemiddelde, mediaan etc. Dat zijn allemaal estimators).

Stel dat je een normale verdeling hebt. Bij een normale verdeling is de standaardfout van de mediaan altijd 25% groter dan de standaardfout van het gemiddelde. Het gemiddelde van de steekproef ligt dichterbij het gemiddelde van de populatie dan de steekproefmediaan. Het steekproefgemiddelde is dan een efficiëntere estimator dan de steekproefmediaan.

Samenvattend: een goede estimator is onpartijdig (unbiased; de steekproefdistributie is gecentreerd rond de parameter) en efficiënt (kleinste standaardfout).

Meestal gebruik je gewoon het steekproefgemiddelde als estimator voor het populatiegemiddelde, de steekproefstandaardafwijking als estimator voor de populatiestandaardafwijking, etc.

R.A. Fisher ontwikkelde de 'meest aannemelijke schatter'. Dit is een schattingsmethode die als schatting van een parameter die waarde kiest, waarvoor de aannemelijkheidsfunctie maximaal is. Hoe aannemelijk een parameterwaarde is, wordt gemeten aan de kans op het vinden van een steekproefuitkomst bij die waarde van de parameter.
Deze manier heeft drie voordelen, met name bij grote steekproeven: 1) ze zijn efficiënt: andere estimators hebben geen kleinere standaardfouten en liggen ook niet dichterbij de parameter, 2) ze zijn niet vertekend (minder vertekening wanneer de steekproef groter wordt), en 3) ze hebben meestal een normale steekproefverdeling.

Intervalschatting

Een betrouwbaarheidsinterval is een intervalschatting voor een parameter. In dit interval zitten betrouwbare schattingen van de parameter. Je kijkt hiervoor naar de distributie van de steekproef, welke vaak een normale verdeling is. Wanneer je een betrouwbaarheidsinterval wilt met 95% zekerheid, valt de schatting van de parameter binnen twee standaardafwijkingen van het gemiddelde. In de praktijk vermenigvuldig je eerst de standaardfout met de z-waarde. De uitkomst tel je dan bij de puntschatting op en trek je van de puntschatting af. Je krijgt twee getallen, die samen het betrouwbaarheidsinterval vormen. Je kunt nu met 95% zekerheid zeggen dat een populatieparameter tussen deze twee getallen ligt. De z-waarde maal de standaardfout noem je ook wel de 'foutmarge' (margin of error).

Dus een intervalschatting is: de puntschatting ± de foutmarge.

Als je een intervalschatting wilt met 95% zekerheid, dan moet je een puntschatting ± een foutmarge doen (die foutmarge moet dan gelijk zijn aan twee standaardfouten).

Betrouwbaarheidsinterval voor een proportie

Nominale en ordinale variabelen zorgen voor categoriale data (Bijv. 'mee eens' – 'niet mee eens'). Als je hier uitspraken over wilt doen, kun je geen gemiddelden berekenen. Je gebruikt dan proporties of percentages. Een proportie valt tussen de 0 en de 1, en een percentage tussen de 0 en de 100.

De onbekende proportie van een populatie wordt aangeduid met het teken: π . De punt schatter van een populatie proportie is de 'steekproef proportie'. Hiermee schat je de populatie proportie. Je geeft de steekproef proportie aan met het teken: $\hat{\pi}$ (want het is een schatting van de daadwerkelijke proportie).

De z-waarde geeft aan hoe vaak de standaardfout vermenigvuldigd moet worden. Het zit zo in elkaar dat de kans onder een normale verdeling binnen z standaardfouten vanaf het gemiddelde gelijk is aan het betrouwbaarheidsniveau. Voor een betrouwbaarheidsniveau van 95% en 99% is z gelijk aan 1.96 en 2.58. De steekproefgrootte n moet zo groot zijn dat op zijn minst 15 observaties in de categorie van de behandelde proportie vallen en 15 buiten de categorie vallen, anders werkt het betrouwbaarheidsinterval niet.

Betrouwbaarheidsinterval voor populatie

proportie $B.I. = \hat{\pi} \pm z \cdot (se)$ waarbij $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

Hieruit kun je afleiden dat een grotere steekproef een meer accuraat betrouwbaarheidsinterval zal geven. Een grotere n zorgt voor een kleinere standaardfout, en

een preciezer betrouwbaarheidsinterval. Meer specifiek: de steekproefgrootte moet verviervoudigen om de precisie te verdubbelen.

Betrouwbaarheidsinterval voor een gemiddelde

Ook hierbij heeft de betrouwbaarheidsinterval de vorm : puntschatting ± foutmarge. De foutmarge bestaat hier uit een t-waarde (in plaats van een z-waarde) maal de standaardfout.

De standaardfout wordt berekend door de standaarddeviatie van de steekproef (s) te delen door de wortel van de steekproefgrootte (n). De puntschatting is in dit geval het steekproefgemiddelde, aangeduid met het teken: $\bar{y}$.

Betrouwbaarheidsinterval voor populatie gemiddelde:

$$B.I. = \bar{y} \pm t_{.025} \cdot (se) \quad \text{met} \quad se = \frac{s}{\sqrt{n}}$$

Kenmerken t-verdeling

- klok-vormig en symmetrisch vanaf het gemiddelde
- Standaarddeviatie is iets groter dan 1. De precieze waarde ervan hangt af van de vrijheidsgraden (df). Deze worden berekend aan de hand van de vrijheidsgraden (n – 1).
- Hoe groter de vrijheidsgraden (df), hoe meer de t-verdeling gaat lijken op een normaal verdeling. De verdeling wordt steeds puntiger. Bij df > 30 zijn ze bijna identiek.

Tabel: Schattingsmethoden voor gemiddelden en proporties.

Parameter	Puntschatting	Geschatte standaardfout	Betrouwbaarheidsinterval
Gemiddelde μ	$\bar{y}$	$se = \frac{s}{\sqrt{n}}$	$\bar{y} \pm t \cdot (se)$
Proportie π	$\hat{\pi}$	$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi} \pm z \cdot (se)$

6. Statistische gevolgtrekking: Significantie toetsen

De vijf delen van een significantie toets

Een significantie toets vergelijkt puntschattingen van parameters met de verwachte waarden van de nulhypothese. Significantie toetsen, ook wel 'hypothese toetsen' of in het kort 'toetsen' genoemd, bestaan uit vijf delen:

- Assumpties. Elke test doet assumpties over het type data (kwantitatief/categorisch), de vereiste randomisatie, de populatie verdeling en de steekproefgrootte.
- Hypothesen. Elke test heeft twee hypothesen, de nulhypothese (H0) en de alternatieve hypothese (Ha). De nulhypothese veronderstelt dat er geen effect is, de alternatieve hypothese stelt dat er ‘een’ effect is.
- Toetsingsgrootheid. Deze geeft aan hoe ver de schatting af ligt van de parameter waarde van H0. Dit wordt vaak weergegeven door het aan standaardfouten tussen de schatting en de H0-waarde.
- P-waarde. Deze geeft de kans dat, in de verdeling gegeven door de nulhypothese, de waarde van de toetsingsgrootheid wordt behaald of overschreden. Hij geeft aan hoe extreem de gevonden waarde is in de verdeling onder de nulhypothese. De p-waarde wordt weergegeven door ‘p’.
- Conclusie. Deze hoort de p-waarde te interpreteren, en zo een uitspraak te doen over H0 (verwerpen/aannemen).

