

## Hoofdstuk 6: Introductie in statistische gevolgtrekkingen

### Inleiding

Statistische gevolgtrekkingen (statistical inference) gaan over het trekken van conclusies over een populatie op basis van steekproefdata. In dit deel zullen in dit verband vooral betrouwbaarheidsintervallen en significantietesten aan de orde komen. Ook gaat het in dit deel alleen over het trekken van statistische gevolgtrekkingen over de populatie als de standaarddeviatie ( $\sigma$ ) van de populatie bekend is.

### Betrouwbaarheidsintervallen

Bij het berekenen van betrouwbaarheidsintervallen proberen we met een bepaalde zekerheid (bijvoorbeeld met een zekerheid van 95%) te stellen dat een populatiewaarde zich tussen twee grenswaarden bevindt. Bij significantietesten is het doel uitzoeken of een bepaalde uitkomst hoogstwaarschijnlijk het gevolg van toeval of van een echt effect is. Als we een therapiemethode voor depressie onderzoeken willen we bijvoorbeeld weten of deze therapie echt effect heeft gehad of dat de vooruitgang bij de deelnemers zo klein is dat er geen echt effect waarneembaar is. Omdat statistische gevolgtrekkingen op steekproevendistributies gebaseerd zijn, wordt vaak gebruik gemaakt van een kansdistributie. We doen bij statistische gevolgtrekkingen alsof de verzamelde data afkomstig is van een random steekproef of een gerandomiseerd experiment.

### 6.1 Schatten met betrouwbaarheid

#### Het populatie- en steekproefgemiddelde

We weten inmiddels dat  $\bar{x}$  een goede schatter is van  $\mu$ , maar hoe precies is de schatting? Je kunt bij de eerste steekproef bijvoorbeeld een gemiddelde van 100 vinden, maar hoogstwaarschijnlijk is dit bij de tweede steekproef niet precies hetzelfde. Om te weten hoe precies onze schatting is, hebben we ook een schatting van de spreiding nodig. Als er weinig spreiding is, dan weten we dat gemiddelden van verschillende steekproeven waarschijnlijk erg dicht bij elkaar liggen. Stel: we hebben een populatie met een standaarddeviatie van 4.5. In dat geval:

- Zegt de 68-95-99.7 regel dat er ongeveer een kans van 0.95 is dat  $\bar{x}$  negen punten (twee standaarddeviaties) van het populatiegemiddelde aflight.
- Gebruiken we het steekproefgemiddelde om het populatiegemiddelde te berekenen en niet andersom. Daarom zeggen we ook wel dat er 0.95 kans is dat het populatiegemiddelde negen punten rond  $\bar{x}$  varieert.
- Betekent dit dat 95% van alle steekproeven de echte  $\mu$  zullen bevatten in het interval van  $\bar{x}-9$  tot  $\bar{x}+9$ . Stel je voor dat onze eigen steekproef een gemiddelde ( $\bar{x}$ ) van 461 geeft. We zeggen dan dat we er 95% zeker (*confident*) van zijn dat het onbekende populatiegemiddelde ( $\mu$ ) tussen  $461-9 = 452$  en  $461+9 = 470$  ligt. Er is echter ook nog 5% kans dat het interval met de grenswaarden 452 en 470 de ware  $\mu$  niet bevat.

### Betrouwbaarheidsintervallen

Het interval 452-470 in het bovenstaande voorbeeld wordt het *95% betrouwbaarheidsinterval voor  $\mu$*  genoemd. De meeste betrouwbaarheidsintervallen hebben de vorm van schatting  $\pm$  foutenmarge. De schatting ( $\bar{x}$  in ons geval) gebruiken we om een onbekende parameter te schatten. De foutenmarge (9 in ons voorbeeld) laat zien hoe zeker we ervan zijn dat onze

schatting van de parameter klopt op basis van de grenswaarden. Betrouwbaarheidsintervallen gaan gepaard met twee belangrijke feiten:

- Betrouwbaarheidsintervallen hebben de vorm van (a,b) waarbij a en b getallen zijn die door middel van de data berekend worden.
- Een betrouwbaarheidsinterval gaat samen met een betrouwbaarheidsniveau (bijvoorbeeld 90 of 95), dat ons duidelijk maakt wat de kans is dat het interval de ware parameter zal bevatten. Als we een betrouwbaarheidsinterval van 90% gebruiken, zeggen we dus eigenlijk dat we er 90% zeker van zijn dat de ware parameter zich zal bevinden tussen de door ons uitgerekenende grenswaarden. In de praktijk wordt het meest gebruik gemaakt van 95% als betrouwbaarheidsniveau; 90% en 99% komen minder vaak voor. Een *betrouwbaarheidsinterval* korten we af met de letter C (van confidence interval).

