Bijlage 3

Statistische toetsing: werkwijze, toetsen, formules, toepassing

In dit boek wordt kennis van statistiek en statistische (hypothese)toetsing in principe bekend verondersteld. Niettemin geven we hier – ter opfrissing – een aantal belangrijke punten kort weer. Wie zich opnieuw (of alsnog) in de statistiek wil verdiepen wordt het boek *Cijfers spreken* (5e druk, Noordhoff Uitgevers, 2011) van Joep Brinkman aanbevolen. Die uitgave behandelt met dezelfde symbolen, formules, principes en uitgangspunten de methodologie en de statistiek waar in *Proeven van succes* van wordt uitgegaan.

1. Statistische (hypothese)toetsing in hoofdlijnen

Zes stappen

In zes stappen uitgedrukt is de algemene wijze van denken en doen bij statistische (hypothese)toetsing als volgt:

1 Je wilt nagaan of (of bewijzen dat) er meer speelt dan alleen maar toeval.
2 Ga er vervolgens van uit dat dat juist níet het geval is. Met andere woorden: *ga ervan uit dat er alleen maar toeval speelt*.
3 Kies het significantieniveau α (doorgaans 1% of 5%). Dit is het risico *ten onrechte* te concluderen dat er meer dan alleen toeval speelt. Als je dit risico wilt beperken kies je een lage α (en omgekeerd).
4 Verzamel gegevens. Bereken met behulp van een toepasselijke theoretische kansverdeling hoe groot de kans is op een zo extreme of nog extremere waarde als in deze 'steekproef' is gevonden. Deze kans is p, de *overschrijdingskans*. Hierbij ga je er steeds vanuit dat stap 2 waar is en er dus louter toevalsomstandigheden zijn.
5 Vergelijk deze overschrijdingskans met α. Als p kleiner is dan α, is het gestelde bij stap 2 (alléén toeval) waarschijnlijk niet waar. In dat geval spreek je van *significantie* en wordt stap 6 gezet. Als p groter is dan α, ga er dan voorlopig vanuit dat het gestelde bij stap 2 waar is, en er dus alleen maar toeval heeft gespeeld.
6 Als het gestelde bij stap 2 onjuist is, moet het tegendeel waar zijn. Dan is de gevonden uitkomst niet alleen door het toeval ontstaan en speelt er meer. Dan is het vermoeden van stap 1 dus juist.

Hypothesen

De tweede stap, waarin wordt aangenomen dat er alleen maar toeval speelt, verdient bijzondere aandacht. Op die aanname is namelijk het verdere rekenwerk gebaseerd. Deze aanname heet de *nulhypothese*, H_0. De nulhypothese heeft altijd de vorm 'er is niets aan de hand, er speelt alleen maar toeval'. Het toetsen vindt steeds plaats door met betrekking tot een soort 'nul'situatie de theoretische kansverdeling (zoals een binomiale verdeling of een t-verdeling) te bepalen. *Toegepast op sensorisch onderzoek*, komt de nulhypothese er meestal op neer dat er geen smaakverschil is, dat er geen verschil in voorkeur is, dat product A even scherp (of zout of hard of…) is

als product B, dat de gemiddelde score op een bipolaire lijnschaal gelijk is aan 50 enzovoort.

Tegenover de nulhypothese staat de *alternatieve hypothese*, H_1. Deze staat bij de toetsingsprocedure hierboven verwoord bij stap 1. In de alternatieve hypothese komt doorgaans tot uitdrukking wat de onderzoeker verwacht, hoopt of wil aantonen. H_1 heeft de vorm 'als de nulhypothese niet opgaat, dan moet er meer aan de hand zijn dan alleen toeval, en moet dus het volgende waar zijn: ...'.

H_0 en H_1 hebben *altijd betrekking op de populatie*. Vandaar dat er doorgaans Griekse symbolen in worden gebruikt.

De nulhypothese en de alternatieve hypothese *moeten elkaar volledig aanvullen en uitsluiten*. Dus als bijvoorbeeld in H_0 het is-gelijk-teken (=) staat, bevat H_1 het is-ongelijk-teken (≠). Als in de nulhypothese ≥ staat, moet de alternatieve hypothese < bevatten. Enzovoort.

Nulhypothesen worden verworpen of niet, alternatieve hypothesen worden geaccepteerd of niet. Dit taalgebruik maakt duidelijk dat nulhypothesen niet kunnen worden bewezen. Nulhypothesen krijgen het voordeel van de twijfel: zij worden gehandhaafd bij gebrek aan bewijs van het tegendeel.