Significantietoets voor een gemiddelde

Bij kwantitatieve variabelen wordt gebruikt gemaakt van het populatie gemiddelde μ. Dit wordt doorgenomen aan de hand van de vijf delen:

1. Assumpties
Er wordt aangenomen dat de data is verkregen uit een willekeurige steekproef, en normaal verdeeld is..

2. Hypothesen
De H0 voor deze toets heeft meestal deze vorm, H0: μ = μ0. Waarbij μ0 de waarde is van het populatiegemiddelde. Deze hypothese geeft meestal aan dat er geen effect of geen verschil is. De Ha geeft dan de overige waarden aan en heeft meestal deze vorm, Ha: μ ≠ μ0.

3. Toetsingsgrootheid
De toetsingsgrootheid is hier de t-score. Deze wordt berekend met deze formule:
$t = \frac{\bar{y} - \mu 0}{se}$ met $se = \frac{s}{\sqrt{n}}$
Het steekproefgemiddelde $\bar{y}$ schat het populatiegemiddelde μ. Onder de aanname dat H0 waar is, zal het gemiddelde van de verdeling van $\bar{y}$ gelijk zijn aan de waarde van μ0. Een waarde van $\bar{y}$ die ver in de staart van de verdeling valt, geeft sterk bewijs tegen H0, omdat het onwaarschijnlijk zou zijn dat je die waarde tegenkomt wanneer μ = μ0.
Hoe verder $\bar{y}$ van μ0 af ligt, des te groter zal de t-score zijn, en daarmee des te sterker het bewijs tegen H0.

4. P-waarde
De p-waarde geeft de kans aan dat je je geobserveerde data vindt als H0 waar is.

Voorbeeld: stel dat t = 0,68 met een steekproefgrootte van n = 186. Het aantal vrijheidsgraden df is dan 185. Dit is een grote steekproef, bijna identiek aan de standaard normaal verdeling.
Deze t-score betekent dat $\bar{y}$ 0,68 standaardfouten boven of onder μ0 ligt en omdat dit een standaard normaal verdeling benadert, kun je deze score opzoeken als z-score in tabel A achterin het boek. Bij deze z-score hoort een kans van bijna 0.025 per staart, dus 0.50 voor twee staarten.
De p-waarde is de kans dat t ≥ 0,68 of t ≤ -0,68. Deze is dus 0.50.

5. Conclusie
Hoe kleiner de p-waarde, des te sterker het bewijs tegen H0. Meestal wordt de H0 verworpen als p< 0.05 of p <0.01. Deze grenswaarde wordt bepaald door het alfa of significantie niveau, weergegeven met α .

Eenzijdige hypothese toetsen

Bij tweezijdige hypothese toetsen bevindt de kritische regio zich aan beide kanten (beide staarten) van de normale verdeling. In de meeste gevallen wordt een hypothese tweezijdig getoetst. In sommige gevallen heeft een onderzoeker echter al een vermoeden over de richting van een effect. Hij kan bijvoorbeeld vermoeden dat een specifiek voedingswaar ervoor zorgt dat mensen aankomen. Ook kan hij denken dat een therapievorm depressie vermindert. In dit soort gevallen is het beter om eenzijdig te toetsen. Op deze manier kan een specifiek vermoeden makkelijker getoetst worden. Bij een eenzijdige toets bevindt de kritische regio zich alleen in één staart van de normale verdeling. Welke staart dit is, hangt af van de alternatieve hypothese. Als er in de alternatieve hypothese staat dat gewicht na inname van een product zal toenemen, bevindt de kritische regio zich in de rechter staart. Als de alternatieve hypothese echter beweert dat gewicht zal afnemen van het consumeren van een product, dan zal de kritische regio zich in de linker staart bevinden. Dit omdat de min- en pluswaarden van z-scores van links naar rechts oplopen.
Let op: Bij tweezijdig toetsen moet de kans op een z-waarde verdubbeld worden. Bij eenzijdig toetsen kan de kans op een z-waarde direct uit tabel A (achterin het boek) gehaald worden.

Eenzijdig en tweezijdig toetsen

Alle onderzoekers zijn het erover eens dat een- en tweezijdige toetsing verschillende dingen zijn. Sommige onderzoekers vinden dat een tweezijdige hypothese toets altijd overtuigender is dan een eenzijdige toets. Dit omdat er bij een tweezijdige toets meer bewijs nodig is om de nulhypothese af te wijzen. Andere onderzoekers prefereren juist eenzijdige toetsen, omdat deze toetsen de uitkomsten zijn van een hele specifieke hypothese. Een eenzijdige toets is volgens hen gevoeliger. Een klein behandelingseffect kan significant zijn bij een eenzijdige toets terwijl hetzelfde effect niet significant is bij een tweezijdige toets. In het algemeen kan gesteld worden dat tweezijdige toetsen gebruikt zouden moeten worden in onderzoekssituaties waarin er geen vermoeden is over de richting van een effect.

Effectgrootte

Sommige onderzoekers hebben kritiek op het proces van hypothesen testen. De grootste kritiek gaat over de interpretatie van een significant resultaat. Er wordt bij het testen van een hypothese namelijk vooral aandacht besteed aan data en niet aan de hypothesen zelf. Als de nulhypothese wordt afgewezen, maken we een statement over de steekproef data en niet over de nulhypothese. Op basis van steekproef data wordt de nulhypothese dus afgewezen of behouden. Of de nulhypothese werkelijk (on)waar is, weten we niet. Een ander kritiekpunt

is dat een significant effect niet meteen zegt dat een behandeling een groot effect heeft. Iets is significant of niet, maar dit zegt niets over de grootte van het effect dat gevonden is. Een significant effect is dus niet hetzelfde als een groot effect. Om meer inzicht te krijgen in de grootte van een significant effect, is Cohen (1988) gekomen met de zogenaamde *effectgrootte.* Zijn maat voor effectgrootte noemen we *Cohen's d.* Deze maat kan berekend worden door eerst het verschil tussen het samplegemiddelde en het oorspronkelijke populatiegemiddelde te vinden (M- μ). Vervolgens wordt deze gedeeld door de standaarddeviatie van de populatie. De uitkomst van Cohen's d is 0.2 bij een klein effect, 0.5 bij een gemiddeld effect en 0.8 bij een groot effect.

Statistische power

Naast het meten van de effectgrootte is het ook mogelijk om de power van een statistische test te meten. De *power* van een test is de kans dat de test de nulhypothese zal afwijzen als deze ook echt fout is. De power gaat dus om het vinden van een effect als deze ook daadwerkelijk bestaat. Effectgrootte en de power van een test hebben echter wel een relatie. Als de effectgrootte stijgt, dan stijgt ook de kans om de nulhypothese af te wijzen. Dit betekent dat de power van een test op dat moment stijgt. Metingen van effectgrootte (zoals Cohen's d) en metingen van power geven beiden een indicatie van de grootte van een effect. De power van een test wordt beïnvloed door drie belangrijke factoren.

1. Allereerst speelt de grootte van een sample (n) een rol. Hoe groter een sample is, hoe groter de kans is om de nulhypothese af te wijzen als deze ook echt fout is. Dit betekent dat de power van een test groter wordt als de grootte van de sample stijgt.
2. Daarnaast wordt de power van een test verlaagd als het alfaniveau verkleind wordt. Als de alfa bijvoorbeeld verlaagd wordt van 5% naar 1% is de kans kleiner dat een effect (dat er in werkelijkheid wel is) gevonden wordt.
3. Ten derde stijgt de power van een test wanneer van een tweezijdige toets een eenzijdige toets wordt gemaakt.

