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 $t \geq 0{,}68$ of $t \leq -0{,}68$. 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.

- 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.
- 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.
- 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	