Een- en tweezijdig toetsen

De vraag of een toets een- of tweezijdig moet worden uitgevoerd, hangt af van de onderzoeksvraag. Als men wil weten óf er een verschil met het toeval bestaat, is de toets tweezijdig. Wie wil weten of er een verschil *in een bepaalde richting* (meer *of* minder kans dan door toeval) bestaat, toetst eenzijdig.
Bij een eenzijdige toets kijkt men naar de linker óf naar de rechter overschrijdingskans. Men noemt de toets dan ook respectievelijk links- of rechtseenzijdig. Je herkent een eenzijdige toets aan de 'ongelijk'-tekens in de hypothesen. De hypothesen van een tweezijdige toets bevatten daarentegen 'is-gelijk'- en 'is-ongelijk'-tekens.

Wie tweezijdig toetst, moet de tweezijdige overschrijdingskans bepalen. Dat is echter meestal lastig of onmogelijk. In de praktijk werkt men meestal omgekeerd en *vergelijkt men bij een tweezijdige toets de eenzijdige overschrijdingskans met ½α*.

Toetsingsgrootheden en kritieke gebieden

Voor het uitvoeren van een toets moet de overschrijdingskans van een *toetsingsgrootheid* worden bepaald (zoals k of z voor de binomiaaltoets, χ^2 voor de χ^2-toets of F voor een variantieanalyse). Voor z en k kan dat met behulp van de tabellen B of C uit het boek. Zulke tabellen zijn echter gauw erg omslachtig en omvangrijk. Daarom bestaan er ook tabellen met alleen de waarden van waaraf de toetsingsgrootheid significant is. Men spreekt van *kritieke waarden*. De tabellen D tot en met J zijn volgens dat principe uitgevoerd. Als een toetsingsgrootheid *op of voorbij* de kritieke waarde komt, luidt de toetsuitslag 'significant' en wordt de nulhypothese dus verworpen. Zodra een in de steekproef gevonden toetsingsgrootheid de kritieke waarde evenaart, is de kans daarop kleiner dan α. Men zegt dan dat de toetsingsgrootheid in het *kritieke gebied* ligt.

Voor het uitvoeren van een toets kunnen dus twee wegen worden bewandeld, die overigens op hetzelfde neerkomen:

1 het berekenen van een overschrijdings*kans*, die wordt vergeleken met de gekozen *kans* α;
2 het vaststellen van de kritieke *waarde* van de toetsingsgrootheid die hoort bij de gekozen α; vervolgens nagaan of de *toetsingsgrootheid* (berekend) uit de steekproef deze kritieke waarde evenaart.

De toetsuitslag is voor beide wegen uiteraard hetzelfde. Toch blijkt het in de praktijk verwarring te wekken. De eerste methode dwingt je er altijd toe de nulhypothese te verwerpen als de gevonden *kans* kléiner is dan α. De tweede leidt er vaak toe, dat je de nulhypothese verwerpt als de *toetsingsgrootheid* gróter is dan de kritieke waarde. Computerprogramma's als SPSS werken door overschrijdingskansen te geven met de eerstgenoemde methode.

2. De binomiaaltoets

- Wordt *toegepast* in situaties waarin herhaalde *gelijke* kansprocessen elk slechts twee mogelijke uitkomsten hebben (zoals: kruis/munt, goed/fout, monster A/monster B).
- Toepassing in hier behandeld sensorisch onderzoek: paarsgewijze vergelijking, driehoekstest, 3-AFC-test, tetrad-test, duo-triotest en twee-uit-vijftest.
- Hierbij geeft n het aantal proefpersonen/tests aan en π de kans per keer op een bepaalde uitkomst (meestal de gokkans bij het niet waarnemen van een verschil). De toetsingsgrootheid is k, dat is het aantal keren dat een bepaalde uitkomst is opgetreden.
- Tabel D en tabel E tonen de kritieke waarden van k voor respectievelijk $\pi = ½$ en $\pi = ⅓$ voor diverse waarden van n.
- In situaties waarin de tabellen D en E niet voorzien, kunnen voor enkele waarden van π en n in tabel C eenzijdige linker overschrijdingskansen worden opgezocht. (Bij die tabel staat aangegeven hoe daaruit rechter overschrijdingskansen kunnen worden afgeleid. Voor tweezijdige toetsing wordt α gehalveerd.)
- In situaties waarin de tabellen C, D en E niet voorzien, kunnen eenzijdige overschrijdingskansen via de *normale verdeling* worden benaderd (mits zowel nπ als n(1–π) minstens 10 bedraagt). Daarvoor wordt z berekend met de formule:

$$z = \frac{k \pm 1/2 - n\pi}{\sqrt{n\pi(1-\pi)}}$$

De zogenoemde continuïteitscorrectie van ½ in de teller is negatief voor de rechter overschrijdingskans en positief voor de linker. De rechter overschrijdingskans van de berekende z kan worden opgezocht in tabel B, de linker kan daarvan worden afgeleid. Voor tweezijdige toetsing wordt α gehalveerd.