Significantie

Hypothese toetsen worden vaak vermeld in wetenschappelijke literatuur. Er wordt bijvoorbeeld laten zien dat een behandeling een significant effect heeft gehad op depressiescores. Een *(statistisch) significant* effect houdt in dat de nulhypothese verworpen is. Het onderzoeksresultaat is dus heel waarschijnlijk niet ontstaan door toeval. Vaak wordt ook een z-score vermeld; bijvoorbeeld z=2.35. Achter de z-score wordt de p- waarde vermeld, bijvoorbeeld $p<0.05$. Wat houdt dit in? Bij een alfa niveau van 5% is de nulhypothese dus verworpen, aangezien het onderzoeksresultaat onder de 5% lag. Er is dus maar 5% kans (p=probability) dat zo'n resultaat verkregen is door alleen toeval verschijnselen (en niet door een echt effect). Omdat dit een kleine kans is, wordt de nulhypothese verworpen. Als de nulhypothese echter niet verworpen is, dan is de gevonden kans groter dan het alfa niveau. Als een therapievorm voor depressie niet effectief gebleken is, staat er bijvoorbeeld $p>0.05$. In de literatuur wordt nooit letterlijk gezegd dat de nulhypothese verworpen is. Dit moet de lezer zelf concluderen wanneer er wordt gesproken over een (statistisch) significant effect. Hoe groter de gevonden z-score is, des te groter de kans op een statistisch significant effect is. Er zijn verschillende factoren die een rol spelen wanneer besloten moet worden of een z-score groot genoeg is om de nulhypothese af te wijzen.

Significantie toets voor een proportie

Bij een categorische variabele kijken we naar de steekproef proportie om de populatie proportie te toetsen.

1. Assumpties
Er worden aannames gemaakt dat het een willekeurige steekproef is, uit een normale verdeling. De steekproefgrootte moet minstens 20 zijn.

2. Hypothesen
We geven steekproef proportie weer met π en populatie proportie met π. De nulhypothese stelt dat er geen effect is of niets aan de hand, dus dat de steekproef proportie gelijk moet zijn aan de populatie proportie, H0: π = π0, waarbij π0 de waarde van de populatie proportie is. De alternatieve hypothese is dan alle andere waarden (bij tweezijdig), Ha: π ≠ π0.

3. Toetsingsgrootheid
We gebruiken nu een z-score. Deze berekenen we als volgt:

$$z = \frac{\hat{\pi} - \pi_0}{se_0} \quad \text{waarbij} \quad se_0 = \frac{\sqrt{\pi_0}(1 - \pi_0)}{n}$$

Deze z-score meet hoe veel standaardfouten de steekproef proportie verwijderd ligt van de populatie proportie. Voor grote steekproeven (>20), en wanneer H0 waar is, is de verdeling van toetsingsgrootheid z gelijk aan die van de normale verdeling.

4. P-waarde
Voor het opzoeken van de p-waarde moet gekeken worden in de distributietabel van de normale verdeling. Deze p-waarde geeft aan hoe groot de kans is dat je je geobserveerde proportie vindt als H0 waar is.

5. Conclusie
Ook bij de toets van een proportie geldt: Hoe kleiner de p-waarde, des te sterker het bewijs tegen H0.

Conclusies en typen fouten in toetsen

Type 1 fout

Het testen van hypothesen is een *inferentieel proces.* Dit betekent dat een beperkte hoeveelheid informatie (namelijk een sample) wordt gebruikt om een algemene conclusie te trekken. Het is mogelijk dat de data ervoor zorgt dat je als onderzoeker denkt dat de nulhypothese afgewezen moet worden terwijl de behandeling eigenlijk geen effect heeft. Dit kan gebeuren omdat samples niet identiek zijn aan populaties. De onderzoeker kan toevallig een extreme sample geselecteerd hebben, waardoor het lijkt dat een behandeling effect heeft gehad terwijl dat niet zo is. Dit noemen we een *type 1 fout.* Zo een fout kan grote gevolgen hebben. Een onderzoeker kan namelijk ten onrechte publiceren dat zijn behandelingsmethode effectief is gebleken. Er is echter maar een hele kleine kans dat een onderzoeker zo een fout maakt. Het alfa niveau laat zien hoe groot de kans is dat een type 1 fout gemaakt wordt. In de meeste gevallen is er dus maar 5% kans dat er zo een fout gemaakt zal worden. Als de onderzoeker strenger wil toetsen, kan een alfa van 2,5 of 1% ook gebruikt worden. Er is dan maar een zeer kleine kans op een type 1 fout. Lagere alfa niveaus geven minder kans op een type 1 fout, maar een lager alfa niveau brengt ook met zich mee dat er relatief meer bewijs uit de data moet blijken om de nulhypothese te kunnen afwijzen.

Type 2 fout

Van een type 2 fout is sprake wanneer een onderzoeker een nulhypothese niet afwijst, terwijl deze echt verkeerd is. De hypothese toets heeft dus een behandelingseffect (dat er in het echt wel is) niet gevonden. Een type 2 fout komt voor wanneer een steekproef gemiddelde zich niet in de kritische regio bevindt, terwijl de behandeling wel een effect heeft gehad op de sample. De gevolgen van een type 2 fout zijn vaak minder ernstig dan de gevolgen van een type 1 fout. Bij een type 2 fout heeft de onderzoeksdata niet kunnen laten zien waar de onderzoeker op had gehoopt. Er is niet een precieze waarde uit te rekenen voor een type 2 fout. Dit in tegenstelling tot de type 1 fout, waarbij het alfaniveau de kans aangeeft op een type 1 fout. De kans op een type 2 fout hangt af van vele factoren. Deze kans wordt aangeduid met de Griekse letter β.

Tabel: Keuzemogelijkheden in een significantietoets

	H0 verwerpen	**H0 niet verwerpen**
H0 in werkelijkheid waar	Type 1 fout	Goede beslissing
H0 in werkelijkheid niet waar	Goede beslissing	Type 2 fout

6.5 Beperkingen van significantie toetsen

Het is belangrijk rekening te houden met het feit dat statistische significantie niet hetzelfde is als praktische significantie. Een significant effect vinden betekent niet dat het een belangrijke vondst is in een praktische zin. Bij grote steekproeven kunnen p-waarden heel klein zijn, zelfs wanneer de puntschatting dicht bij de H0-waarde valt. De grootte van P geeft simpelweg aan hoeveel bewijs er is tegen H0, niet hoe ver de parameter verwijderd is van H0.

Daarbij is het misleidend om alleen onderzoeken te rapporteren die significante effecten hebben gevonden. Zo kan er 20 keer hetzelfde onderzoek zijn uitgevoerd, met maar 1 keer een significant effect gevonden te hebben. Als alleen dat onderzoek wordt gerapporteerd, ontstaat er een verkeerd beeld over de situatie. Dit resultaat kan immers gewoon per toeval gevonden zijn.

Ook moet de p-waarde niet geïnterpreteerd worden als de kans dat H0 waar is. In de werkelijkheid is H0 geen kwestie van kansen: het is waar, of het is niet waar. Bepaalde resultaten kunnen niet ‘significanter’ zijn dan andere.