### **Betrouwbaarheidsinterval voor een populatiegemiddelde**

Een betrouwbaarheidsniveau gaat samen met een z-waarde. Een bijbehorende z-waarde kan altijd gevonden worden met tabel D achterin het boek. Een 95% betrouwbaarheidsinterval gaat bijvoorbeeld samen met een z van 1.96. Er is een kans van 95% dat  $\bar{x}$  tussen  $\mu - z^*(\sigma/\sqrt{n})$  en  $\mu + z^*(\sigma/\sqrt{n})$  ligt. Dit is precies hetzelfde als zeggen dat het onbekende populatiegemiddelde  $\mu$  tussen  $\bar{x} - z^*(\sigma/\sqrt{n})$  en  $\bar{x} + z^*(\sigma/\sqrt{n})$  ligt. De schatter van de onbekende  $\mu$  is  $\bar{x}$  en de *foutenmarge* m is:  $z^*(\sigma/\sqrt{n})$ . Een korte samenvatting van deze informatie volgt hieronder.

- Als een SRS van grootte n uit een populatie met een onbekende  $\mu$  en bekende  $\sigma$  wordt geselecteerd, dan is de foutenmarge dus:  $m = z^*(\sigma/\sqrt{n})$ .
- In de bovenstaande formule staat  $z^*$  voor de waarde van de normaalverdeelde curve met gebied C (dus bijvoorbeeld 95%) tussen de grenswaarden  $-z^*$  en  $z^*$ . Het niveau C betrouwbaarheidsinterval voor  $\mu$  is  $\bar{x} \pm m$ . Dit interval is een precieze schatter wanneer de populatie normaalverdeeld is en is een bijna precieze schatter in andere gevallen, maar n moet dan wel groot zijn.

### **Feiten over betrouwbaarheidsintervallen**

Een hoge betrouwbaarheid is altijd gewenst, maar een klein foutenmarge ook. Een hoge betrouwbaarheid (confidence) zegt dat onze methode bijna altijd juiste antwoorden geeft. Een klein foutenmarge zegt dat we de parameter behoorlijk zeker kunnen schatten. Als een onderzoeker besluit dat de foutenmarge te groot is, dan kan hij of zij drie dingen doen:

- Een kleiner betrouwbaarheidsniveau (C) gebruiken.  $Z^*$  zal kleiner worden wanneer we een kleiner betrouwbaarheidsniveau gebruiken. Daarom zal een kleinere  $z^*$  leiden tot een kleinere foutenmarge (maar alleen als niets met n en  $\sigma$  gebeurd is).
- De steekproefgrootte n laten toenemen. De foutenmarge wordt dan kleiner voor welk betrouwbaarheidsniveau dan ook.
- De standaarddeviatie ( $\sigma$ ) verkleinen. Dit kan niet altijd, omdat we populatiegegevens niet makkelijk kunnen veranderen.

### **Een steekproefgrootte kiezen**

Een onderzoeker kan ook van tevoren vaststellen wat de foutenmarge mag zijn. Op basis van dat gegeven kan hij of zij de steekproefgrootte n bepalen. Dit kan met de volgende formule:

- $n=(z^*\sigma/m)^2$ . Uit deze formule volgt dat de *steekproef* de foutenmarge bepaalt. De grootte van de *populatie* heeft geen invloed op de steekproefgrootte die we nodig hebben.

### Voorzichtigheid

Met de gegeven formule voor betrouwbaarheidsintervallen,  $\bar{x} \pm z^* \sigma / \sqrt{n}$ , gaan een aantal waarschuwingen gepaard.