Voorbeelden:

- *Driehoekstest (dus π = 1/3) met (n =) 40 panelleden en k = 22 (goede antwoorden).* Toetsing met α = 5% (langs elk van de drie genoemde wegen):
  - Tabel E laat zien dat k = 19 genoeg is om de nulhypothese te verwerpen. Met k = 22 is er dus een significant verschil.
  - Uit tabel C (bij n = 40 en π = 1/3) valt af te leiden dat de rechter overschrijdingskans 100% – 99,61% = 0,39% bedraagt. Deze is kleiner dan α, dus wordt de nulhypothese verworpen.
  - De normale benadering mag zo nodig worden toegepast. Invullen van de formule levert op: $z = (22–1/2–40/3)/\sqrt{(40×1/3×2/3)} = +2{,}74$. Tabel B laat zien dat hier een overschrijdingskans bij hoort van 0,31%. Deze is kleiner dan α, dus wordt de nulhypothese verworpen. (Merk het verschil op van deze benadering met de exacte werkwijze volgens tabel C.)
- *Paarsgewijze vergelijking (dus π = 1/2) met (n =) 76 panelleden en k = 45.* Tweezijdige toetsing met α = 1%.
  Tabel D noch tabel C voorziet in n = 76. De normale benadering kan en mag echter worden toegepast. $z = (45–1/2–76/2)/\sqrt{(76×1/2×1/2)} = +1{,}49$. Tabel B laat zien dat hier een rechter overschrijdingskans bij hoort van 6,81%. Deze is groter dan ½α, dus de nulhypothese wordt niet verworpen. Er kan geen verschil worden aangetoond.

3. De χ^2-toets (chi-kwadraat-toets)

- Wordt – als het om sensorisch onderzoek gaat – *toegepast* in situaties waarin gegevens bestaan uit *absolute frequenties* die zijn ondergebracht in een 'kruistabel'. Het gaat dan om het verband tussen twee variabelen die beide op nominaal niveau zijn gemeten.
- Toepassing in hier behandeld sensorisch onderzoek: A/niet-A-test en eenvoudige verschiltest. Hiervan worden de uitkomsten in 2-bij-2-tabellen ondergebracht.
- De gegevens staan in een kruistabel met werkelijk gevonden frequenties W. Daarnaast wordt een even grote tabel met verwachte frequenties V opgezet. Deze V's zijn de (niet op hele getallen afgeronde) frequenties die de 'eerlijkste' verdeling van de variabelen over elkaar weergeven, ervan uitgaande dat de nulhypothese waar is en er dus geen enkel verband tussen de variabelen bestaat.
- De V's worden berekend aan de hand van de *randtotalen* van de tabel met W's: de V van elke cel bereken je door het bijbehorende kolomtotaal te vermenigvuldigen met het bijbehorende rijtotaal, en dit product te delen door het algemene totaal:

Werkelijke frequenties W_1 t/m W_4:		
W_1	W_2	$A = W_1 + W_2$
W_3	W_4	$B = W_3 + W_4$
$C = W_1 + W_3$	$D = W_2 + W_4$	$n = A + B\ (= C + D)$

Verwachte frequenties V_1 t/m V_4:		
$V_1 = (A×C)/n$	$V_2 = (A×D)/n$	A
$V_3 = (B×C)/n$	$V_4 = (B×D)/n$	B
C	D	n

- De omvang van een kruistabel wordt getypeerd door het aantal zogeheten *vrijheidsgraden (degrees of freedom), afgekort tot df*. Df is het aantal kolommen min 1 vermenigvuldigd met het aantal rijen min 1. Voor een 2-bij-2-tabel geldt dus df = 1.
- Voor df = 1 is voor de berekening van χ^2 een *continuïteitscorrectie* nodig. Hiervoor wordt elke W met ½ in de richting van de bijbehorende V gebracht. Als W groter is dan V wordt W dus met ½ verminderd, als W kleiner is wordt er ½ bij opgeteld.
- De toetsingsgrootheid χ^2 wordt dan berekend met de formule:

$$\chi^2 = \sum \frac{(W-V)^2}{V}$$

- Tabel F toont de kritieke waarden van χ^2 voor verschillende waarden van df.
- χ^2 kan nooit negatief zijn. Voor een eenzijdige toets moet daarom nog goed naar de kruistabel worden gekeken om na te gaan of een significante waarde van χ^2 wel overeenkomt met de in H_1 geformuleerde richting van het verband.
- Voorbeeld: de A/niet-A-test van paragraaf 6.7