Tabel van significantie toetsen

Parameter	Gemiddelde	Proportie
Assumpties	Willekeurig getrokken steekproef, kwantitatieve variabele, normaal verdeelde populatie.	Willekeurig getrokken steekproef, categorische variabele, een steekproefgrootte van minstens 20.

Hypothesen	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$ $H_a: \mu > \mu_0$ $H_a: \mu < \mu_0$	$H_0: \pi = \pi_0$ $H_a: \pi \neq \pi_0$ $H_a: \pi > \pi_0$ $H_a: \pi < \pi_0$
Toetsingsgrootheid	$t = \frac{\bar{y} - \mu_0}{se}$ met $se = \frac{s}{\sqrt{n}}$ en $df = n - 1$	$z = \frac{\hat{\pi} - \pi_0}{se_0}$ waarbij $se_0 = \frac{\sqrt{\pi_0}(1 - \pi_0)}{n}$
P-waarde	Kans van twee staarten in steekproef verdeling voor tweezijdige toets ($H_0: \pi \neq \pi_0$ of $H_a: \mu \neq \mu_0$) en kans van één staart in steekproef verdeling voor eenzijdige toets.	
Conclusie	Verwerp H0 als p-waarde kleiner dan of gelijk het alfa niveau, α is. Zoals een α van 0.05	

7. Het vergelijken van twee groepen

Het vergelijken van groepen

Vaak worden twee groepen met elkaar vergeleken. Bij kwantitatieve variabelen kijk je dan naar gemiddelden, en bij categoriale variabelen kijk je dan naar proporties. Wanneer je twee groepen met elkaar vergelijkt, creëer je een binaire variabele: een variabele met twee categorieën (soms ook wel dichotoom genoemd). Stel bijvoorbeeld dat je mannen en vrouwen vergelijkt, dan creëer je een binaire variabele 'geslacht' met de categorieën mannen en vrouwen. Het vergelijken van deze groepen is een voorbeeld van een bivariate statistische methode.

Twee groepen kunnen afhankelijk en onafhankelijk van elkaar zijn. De groepen zijn afhankelijk wanneer de respondenten van nature 'matchen' met elkaar, bijvoorbeeld wanneer je dezelfde groep gebruikt voor en na een meting. Stel dat je wilt weten of studenten betere resultaten hebben na een bepaald lesprogramma, dan is de kans groot dat de studenten die al beter presteerden voor het lesprogramma ook beter presteren na het programma. De twee resultaten zijn dus afhankelijk van elkaar. Groepen zijn onafhankelijk wanneer er geen sprake is van 'matching' tussen de groepen, bijvoorbeeld wanneer je gebruik maakt van randomisatie.

Standaardfout van groepsverschil

Stel dat we twee groepen met elkaar vergelijken: mannen en vrouwen en hun tijdsbesteding aan koken. Mannen en vrouwen zijn twee groepen, met allebei een ander populatiegemiddelde en een andere schatting daarvan. Je hebt dan ook twee standaardfouten. De standaardfout geeft namelijk aan hoe precies je schatting van de parameter is. Omdat we het verschil tussen mannen en vrouwen in de populatie willen weten, heeft ook dit verschil een standaardfout (want je schat het populatieverschil met je steekproefverschil).

De formule voor de geschatte standaardfout (van het verschil) is:
$\sqrt{(se_1)^2 + (se_2)^2}$ met $se = \frac{s}{\sqrt{n}}$. Hierbij is se1 de standaardfout van groep 1 (mannen) en se2 de standaardfout van groep 2 (vrouwen). Omdat je hier met twee groepen werkt, heb je ook twee keer een 'n', namelijk het aantal mannen en het aantal vrouwen. Dit geven we aan met n1 en n2. De 's' staat voor de standaard deviatie van de groep, en daar heb je er hier twee van. Deze geven we weer met s1 en s2. De formule voor de geschatte standaardfout kun je dan ook zo schrijven: $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

Betrouwbaarheidsinterval van groepsverschil

Het betrouwbaarheidsinterval bestaat uit je puntschatting van het verschil ± de t-score maal de standaardfout. De formule ziet er voor het groepsverschil zo uit:

$(\bar{y}_2 - \bar{y}_1) \pm t\,(se)$ waarbij $se = \frac{s}{\sqrt{n}}$

Wanneer het betrouwbaarheidsinterval positieve waarden aangeeft, dan betekent dat dat μ□ - μ□ positief is, en dus dat μ□ groter is dan μ□. Wanneer het betrouwbaarheidsinterval negatieve waarden heeft, betekent het dan ook dat μ□ kleiner is dan μ□.

Significantie toets van groepsverschil

We kunnen testen of de twee groepen significant van elkaar verschillen. Normaal wordt toetsingsgrootheid t berekend door de geschatte parameter min de nulhypothese te doen en die te delen door de standaardfout van de schatting. De geschatte parameter is hier het verschil tussen de twee groepen (dus y2-y1). Je nulhypothese stelt dat er 'niets' aan de hand is en dat er geen verschil is tussen mannen en vrouwen: het verschil is 0. De standaardfout werd berekend met de formule: $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

De formule voor toetsingsgrootheid t ziet er dan zo uit:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}$$

Voorbeeld

We gaan met een voorbeeld de standaardfout berekenen, het betrouwbaarheidsinterval en vervolgens het verschil toetsen. Stel dat we het verschil tussen mannen en vrouwen in de tijd (aantal minuten per dag) die zij besteden aan huishoudelijk werk bekijken. Dit zijn de gegevens:

Geslacht	Steekproefgrootte	Gemiddelde	Standaarddeviatie
Mannen	1219	23	32
Vrouwen	733	37	16

Standaardfout:

De formule voor de standaardfout (se) is: $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

Als we dit invullen met onze gegevens: $(32)^2/1219 + (16)^2/733 = 1{,}09$

Betrouwbaarheidsinterval:

De formule voor de betrouwbaarheidsinterval is: $(\bar{y}_2 - \bar{y}_1) \pm t\,(se)$. Bij een alpha van .05 moeten we gebruik maken van de t-waarde ±1,96. We hebben berekend dat de standaardfout 1,09 is. De gemiddeldes van de mannen en vrouwen zijn gegeven. We vullen de formule in: (37 – 23) ± 1,96 (1,09) = (12 ; 16).

Significantie toets

De formule voor de t-toets is $t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}$.

We kunnen de formule weer gewoon invullen: $\bar{y}_2 - \bar{y}_1$ = 37 – 23 = 14. Dus 14 – 0 / 1,09 = 12,8. Wanneer we de t-waarde 12,8 opzoeken in de tabel van de t-distributie, dan zien we dat deze een p-waarde heeft van <.000. Het is dus een significant verschil.

8. Analyseren van samenhang tussen categorische variabelen

Er bestaat een samenhang tussen twee variabelen als de verdeling van de respons (afhankelijke) variabele verandert op het moment dat de waarde van de verklarende (onafhankelijke) variabele verandert. We gaan nu kijken hoe je zo'n verband kunt vaststellen bij twee categorische, dus ordinale of nominale, variabelen.

Terminologie voor categorische data-analyse en statistische onafhankelijkheid.

Marginale verdeling

Categorische data worden vaak weergegeven in een kruistabel. Bijvoorbeeld deze:

Partijvoorkeur				
Geslacht	Democraten	Onafhankelijk	Republikeins	Totaal
Vrouw	573	516	422	1511
Man	386	475	399	1260
Totaal	959	991	821	2771

Deze tabel heeft twee rijen (man-vrouw) en drie kolommen (D-O-R). De rij- en kolomtotalen noemen we de marginale verdeling. Wil je van deze tabel bijvoorbeeld de marginale verdeling van partijvoorkeur weergeven, dan schrijf je (959, 991, 821).