- De data moeten het resultaat zijn van het trekken van een SRS uit de populatie. Het beste is om een gerandomiseerde SRS te trekken, maar het is voor de betrouwbaarheid van de onderzoeksresultaten ook toereikend als we kunnen aannemen dat de data het resultaat is van onafhankelijke observaties uit de populatie.
- De formule is niet te gebruiken wanneer de steekproef niet op basis van een SRS getrokken is.
- Data moet goed verzameld worden, zodat er geen sprake is van bias of onbekende steekproefgrootte. Als dat wel het geval is, dan geeft de bovenstaande formule geen betrouwbare resultaten, simpelweg omdat de data zelf ook niet betrouwbaar is.
- Omdat  $\bar{x}$  niet robuust is, hebben uitbijters een groot effect op de betrouwbaarheidsinterval.
- Als de steekproefgrootte klein is en de populatie niet normaal verdeeld is, dan zal het ware betrouwbaarheidsniveau anders zijn dan de waarde C die gebruikt is om het interval te berekenen. Wanneer  $n \geq 15$  is, dan zal het betrouwbaarheidsniveau niet erg beïnvloed worden door het feit dat de populatie niet normaal verdeeld is. Een grote steekproefgrootte is dus gewenst als de populatie niet normaal verdeeld is.
- De formule gaat er vanuit dat de standaarddeviatie van de populatie ( $\sigma$ ) bekend is. In de praktijk is dit echter bijna nooit het geval. In het volgende deel zal aan de orde komen hoe betrouwbaarheidsintervallen berekend moeten worden als de standaarddeviatie van de populatie onbekend is. Als de steekproef groot is, dan zal de standaarddeviatie van de steekproef ( $s$ ) dichtbij de onbekende  $\sigma$  liggen. Het interval  $\bar{x} \pm z^* s / \sqrt{n}$  is dan een benadering van het betrouwbaarheidsinterval van  $\mu$ .

## 6.2 Het toetsen van significantie

### De redenering achter het testen van significantie

Een significantietoets voeren we uit om geobserveerde data te vergelijken met een vooraf opgestelde hypothese waarvan we de juistheid willen toetsen. Een hypothese is een statement over populatiegegevens (parameters). De uitkomsten van een significantietoets worden weergegeven in de vorm van kansen. We kunnen op basis van de uitkomsten van een significantietoets berekenen hoe groot de kans is dat de gevonden resultaten het gevolg zijn van toeval.

### Hypothesen

De eerste stap bij het toetsen van significantie is het bedenken van een stelling waar we bewijs *tegen* hopen te vinden.

- De hypothese die bij een significantietoets getoetst wordt, wordt de *nulhypothese* ( $H_0$ ) genoemd. De significantietoets gaat over hoe sterk het bewijs tegen de nulhypothese is. In de meeste gevallen is de nulhypothese een statement in termen van 'geen effect' of 'geen verschil'. Een (nul)hypothese wordt altijd in parameters genoteerd.

- Daarnaast wordt er een *alternatieve hypothese* ( $H_a$ ) geformuleerd. Deze stelt dat er wel een verschil of verandering is. We willen bewijs vinden dat de alternatieve hypothese *steunt*. Vaak beginnen onderzoeken met het formuleren van deze alternatieve hypothese. Vervolgens formuleren ze de hypothese waarvan ze hopen dat deze niet klopt (de nulhypothese).
- De alternatieve hypothese  $H_a$  kan *eenzijdig* of *tweezijdig* zijn. Een alternatieve hypothese is tweezijdig wanneer een onderzoeker geen vermoeden heeft over de richting van een effect. Het is verkeerd om eerst naar de data te kijken en daarna een alternatieve hypothese te formuleren die bij de data past. Als je geen idee hebt van de richting van een mogelijk effect, dan is het goed om tweezijdig te toetsen. Als je als onderzoeker wel een richting vermoedt (bijvoorbeeld dat een therapievorm depressie *vermindert*) dan is het wel geoorloofd om eenzijdig te toetsen.

### Teststatistieken

Een significantietoets is gebaseerd op een statistiek die een parameter schat. Deze parameter is in de nulhypothese weergegeven. Wanneer de nulhypothese waar is, verwachten we dat deze schatting een waarde aanneemt die dicht bij de parameter uit de nulhypothese ligt. Schattingen van de parameter die ver van de nulhypothese liggen, geven juist bewijs *tegen* de nulhypothese. Om uit te zoeken hoe ver de schatting van de echte parameter verwijderd is, is het van belang om de schatting te standaardiseren. In de meeste gevallen heeft de teststatistiek de volgende vorm:

- $Z = (\text{schatting} - \text{waarde uit de hypothese}) / \text{standaarddeviatie van de schatting}$ .
- Een *teststatistiek* meet in hoeverre de nulhypothese en de verzamelde data overeenkomstig zijn. Dit gegeven gebruiken we voor de significantietoets. Een teststatistiek zien we als een random variabele.

### P-waarden

Een significantietoets laat zien wat de kans is dat een bepaald resultaat (of een extremer resultaat) gevonden wordt. 'Extreem' is in dit geval 'ver van wat we zouden verwachten als de nulhypothese waar zou zijn'.