Werkelijke frequenties W:			
Gegeven:	**Antwoord panellid:**		
	'M'	**'Niet M'**	**Totaal**
merk M	15	15	30
merk P	9	21	30
Totaal	24	36	60

Verwachte frequenties V:		
Antwoord panellid:		
'M'	**'Niet M'**	**Totaal**
12*	18	30
12	18	30
24	36	60

Gecorrigeerde W's:		
Antwoord panellid:		
'M'	**'Niet M'**	**Totaal**
14,5	15,5	30
9,5	20,5	30
24	36	60

* = 30×24/60 enzovoort

$\chi^2 = (14{,}5-12)^2/12 + (15{,}5-18)^2/18 + (9{,}5-12)^2/12 + (20{,}5-18)^2/18 = 1{,}74.$

Tabel F toont voor eenzijdige toetsing (df = 1) voor $\alpha = 1\%$ en $\alpha = 5\%$ achtereenvolgens 5,41 en 2,71 als kritieke waarden voor χ^2. De gevonden χ^2 is kleiner, de resultaten zijn dus niet significant (afwijkend van het toeval), de nulhypothese blijft staan, er kan niet worden aangetoond dat liefhebbers van pilsmerk M pils van hun merk kunnen onderscheiden van dat van merk P.

4. Rangordetoets

- Wordt – als het om sensorisch onderzoek gaat – *toegepast* in situaties waarin gegevens bestaan uit de volgordes van door panelleden op basis van een of ander criterium gerangordende monsters of producten. Het gaat daarbij dus om drie of meer producten.
- Het is een toets die rechtstreeks op de getabelleerde gegevens wordt toegepast zonder berekening van een enkele toetsingsgrootheid. In paragraaf 6.9 worden uitgebreid de procedure en het gebruik van tabel G beschreven aan de hand van een voorbeeld.

5. De t-toets

- Wordt *toegepast* in situaties waarin moet worden bepaald of *gemiddelden* verschillen.
- Toepassing bij sensorisch onderzoek: daar waar gegevens het berekenen van gemiddelden en standaarddeviaties toelaten. Met name mogelijk na gebruik van lijnschalen en soms mogelijk geacht bij de verschil-met-referentietest.
- Er zijn drie varianten:
  1. een variant waarmee wordt nagegaan of het gemiddelde in één steekproef afwijkt van een bepaald gegeven getal;
  2. een variant waarmee wordt nagegaan of het gemiddelde in een steekproef afwijkt van dat van een daarvan *onafhankelijke* andere steekproef;
  3. een variant waarmee wordt nagegaan of het gemiddelde in een steekproef afwijkt van dat van een daarvan *afhankelijke* andere steekproef (waarbij dus sprake is van gepaarde waarnemingen).
- De toetsingsgrootheid is t. Deze kan worden berekend met de volgende formules:
  - voor variant 1: $t = \dfrac{\bar{x} - a}{s/\sqrt{n}}$

    Hierbij is $\bar{x}$ het gemiddelde over n gegevens, s is er de standaarddeviatie van en a is het gegeven getal.

  - voor variant 2: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{(n_1-1)s_1^{\,2} + (n_2-1)s_2^{\,2}}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

    De indices 1 en 2 geven de betreffende steekproef aan.

  - voor variant 3: $t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$

    Hierbij is d het verschil tussen de meetwaarden van *elk paar*. Er zijn dus evenveel d's als gegevensparen. Het gemiddelde van al die d's is $\bar{d}$, terwijl s_d de standaarddeviatie van de d's is.

- Van belang is verder ook hier het aantal vrijheidsgraden, df. Hiervoor geldt:
  voor variant 1: $df = n - 1$
  voor variant 2: $df = n_1 + n_2 - 2$
  voor variant 3: $df = n - 1$ (waarbij n het aantal *paren* van waarnemingen is!)