Het maken van zo'n kruistabel is de eerste stap bij het doen van data-analyse met categorische variabelen. De tweede stap is om de absolute getallen om te zetten in percentages.

Conditionele verdeling

We willen weten of er een samenhang bestaat tussen geslacht en stemgedrag. We willen dus weten of het stemgedrag anders is voor mannen en voor vrouwen. We moeten dan kijken naar het stemgedrag van de mannen, en dat van de vrouwen. Daarom gebruiken we bij het berekenen van de percentages niet de totale groep.

De tabel met de percentages ziet er zo uit:

Partijvoorkeur					
Geslacht	Democraten	Onafhankelijk	Republikeins	Totaal	n

Vrouw	38%	34%	28%	100%	1511
Man	31%	38%	32%	100%	1260

*Toelichting Tabel: Vrouw-Democraten : 573/1511*100 = 38%. Man-Democraten : 959/1260*100 = 31%. Etc.*

Nu heb je de relatieve data verdeling van partijvoorkeur, afhankelijk van geslacht. Deze sets van percentages noemen we de conditionele distributie van partijvoorkeur. De conditionele verdeling van de vrouwen is (38, 34 28) voor (D,O,R). De conditionele distributie van de mannen is (31, 38, 32) voor (D,O,R).

Je kunt natuurlijk ook conditionele distributies maken voor geslacht per partijvoorkeur. Dan kreeg je bij de vrouwen (573/959*100) 60% en voor mannen (386/959*100) 40%. Meestal maak je echter een conditionele distributie voor de afhankelijke variabele.

Dus wanneer er wordt gevraagd om een conditionele distributie te maken van de response variabelen, binnen de categorieën van de verklarende variabele, doe je dat zoals bovenstaand.

Joint distribution (simultane verdeling)

Je kunt de percentages ook op een andere manier weergeven. Je berekent dan de percentages ten opzichte van de totale steekproef. Dan zou het er zo uit zien:

Partijvoorkeur			
Geslacht	Democraten	Onafhankelijk	Republikeins
Vrouw	21%	19%	15%
Man	14%	17%	14%

*Toelichting: Vrouw-Democraten : 573/2771*100 = 21%. Man-Democraten : 386/2771*100 = 14%. Etc.*

Deze verdelingen noemen we de simultane verdelingen. Maar wanneer je kijkt naar een respons (afhankelijke) en verklarende (onafhankelijke) variabele is het zinniger om te kijken naar conditionele verdelingen dan naar simultane verdelingen.

Statistisch (on)afhankelijk

Twee categorische variabelen zijn statistisch onafhankelijk wanneer de kans op het voorkomen van de ene ‘gebeurtenis’ los staat van de kans dat de andere ‘gebeurtenis’ voor komt. Anders gezegd: ze zijn statistisch onafhankelijk wanneer de kansverdeling van de mogelijke uitkomsten van de ene variabele niet wordt beïnvloedt door de uitkomsten van de andere variabele. Gebeurt dat wel, dan zijn ze statistisch afhankelijk.

Stel dat bij ons voorbeeld de twee variabelen geslacht en partijvoorkeur onafhankelijk van elkaar zouden zijn, dan zouden de percentages zo zijn verdeeld dat je bij Democraten een even groot percentage mannen als vrouwen hebt, en bij Onafhankelijke en Republikeinen ook. Maar dat is niet het geval.

Chi-kwadraat toets

Wanneer we zeggen dat twee variabelen onafhankelijk zijn, hebben we het over variabelen in de populatie. We verwachten wel dat de verdeling in de steekproef min of meer gelijk is aan die in de populatie, maar dat is die nooit helemaal. We willen dus kijken of het waarschijnlijk is dat we deze verschillen bij toeval tegenkomen in de steekproef. Dus dat de variabelen in de populatie wel onafhankelijk zijn, maar vanwege de steekproeffout de verdeling toch niet helemaal gelijk is. We toetsen dan

H0: de variabelen zijn statistisch onafhankelijk

Ha: de variabelen zijn statistisch afhankelijk

Chi-kwadraat

We berekenen dit met de Chi-kwadraat. De Chi-kwadraat toets vergelijkt de geobserveerde frequenties met de frequenties die voldoen aan H0 (de verwachte frequenties). Je kunt deze het beste ook in een tabel zetten. Hierbij blijven de rij- en kolomtotalen hetzelfde, maar voldoen de getallen aan onafhankelijkheid. De tabel ziet er zo uit:

Partijvoorkeur				
Geslacht	Democraten	Onafhankelijk	Republikeins	Totaal
Vrouw	573 (522,9)	516 (540,4)	422 (447,7)	1511
Man	386 (436,1)	475 (450,6)	399 (373,3)	1260
Totaal	959	991	821	2771

*Toelichting: De getallen in de tabel zonder haakjes zijn de geobserveerde frequenties. De getallen tussen de haakjes zijn de verwachte frequenties als H0 waar is. Vrouw-Democraten : 1511/2771*959 = 522,9. Man-Democraten : 1260/2771*959 = 436,1. Etc. Je kunt ook rij-totaal*kolomtotaal/steekproeftotaal. Dus voor Vrouw-Democraten: (1511*959)/2771 = 522,9.*

Geobserveerde frequenties noteren we met 'f□'. De verwachte frequenties noteren we met 'f□'.

Deze gebruik je voor het berekenen van de Chi-kwadraat, dat we weergeven met het symbool X^2. De formule voor X^2 is: $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$.

Dus je berekent de verschillen tussen de geobserveerde en de verwachte frequentie. Deze kwadrateer je. Deze deel je door de verwachte frequentie. En dat doe je voor elke cel (= optellen).

Interpretatie van X²

Wanneer H0 waar is, dan zullen de geobserveerde frequenties (f□) dicht liggen bij de verwachte frequenties (f□), dan zal X² klein zijn.

Wanneer H0 niet waar is, dan zullen f□ en f□ niet dicht bij elkaar liggen, waardoor de X² groot zal zijn.

Hoe groter X², hoe groter de kans wordt dat je H0 kunt gaan verwerpen. Het wordt dan onwaarschijnlijker dat de verschillen die je hebt gevonden toevallig zijn.

Kenmerken van X²

De Chi-kwadraat verdeling geeft aan hoe groot X² moet zijn voordat je H0 kunt verwerpen. De verdeling heeft een aantal kenmerken:

- De verdeling is altijd positief. X² kan nooit negatief zijn.
- De verdeling is rechts scheef (lange staart rechts)
- De precieze vorm van de verdeling hangt af van het aantal vrijheidsgraden (df). Voor de Chi-kwadraat verdeling geldt : $\mu = df$; $\sigma = 2df$. De verdeling is rechts scheef, en wordt 'platter' naarmate df groter wordt.
- Hoe groter de kruistabellen zijn, hoe meer vrijheidsgraden je hebt, hoe groter je X² is.
- Hoe groter X², hoe groter de kans dat je H0 kunt gaan verwerpen.