- De *p-waarde* is de kans dat een teststatistiek een extreme(re) waarde aanneemt dan uit de gevonden data blijkt, als de nulhypothese waar zou zijn. Hoe kleiner de p-waarde, hoe sterker het bewijs tegen de nulhypothese. De p-waarde wordt berekend aan de hand van de steekproevendistributie van de teststatistiek.

### Statistische significantie

We kunnen de berekende p-waarde vergelijken met een vaste waarde waarvan we besloten hebben dat deze beslissend is. Deze waarde geeft van tevoren aan hoeveel bewijs tegen de nulhypothese nodig is om deze te kunnen verwerpen.

### Significantieniveau:

De beslissende waarde van  $p$  wordt het *significantieniveau* genoemd. Deze korten we af met  $\alpha$ . Als we  $\alpha=0.05$  gebruiken, dan willen we dat data zo'n bewijs tegen de nulhypothese levert dat er maar 5% kans is dat de gevonden resultaten het gevolg zijn van toeval en niet van een echt effect. Bij een alfa ( $\alpha$ ) van 0.01 willen we nog sterker bewijs om de nulhypothese te verwerpen dan bij een alfa van 0.05. Als een resultaat statistisch significant blijkt te zijn, dan zie je in wetenschappelijke literatuur bijvoorbeeld  $P<0.01$  of  $P<0.05$  staan. ( $P$  is de p-waarde).

### P-waarden en statistische significantie:

We hebben meer aan p-waarden dan aan het feit of iets statistisch significant is gebleken, omdat we resultaten zelf significant kunnen maken door grotere p-waarden als beslissend te bestempelen. Een resultaat van  $P=0.03$  is bijvoorbeeld significant op  $\alpha=0.05$  niveau, maar niet op  $\alpha=0.01$  niveau.

- Als de p-waarde even klein of kleiner dan  $\alpha$  is, dan zeggen we dat de data *statistisch significant op het niveau van  $\alpha$*  is. Als er een tweezijdige significantietoets uitgevoerd wordt, moet de gevonden p-waarde maal twee gedaan worden om te kijken of deze kleiner dan  $\alpha$  is. Bij eenzijdige toetsen hoeft dit niet.

### Het uitvoeren van een significantietoets

Het uitvoeren van een significantietoets gaat door middel van vier stappen.

1. Formuleren van de nulhypothese en de alternatieve hypothese.
2. Berekenen van de waarde van de teststatistiek.
3. P-waarden vinden die bij de geobserveerde data horen.
4. Trekken van een conclusie door een significantieniveau  $\alpha$  vast te stellen.  
Deze bepaalt hoeveel bewijs er tegen de nulhypothese nodig is om deze te verwerpen. Als de p-waarde kleiner of gelijk is aan  $\alpha$ , dan moet geconcludeerd worden dat de alternatieve hypothese klopt. Als de p-waarde groter dan  $\alpha$  is, dan moet geconcludeerd worden dat de gevonden data niet genoeg bewijs levert om de nulhypothese te verwerpen.

### Het toetsen van het populatiegemiddelde

Voor een toets van het populatiegemiddelde  $\mu$  is de nulhypothese: het ware populatiegemiddelde is gelijk aan  $\mu_0$ . Dus:

- $H_0: \mu = \mu_0$ . In dit verband heeft  $\mu_0$  een specifieke waarde die we willen onderzoeken. De significantietoets is gebaseerd op een schatter van de parameter, dus op het steekproefgemiddelde ( $\bar{x}$ ). Onze teststatistiek meet het verschil tussen de steekproefschatting en de parameter uit de nulhypothese in termen van standaarddeviaties van de teststatistiek:
- $Z = (\text{schatting} - \text{waarde uit de hypothese}) / \text{standaarddeviatie van de schatting}$ .
- De standaarddeviatie van  $\bar{x}$  is  $\sigma / \sqrt{n}$ . Daarom is de teststatistiek:
- $z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$ . In dit geval is de standaarddeviatie van de populatie dus bekend.

## 6.3 Gebruik en misbruik van testen

### Voorzichtigheid

Een significantietoets uitvoeren is vaak gemakkelijk en tegenwoordig worden hiervoor vaak computerprogramma's gebruikt. Het gebruik van een significantietoets is echter niet altijd even gemakkelijk.

- Het is in wetenschappelijke literatuur vaak normaal om de gevonden p-waarde te noteren en erbij te zetten of de resultaten significant zijn gebleken. *Er is echter geen scherpe lijn tussen significant en niet-significant te trekken*. Of data als significant of niet-significant beoordeeld worden, hangt samen met de  $\alpha$  die van tevoren gekozen is.