- Tabel H toont de kritieke waarden van t voor diverse waarden van α en df. Merk op dat t zowel negatief als positief kan zijn. Om H_0 te verwerpen moet bij een eenzijdige toetsing de gevonden t soms juist positief, soms juist negatief zijn, afhankelijk van de hypothesen.
- Toepassing van de toets is gebonden aan bepaalde voorwaarden, die vooral de normaliteit van de verdelingen en de variantie van de verdelingen betreffen. Zie daarvoor meer statistisch gespecialiseerde literatuur.
- Voorbeelden van de toepassing:
  - *Variant 1.* Een onderzoeker heeft een panel van 81 consumenten gevraagd de textuur van een product (hedonisch) te beoordelen op een bipolaire lijnschaal (veel te zacht/veel te hard) van 100 mm. 'Precies goed' zit in het midden, op 50 mm dus. Hij vindt een gemiddelde van 55 mm, wat er op duidt dat het product wat te hard zou worden bevonden. De standaarddeviatie van de scores bedraagt 18 mm. De vraag is of 55 significant (met $\alpha = 5\%$) boven het 'precies goed'-punt van 50 ligt. Dus:
    $H_0: \mu \leq 50$
    $H_1: \mu > 50$

    De gevonden t bedraagt $(55-50)/(18/\sqrt{81}) = +2{,}50$. $Df = 81-1 = 80$. De kritieke waarde van t ligt dan tussen 1,658 en 1,671 en wordt dus overschreden. H_0 wordt daarom verworpen, H_1 geaccepteerd.
  - *Variant 2.* Een sensoricus wil (met $\alpha = 5\%$) weten of mannen en vrouwen verschillen in hun oordeel over een product. Een panel van 40 mannelijke en 60 vrouwelijke consumenten geeft een totaaloordeel in de vorm van een rapportcijfer: mannen gemiddeld een 7,1 ($s = 1{,}2$), vrouwen een 7,7 ($s = 1{,}4$).
    $H_0: \mu_{mannen} = \mu_{vrouwen}$

    $H_1: \mu_{mannen} \neq \mu_{vrouwen}$

    De gevonden t bedraagt

$$\frac{7{,}1 - 7{,}7}{\sqrt{\frac{(40-1)1{,}2^2 + (60-1)1{,}4^2}{60+40-2}\left(\frac{1}{60}+\frac{1}{40}\right)}} = -2{,}220$$

    (Let op de volgorde van de bewerkingen.) $Df = 60 + 40 - 2 = 98$. De kritieke waarde van t ligt dan tussen −1,980 en −2,000. H_0 wordt dus verworpen, H_1 geaccepteerd.

- *Variant 3.* Een kwaliteitsonderzoeker wil met een analytisch panel van 16 personen met behulp van lijnschalen van 100 mm (0 = helemaal niet zout, 100 = heel sterk zout) nagaan of product A zouter is dan B ($\alpha = 5\%$). Hij trekt voor elk panellid diens score van product B af van diens score van product A. Aldus ontstaat een serie van 16 verschilscores, met een gemiddelde van *min* 1,8 mm en een standaarddeviatie van 3,6 mm. Dus:

  H_0: $\mu_A \leq \mu_B$
  H_1: $\mu_A > \mu_B$

  De gevonden t bedraagt $-1{,}8/(3{,}6/\sqrt{16}) = -2{,}000$. Df = 16–1 = 15. De kritieke waarde van t is hier **plus** 1,753. H_0 wordt dus niet verworpen, H_1 wordt niet geaccepteerd.

6. Variantieanalyse

- Een ingewikkelde techniek, waar we hier maar globaal op ingaan.
- Eigenlijk gaat het om een uitbreiding van de t-toets, die wordt *toegepast* in situaties waarin moet worden bepaald of *meer dan twee gemiddelden* verschillen.
- Wordt in sensorisch onderzoek toegepast als bij de t-toets, maar dan als er meer dan twee gemiddelden in het geding zijn. De toets heet nu ook F-toets.
- De toetsingsgrootheid F (of F-ratio) wordt als volgt berekend:

$$\text{F-ratio} = \frac{\text{TKS}/(r-1)}{\text{BKS}/(n-r)}$$

  Waarbij r het betrokken aantal producten is, n het totaal aantal waarnemingen/scores is in het onderzoek (dus over alle r producten *samen*), BKS de zogeheten binnenkwadratensom (de som van de gekwadrateerde afwijkingen van elke score ten opzichte van zijn eigen productgemiddelde) en TKS de tussenkwadratensom (de som van de gekwadrateerde afwijkingen van de productgemiddelden ten opzichte van het algemene gemiddelde, nadat elk van die kwadraten vermenigvuldigd is met het bijbehorend aantal waarnemingen).
- Er zijn bij elke F twee soorten vrijheidsgraden in het geding: een voor de teller en een voor de noemer. Df bedraagt voor de teller r–1, voor de noemer n–r.
- Tabel I bevat kritieke waarden van F voor verschillende combinaties van vrijheidsgraden.
- Variantieanalyse is, door de aard van de hypotheses, een tweezijdige toets. Dat wil zeggen dat er is-gelijktekens (=) in de nulhypothese staan.