Nu gaan we terug naar ons voorbeeld. We willen toetsen of er een relatie is tussen partijvoorkeur en geslacht. H0 : geslacht en partijvoorkeur zijn statistische onafhankelijk. Ha : geslacht en partijvoorkeur zijn statistisch afhankelijk. Wanneer je de X² uitrekent dan is deze 16,2. Dat kun je zelf narekenen. Of deze significant is moet je opzoeken in de tabel met alle p-waarden van de Chi-kwadraat verdeling. We hebben df = 2. Nu moet je in de tabel kijken, in de rij van df = 2, tussen welke twee getallen jouw Chi-waarde ligt. Kies de laagste. Als je dan omhoog gaat, heb je je p-waarde.

In ons voorbeeld zien we dat X² groter is dan 13,82. Dan omhoog zien we dat die hoort bij een p-waarde van 0.001. We verwerpen H0 en concluderen dat het erg onwaarschijnlijk is dat we deze verschillen per toeval zouden tegenkomen, en dat er dus wel degelijk een verband bestaat tussen geslacht en partijvoorkeur.

Je kunt alleen een Chi-kwadraat toets doen wanneer de verwachte frequenties (f□) in elke cel groter zijn dan 5.

Vrijheidsgraden bij een Chi-kwadraattoets

Het aantsl vrijheidsgraden bij een Chi-kwadraattoets (df) bereken je door : (r – 1) * (c – 1). Dit betekent dat je het aantal rijen neemt -1. Dat vermenigvuldig je door het aantal kolommen – 1. In onze tabel hadden we twee rijen en drie kolommen. Dus 1 * 2 = 2.

Residuen

De Chi-kwadraat toets zegt echter niks over de richting of de sterkte van de samenhang. Deze toets geeft alleen aan of de variabelen een verband hebben met elkaar. Er kan geen uitspraak worden gedaan over significante verschillen.

Daarom kijken we naar residuen. Een residu is het verschil tussen de geobserveerde en verwachte frequentie in een cel: f□ - f□. Kijken we bijvoorbeeld naar het residu van vrouw-democraten, dan is dat de geobserveerde frequentie (573) minus de verwachte frequentie (522,9) = 50,1.

Maar hoe weet je nu of een residu groot genoeg is, dat het onwaarschijnlijk is dat het toeval is dat je die tegenkomt? Daarvoor gebruik je de gestandaardiseerde residuen. Je krijgt dit gestandaardiseerde residu (z) door het residu te delen door de standaardfout. Deze standaardfout is de fout die je verwacht wanneer H0 waar zou zijn.

Nu kunnen we de formule voor gestandaardiseerde residuen weergeven:

$$Z = \frac{f_0 - f_e}{se} = \frac{f_0 - f_e}{\sqrt{f_e(1 - \text{rijproportie})(1 - \text{kolomproportie})}}$$

Van deze gestandaardiseerde residuen weten we dat deze normaal verdeeld zijn, met een gemiddelde van 0, met een standaarddeviatie van 1. Wanneer het gestandaardiseerde residu boven of onder de 3 komt, is dat genoeg bewijs voor een bestaand effect in die cel.

Associatiematen voor ordinale variabelen

Nu kijken we naar associatiematen voor kruistabellen met ordinale variabelen. Bij ordinale variabelen kan zich een positief of een negatief verband voordoen. Een positief verband is wanneer iemand hoog op x scoort, ook hoog op y scoort, en wie laag op x scoort, laag op y scoort. Een negatief verband is wanneer iemand hoog op x scoort en laag op y scoort, en wie laag op x scoort, hoog op y scoort.

Concordante en discordante paren

We gaan dit onderzoeken aan de hand van concordante en discordante paren. Een paar van observaties is concordant wanneer de persoon die, ten op zichte van de persoon in een lagere klasse, hoger scoort op de ene variabele ook hoger scoort op de andere variabele. Een paar van observaties is discordant wanneer de persoon die hoger scoort op de ene variabele lager scoort op de andere variabele. Hieronder staat de berekening van het aantal concordante en het aantal discordante paren.

Om deze berekening te begrijpen, moet je goed kijken welke getallen uit de tabel worden vermenigvuldigd. De vermenigvuldigingen worden uiteindelijk bij elkaar opgeteld. Als je het eenmaal ziet, kun je het makkelijk zelf doen.

Concordante paren

Stel we hebben de volgende tabel, met betrekking tot ‘geluk’ en ‘inkomen’. In totaal hebben 67 mensen een beneden gemiddeld inkomen, 68 mensen hebben een gemiddeld inkomen en 22 mensen hebben een boven gemiddeld inkomen. We gaan kijken of mensen gelukkiger worden met naarmate het inkomen stijgt en ongelukkiger naarmate het inkomen daalt. Dit bekijken we door middel van het aantal concordante paren.

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**

Beneden gemiddeld	16 (24%)	36 (54%)	15 (22%)	**67 (100%)**
Gemiddeld	11 (16%)	36 (53%)	21 (31%)	**68 (100%)**
Boven gemiddeld	2 (9%)	12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Je berekent de concordante paren (C) als volgt. Je begint linksboven in de hoek (Beneden gemiddeld, niet erg gelukkig). Je streept alles weg dat in dezelfde rij staat, en alles weg dat in dezelfde kolom staat:

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld	16 (24%)			**67 (100%)**
Gemiddeld		36 (53%)	21 (31%)	**68 (100%)**
Boven gemiddeld		12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Vervolgens gaan we wat er in de cel linksboven staat (16) vermenigvuldigen met alles in de overgebleven cellen. De overgebleven cellen bevatten namelijk allemaal personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele, ten op zichte van alle personen in de cel linksboven (16). De vermenigvuldiging van deze cellen met de cel linksboven, is het vormen van concordante paren in de klasse 'beneden gemiddeld' en 'niet erg gelukkig' (de cellen 'totaal' worden hierbij genegeerd). Die paren bereken je als volgt:

16 * (36 + 21 + 12 + 8) = 1232

Nu gaan we een cel naar rechts (beneden gemiddeld, redelijk gelukkig), en doen we hetzelfde: strepen alles in dezelfde rij weg, en alles in dezelfde kolom weg. De overgebleven cellen bevatten namelijk allemaal personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele, ten op zichte van alle personen in de tweede cel van links (36).

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld		36 (54%)		**67 (100%)**
Gemiddeld			21 (31%)	**68 (100%)**
Boven gemiddeld			8 (36%)	**22 (100%)**

Totaal	**29**	**84**	**44**	**157**

Het is belangrijk dat je alleen naar de getallen aan de rechterkant van je cel kijkt. Nu doen we weer het getal uit onze cel, maal de overgebleven getallen:

36 * (21 + 8) = 1044.

Als we nog een cel naar rechts gaan (beneden gemiddeld, heel erg gelukkig), dan zien we dat er niks overblijft aan de rechterkant, dus deze cel heeft geen concordante paren. Dus gaan we een rij naar beneden en beginnen we weer aan de linkerkant, bij cel (gemiddeld, niet erg gelukkig). We doen weer hetzelfde: doorstrepen van alles in dezelfde rij en kolom, en we kijken alleen naar wat er rechts (en onder) overblijft:

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld				**67 (100%)**
Gemiddeld	11 (16%)			**68 (100%)**
Boven gemiddeld		12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

We gaan deze cel (11) vermenigvuldigen met de overgebleven cellen. De overgebleven cellen bevatten namelijk allemaal personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele, ten op zichte van deze cel (11).

11 * (12 + 8) = 220

We gaan weer een cel naar rechts (gemiddeld, redelijk gelukkig), en doen we hetzelfde: strepen alles in dezelfde rij weg, en alles in dezelfde kolom weg. De overgebleven cel bevat namelijk personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele.