- Als de nulhypothese wordt verworpen, betekent dit dat er sprake is van een effect en dat de onderzoeksresultaten hoogstwaarschijnlijk niet het gevolg zijn van toevalsverschijnselen. Dit zegt echter helemaal niets over hoe groot het effect is. Er kan dus sprake zijn van een zeer klein effect, maar ook van een groot effect. *Als er grote steekproeven worden getrokken, dan zijn kleine afwijkingen van de nulhypothese al snel significant.*
- In de praktijk worden betrouwbaarheidsintervallen te weinig gebruikt, terwijl significantietoetsen juist te vaak worden uitgevoerd.
- Het hoeft niet per se zo te zijn dat niet-significante resultaten betekenisloos zijn. Soms zijn resultaten net niet significant. Dit zegt ook weer iets.
- Als je een onderzoek wilt uitvoeren is het van belang om een toetsmethode te gebruiken waarvan je zeker weet dat deze een effect kan vaststellen als deze ook daadwerkelijk aanwezig is in de data.
- Onderzoekers moeten oppassen dat ze objectief blijven, want vaak willen ze een effect vinden. Dit omdat ze bijvoorbeeld willen bewijzen dat hun therapiemethode voor depressie effectief is.

## 6.4 Statistische power

### Power

Als we een  $\alpha$  van 5% bij een significantietoets gebruiken, zijn er we er 95% zeker van dat als de nulhypothese in werkelijkheid verkeerd is, dat we dat dan ook echt zullen vinden.

- De kans dat een significantietoets met een vaste  $\alpha$  de nulhypothese zal afwijzen als de alternatieve hypothese in werkelijkheid juist is, noemen we de *power* van de toets.

### Stappen om de power te berekenen

Het berekenen van de power van een test gaat in drie stappen:

1. Formuleren van de nulhypothese en de alternatieve hypothese.
2. Vinden van de waarden van  $\bar{x}$  die leiden tot het verwerpen van de nulhypothese.
3. Kans berekenen dat de waarden van  $\bar{x}$  gevonden zullen worden als de alternatieve hypothese waar is. Zie voor voorbeelden op bladzijde 386 en 387.

### De power verhogen

Stel je voor dat je als onderzoeker ontdekt dat de power van je toets te klein is. Wat kun je dan doen?

- Het is mogelijk om  $\alpha$  te verhogen.
- Het is ook mogelijk om een alternatieve hypothese te formuleren die verder van de waarde van de nulhypothese ligt. Waarden van  $\mu$  in de alternatieve hypothese die dichtbij de waarden van de nulhypothese liggen zijn moeilijker te bewijzen dan waarden die verder van de nulhypothese liggen.
- Daarnaast is het aan te raden om de steekproefgrootte toe te laten nemen. Meer data zorgt ervoor dat er meer informatie over  $\bar{x}$  beschikbaar is. Dit zorgt er weer voor dat er een grotere kans is dat we onderscheid kunnen maken tussen waarden van  $\mu$ .
- Tot slot kan  $\sigma$  verlaagd worden. Dit heeft hetzelfde effect als een stijging in de steekproefgrootte. De  $\sigma$  kan op twee manieren verlaagd worden: (1) door het metingsproces te verbeteren en (2) door je als onderzoeker te beperken tot een subpopulatie.

### **Twee soorten fouten**

Zelfs significantietoetsen geven niet honderd procent de garantie dat de juiste conclusie over hypothesen worden getrokken. Er kunnen twee soorten fouten gemaakt worden:

- Een *type I fout*: Afwijzen van de nulhypothese en het aannemen van de alternatieve hypothese. Dit terwijl in werkelijkheid de nulhypothese juist is.
- Een *type II fout*: Aannemen van de nulhypothese en afwijzen van de alternatieve hypothese. Dit terwijl de alternatieve hypothese in werkelijkheid juist is.

Het is niet mogelijk om beide fouten tegelijkertijd te maken. Het significantieniveau  $\alpha$  van een significantietoets is de kans op een Type I fout. Bij een  $\alpha$  van 5% is er dus 5% kans dat we de nulhypothese ten onrechte verwerpen. De power van een significantietoets met een vast significantieniveau  $\alpha$  om de alternatieve hypothese te detecteren is  $1 -$  de kans op een Type-II fout. In de praktijk worden type-I fouten erger gevonden, omdat er dan wordt gedacht dat er een effect is terwijl dat in werkelijkheid helemaal niet het geval is.