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld				**67 (100%)**

Gemiddeld		36 (53%)		**68 (100%)**
Boven gemiddeld			8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

We gaan deze cel (36) weer vermenigvuldigen met de overgebleven cel:

36 * 8 = 288.

Deze vier cellen (16, 36, 11, 36) waren de enige cellen die we konden paren met cellen rechtsonder. Dat betekent dat we nu alle concordante paren hebben berekend. Deze paren moet je optellen, dat geeft het totaal aantal concordante paren:

1232 + 1044 + 220 + 288 = 2784 concordante paren. Dus C = 2784

Discordante paren

Voor het berekenen van de discordante paren (D) doen we in principe hetzelfde, maar dan begin je rechtsboven in de hoek, en kijk je naar alles wat er linksonder overblijft. De cellen 'totaal' worden ook hierbij genegeerd. Hieronder is de hele tabel gegeven, zodat je zelf kan gaan wegstrepen.

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld	16 (24%)	36 (54%)	15 (22%)	**67 (100%)**
Gemiddeld	11 (16%)	36 (53%)	21 (31%)	**68 (100%)**
Boven gemiddeld	2 (9%)	12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Het eerste disconcordante paar wordt gevormd met de cel helemaal rechtsboven (15). Streep alle cellen in dezelfde rij en in dezelfde kolom weg. Vermenigvuldig daarna de cel (15) met alle cellen linksonder.

15 * (11 + 36 + 2 + 12) = 915

We schuiven een kolom naar links. Hiermee (36) wordt het volgende discordante paar gevormd. Streep alle cellen in dezelfde rij en in dezelfde kolom weg en vermenigvuldig met de overgebleven cellen:

36 * (11 + 2) = 468

We schuiven een rij naar onder. Begin weer helemaal rechts en vermenigvuldig deze cel (21) met alle cellen linksonder:

21 * (2 + 12) = 294

We schuiven een kolom naar links. Met deze cel (36) wordt het laatste discordante paar gevormd:

36 * 2 = 72

Deze vier cellen (15, 36, 21, 36) waren de enige cellen die we konden paren met cellen linksonder. Dat betekent dat we nu alle discordante paren hebben berekend. Deze paren moet je optellen, dat geeft het totaal aantal discordante paren:

915 + 468 + 294 + 72 = 1749 discordante paren. Dus D = 1749

Wanneer je een positief verschil hebt tussen de concordante en discordante paren (C – D) is er een positieve relatie tussen de twee variabelen (inkomen en geluk) . Wanneer C – D negatief is, is er een negatieve relatie tussen de twee variabelen.

Gamma

Omdat je bij grotere steekproeven ook meer paren hebt en vaker ook grotere verschillen tussen C en D, standaardiseren we dit verschil. Dit standaardiseren geeft de gamma (γ). De formule hiervoor is $\gamma = \frac{C-D}{C+D}$. We weten een aantal dingen over deze gamma: de waarde ervan valt tussen -1 en 1; gamma geeft aan of de relatie positief dan wel negatief is; hoe groter gamma, hoe sterker de samenhang tussen twee variabelen.

9. Lineaire Regressie en Correlatie

Lineaire verbanden

In dit hoofdstuk worden methoden gepresenteerd waarmee je kwantitatieve respons variabelen (afhankelijk) en verklarende variabelen (onafhankelijk) kunt analyseren. Dit gebeurt aan de hand van regressie analyse. Regressie analyse omvat drie analyses: 1) onderzoeken of er een verband bestaat tussen de variabelen, 2) de sterkte van dit verband bepalen, en 3) het maken van een regressieformule om zo de waarde van de response variabele te kunnen voorspellen aan de hand van de verklarende variabele.

Bij een lineair verband wordt de respons variabele weergegeven met 'y' en de verklarende variabele met 'x'. Met een lineaire functie kunnen we een rechte lijn trekken door de datapunten in een grafiek. Deze functie heeft deze vorm: y = a + b (x). Hierbij is 'α' de intercept, en 'β' de hellingscoëfficiënt.

Interpreteren van y-intercept en hellingscoëfficiënt

De y-intercept is de waarde van y wanneer x = 0. Want als x = 0, dan vervalt b(x), en hou je alleen y = β over. Daarbij geeft de y-intercept aan waar de lijn op de y-as begint.

De hellingscoëfficiënt geeft de verandering aan in y, bij een toename van 1 punt bij x. Wanneer x er 1 punt bij krijgt, verandert y met b.

Over het algemeen is het zo dat hoe groter β, hoe steiler de regressielijn. Als β positief is, betekent dat dat wanneer x hoger wordt, y ook hoger wordt. Dit kenmerkt een positief verband. Wanneer β negatief is, betekent het dat wanneer x hoger wordt, y lager wordt. Dit is een negatief verband. Wanneer β = 0, betekent het dat de waarde van y constant is en niet verandert wanneer x verandert. Dit kan betekenen dat de twee variabelen onafhankelijk van elkaar zijn.

De best passende regressielijn

Bij regressieanalyse beschouwen we a en b als onbekende parameters, die we gaan schatten aan de hand van de data. De eerste stap hierbij is het plotten van de data in een scatterplot. Zo kun je zien of het wel logisch om een lineaire formule te gaan maken. Wanneer de data immers een U vorm heeft, heeft het geen zin om daar een lineaire lijn door te trekken.

Omdat we y gaan benaderen met de regressie-analyse en het een schatting is, geven we dit aan met een dakje boven de y: $\hat{y}$. De regressie formule ziet er zo uit: $\hat{y}$ = a + b(x). Deze lijn zal de 'beste' lijn weergeven, in de zin dat deze het dichtste ligt bij alle datapunten. Dit wordt later toegelicht.

Effecten van outliers

Een regressie outlier (uitschieter) is een datapunt dat ver buiten de trend van de andere datapunten valt. Zo'n datapunt wordt invloedrijk genoemd wanneer het verwijderen ervan een grote verandering teweeg brengt in de regressieformule. Dit effect is kleiner bij een grote dataset. Het is soms beter om deze outliers te verwijderen.

Residuen

De regressieformule geeft een schatting van de y-waarden. Deze zullen niet helemaal overeenkomen met de daadwerkelijke (geobserveerde) y-waarden. Door het verschil tussen de geschatte waarden en de geobserveerde waarden te bekijken, kun je zien hoe ‘goed’ de regressielijn is. Het verschil tussen deze twee heet een residu. Het is het verschil tussen een geobserveerde waarde (y) en de voorspelde waarde (ŷ). Wanneer de geobserveerde waarde groter is dan de voorspelde waarde heb je een positief residu. Wanneer de geobserveerde waarde kleiner is dan de voorspelde waarde heb je een negatief residu. Hoe kleiner de absolute waarde van het residu, hoe beter de voorspelling, en dus de regressielijn.

Sum of Squared Errors/Residual Sum of Squares

De beste regressielijn is die met de kleinste residuen. Om die te vinden, worden de residuen van de datapunten gekwadrateerd en opgeteld. Dit noemen we de SSE. De SSE staat voor 'sum of squared errors'. De formule is $SSE = \sum (y - \hat{y})^2$. De beste regressielijn heeft de kleinste SSE van alle andere mogelijke lijnen. De SSE van de beste regressielijn heeft zowel negatieve als positieve residuen (die door het kwadrateren allemaal positief worden), waarvan samen de som en het gemiddelde 0 zijn. Daarbij loopt de regressielijn altijd door het punt van het gemiddelde van x en het gemiddelde van y, dus door het punt ($\bar{x}$, $\bar{y}$).

Het lineaire regressie model

Bij een regressieformule y = a + b(x) hoort bij elke x-waarde eenzelfde y-waarde. Dit heet een deterministisch model. Zo werkt het in de werkelijkheid niet. Stel bijvoorbeeld dat we inkomen (y) willen voorspellen aan de hand van opleidingsniveau (x), dan zien we dat niet iedereen met dezelfde opleiding ook hetzelfde inkomen heeft. In plaats van een deterministisch model kun je dan beter gebruik maken van een probabilistisch model. Deze ‘conditionele distributie’ refereert naar de variabiliteit in de y-waarden op een vaste waarde voor x.

Daarom veranderen we de formule nu naar E(y) = a + b(x). Hierbij geeft E(y) aan dat we kijken naar de conditionele distributie van y, en we proberen het gemiddelde van y te voorspellen.

Variatie op de regressielijn

Het lineaire regressiemodel kent nog een parameter, namelijk σ. Deze beschrijft de standaard afwijking van elke conditionele distributie. Het meet de variabiliteit van de y-waarden voor alle personen met die bepaalde x-waarde. We noemen σ de conditionele standaarddeviatie.

Omdat we deze echte standaardafwijking niet weten, gebruiken we wat we weten uit de steekproef, namelijk ‘s’. De formule van ‘s’ is $s = \sqrt{\frac{SSE}{n-2}}$, waarbij $SSE = \sum (y - \hat{y})^2$.

Als je deze ‘s’ kwadrateert heb je de zogenaamde ‘Mean Square Error’ of MSE.

Gestandaardiseerde regressie coefficient / Pearson correlatie

Nu gaan we kijken hoe sterk het eventuele verband is tussen x en y. We gebruiken daarvoor een gestandaardiseerde versie van correlatie en geven deze met 'r' aan. Deze ‘r’ wordt ook wel de gestandaardiseerde regressie coëfficiënt, of Pearson correlatie genoemd. Deze wordt berekend als volgt:

$$r = \left(\frac{S_x}{S_y}\right) b$$

Hierbij is S_x de steekproef deviatie van ‘x’ en S_y de steekproef deviatie van ‘y’. De formules van S_x en S_y, zijn als volgt:

$$S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \text{ en } S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

Deze correlatie heeft een aantal kenmerken:

- Je kunt de correlatie alleen gebruiken wanneer een lineair verband zinvol is.
- ‘r’ valt tussen 1 en -1.
- ‘r’ is positief/negatief gelijk aan ‘b’. Als ‘b’ positief is (en er een positief verband is) is ‘r’ ook positief en als ‘b’ negatief is (en er een negatief verband is) is ‘r’ ook negatief.
- Hoe groter ‘r’, hoe sterker het lineaire verband.

r-kwadraat

De 'coëfficiënt of determination' of r^2 is gerelateerd aan ‘r’ en geeft aan hoe goed y voorspeld kan worden door x. r-kwadraat heeft een aantal kenmerken die sterk overeenkomen met r:

- Omdat r tussen 1 en -1 valt, moet r^2 wel tussen 0 en 1 liggen.
- Als SSE = 0, dan r^2 = 1. Alle punten moeten op de lijn vallen.
- Als b = 0, r = 0, r^2 = 0.
- Hoe groter r^2, hoe sterker het lineaire verband.

Significantie toetsen bij regressie

t-toets voor onafhankelijkheid

Het principe bij deze test is hetzelfde als die bij de Chi-kwadraattoets, namelijk kijken of de variabelen onafhankelijk zijn. Je gaat ervan uit dat het zo is, dus je H0: β = 0 en Ha: $\beta \neq 0$. Bij het doen van deze test wordt er van uitgegaan dat er aan een aantal assumpties is voldaan:

- Randomisatie
- Het gemiddelde van y wordt benaderd door de formule E(y) = a + b (x)
- De conditionele standaard deviatie is gelijk voor elke x-waarde
- De conditionele distributie van y voor elke x-waarde is normaal verdeeld

Manier 1:

De t-score wordt berekend door b te delen door de standaardfout van b. De formule voor t is $t = \frac{b}{se}$.

De opbouw van deze formule is niet belangrijk. De vorm van de formule is gelijk aan die van elke t-score. Namelijk de schatting minus de nulhypothese (die hier 0 is, en dus gewoon verdwijnt), gedeeld door de standaardfout van de schatting.

Voor het opzoeken van de p-waarde gebruik je df = n – 2.

Manier 2:

Je kunt de t-score ook berekenen met deze formule:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}.$$

Er zal hetzelfde getal uitkomen als bij manier 1. Hiermee test je of de correlatie die is gevonden significant is, en daarmee test je dus ook of de variabelen onafhankelijk van elkaar zijn.

Betrouwbaarheidsinterval

Je kunt ook een betrouwbaarheidsinterval voor de hellingscoëfficiënt maken. Je doet dan de hellingscoëfficiënt plus en min de waarde van t, vermenigvuldigd met se.

$$\text{B.I. voor } \beta = b \pm t\,(\text{se})$$

Wat je invult voor t is afhankelijk van de precisie en zekerheid die je wilt gebruiken. Je moet dan wel opletten dat je voor df = n – 2 gebruikt.

Assumpties

Het gemiddelde van y kan worden benaderd met een lineair model

Het is belangrijk dat je altijd eerst een scatterplot maakt om te kijken of het wel zinvol is om een lineair model te maken voor jouw onderzoek. Als je dit niet doet, kun je het gevaar lopen een lineair verband te ontdekken in data die helemaal niet lineair is. Als je data bijvoorbeeld een U vorm heeft, kan SPSS alsnog een lineair verband ontdekken: de y verandert immers als de x verandert.

Niet extrapoleren

Het is niet verstandig om je regressieformule te gebruiken voor zelfbedachte datapunten buiten je dataset. Dit omdat het verband wellicht bij die datapunten helemaal niet meer lineair is, of omdat de schattingen van y dan niet realistisch zijn.

Outliers

Sommige outliers kunnen grote effecten hebben op de regressielijnen en de correlaties. Het is soms nodig dat je bepaalde outliers eruit haalt om nog een keer je analyses te lopen. Één punt kan al veel invloed hebben, in het bijzonder bij een kleine steekproef.

Steekproefgrootte

Bij een kleine steekproef heb je minder variatie in je x-waarden dan bij een grote steekproef. Je loopt dan het risico dat je niet een realistische dataset hebt verkregen, en dat je analyses niet veel waard zijn. Je kunt correlaties het beste gebruiken bij datasets die willekeurig verkregen zijn, en representatief zijn wat betreft de variatie in de x- en y-waarden.

Tabel: Onafhankelijkheidstesten en associatiematen

	Meetniveau

	Nominaal	**Ordinaal**	**Interval**
Nulhypothese	H0: onafhankelijkheid	H0: onafhankelijkheid	H0: onafhankelijkheid (b = 0)
Test statistiek	$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$z = \frac{\hat{y}}{se}$	$t = \frac{b}{se}$, $df = n - 2$
Meting samenhang	$\hat{\pi}_2 - \hat{\pi}_1$ Odds ratio	$\hat{\gamma} = \frac{C - D}{C + D}$	$r = b(\frac{s_x}{s_y})$