5.4 Populatieproporties schatten

Inleiding

Je wilt een schatting maken van het percentage havoleerlingen dat doorstroomt naar vwo 5. Dat doe je door één keer een grote steekproef te trekken, want alle havoleerlingen in heel Nederland bevragen is ondoenlijk. En ook veel steekproeven nemen is nauwelijks haalbaar en zeer kostbaar. Maar hoe betrouwbaar is je schatting op grond van één steekproef?

Je leert in dit onderwerp

- betrouwbaarheidsintervallen van een steekproefproportie bepalen om een populatieproportie te kunnen schatten;;
- een populatieproportie toetsen.

Voorkennis

- de normale verdeling met zijn vuistregels over het gemiddelde en de standaardafwijking;
- kansen en grenswaarden berekenen met de normale verdeling;
- hypothese toetsen met de normale verdeling;
- werken met betrouwbaarheidsintervallen bij het schatten van een populatiegemiddelde.

Verkennen

Opgave V1

Je wilt een schatting maken van het percentage havoleerlingen dat doorstroomt naar vwo 5. Dat doe je door één keer een grote steekproef te trekken, want alle havoleerlingen in heel Nederland bevragen is ondoenlijk. En ook veel steekproeven nemen is nauwelijks haalbaar en zeer kostbaar.

a Je vraagt in twee vwo 5 klassen hoeveel leerlingen er vanuit havo 5 komen. Van de 58 leerlingen blijken dat er 3 te zijn. Kun je nu gewoon vaststellen dat 5,1% van de havo leerlingen doorstroomt naar vwo?

b Maar hoe betrouwbaar is je schatting op grond van één steekproef?

Uitleg 1

Bij statistische onderzoeken komen regelmatig vragen voor met maar twee mogelijke antwoorden, bijvoorbeeld:

- Ben je man of vrouw?
- Ben je ouder dan 40 jaar?

Als 39% van de ondervraagden op de vraag ‘Ben je man?’ met ‘ja’ antwoordt is het deel van de ondervraagden dat man is $p_{\text{steekproef man}} = 0{,}39$. Dit heet de steekproefproportie mannen.

Bij ‘nee’ hoort in dit geval $p_{\text{steekproef vrouw}} = 0{,}61$. Deze twee waarden van p zijn bij elkaar altijd 1.

Elke keer dat er een steekproef wordt genomen, kan het deel van de mensen dat ‘ja’ antwoordt, een iets andere waarde hebben. Omdat deze proporties p een steekproevenverdeling vormen, heeft p een normale verdeling mits de steekproef voldoende groot is.
Deze steekproevenverdeling heeft een standaardafwijking σ die je kunt berekenen uit $p_{\text{steekproef}}$ en het aantal mensen n in de steekproef. $p_{\text{steekproef}}$ noteer je ook wel als $\widehat{p}$.

Dan is: $\sigma = \sqrt{\frac{\widehat{p} \cdot (1 - \widehat{p})}{n}}$

Stel dat in een steekproef van 50 uit een grote populatie 72% een bepaalde eigenschap heeft.

Dan is $n = 50$ en $\widehat{p} = 0{,}72$, zodat $\sigma = \sqrt{\frac{0{,}72 \cdot (1 - 0{,}72)}{50}} \approx 0{,}063$.

Uit de vuistregels van de normale verdeling volgt dat 95% van de proporties in de steekproevenverdeling tussen $0{,}72 - 2 \cdot 0{,}063 \approx 0{,}59$ en $0{,}72 + 2 \cdot 0{,}063 \approx 0{,}85$ liggen. (In plaats van 2 kun je ook de z-waarde 1,96 nemen.)

Dit betekent dat met 95% betrouwbaarheid bij elke nieuwe steekproef het deel van de populatie met de eigenschap tussen 0,59 en 0,85 ligt. Het 95%-betrouwbaarheidsinterval van de populatieproportie is dus $[0{,}59; 0{,}85]$.

Opgave 1

Bereken in de volgende gevallen de steekproefproportie en de bijbehorende standaardafwijking. Geef vervolgens het 95%-betrouwbaarheidsinterval voor de populatieproportie.

a Een steekproef van 200 personen, waarvan 130 personen 'ja' antwoorden.

b Een populatie van 11000 personen, waarvan 12% als steekproef wordt genomen.
63% van de mensen uit de steekproef zegt '1 of meer kinderen te hebben'.

Opgave 2

Er wordt een onderzoek gedaan naar welk deel van de festivalgangers boven de 40 jaar is. Bij een festival is door middel van een steekproef van 150 bezoekers de leeftijd gevraagd. Van deze groep blijken 50 bezoekers 40 jaar of ouder te zijn. Bereken de standaardafwijking van de steekproevenverdeling. Rond af op vier decimalen. Schat hiermee welk percentage van de festivalgangers boven de 40 jaar is.

Uitleg 2

In een vwo 5 klas zit 1 leerling die na zijn havodiploma doorstroomde.

Omdat er in totaal 30 leerlingen in deze klas zitten, is de proportie havogediplomeerden $\frac{1}{30} \approx 0{,}033$.

Een aantal jaar geleden heeft een onderzoeksbureau door middel van een steekproef berekend dat de proportie havogediplomeerden in vwo 5- en 6-klassen in Nederland met een betrouwbaarheid van 95% tussen de 7,2% en de 15,2% ligt. Het 95%-betrouwbaarheidsinterval is dus $[7{,}2; 15{,}2]$.

Je kunt op basis van het resultaat van 1 havogediplomeerde in de steekproef van 30 toetsen of de huidige proportie havogediplomeerden niet lager is dan 0,072 (de linkergrens van het betrouwbaarheidsinterval).

De nulhypothese is H_0: $p = 0{,}072$.
De alternatieve hypothese is H_1: $p < 0{,}072$.
Neem bijvoorbeeld een significantieniveau van 5%.

Je berekent nu $\text{P}(X \leq 1)$, ervan uitgaande dat ieder van de 30 leerlingen een kans van 0,072 heeft om havo gediplomeerd te zijn. Deze kans is ongeveer 0,354.

Omdat $0{,}354 > 0{,}05$ is er geen reden om de nulhypothese te verwerpen.

Opgave 3

Gebruik de gegevens uit **Uitleg 2**.

a Wat voor soort toets wordt er genomen? Links-, rechts- of tweezijdig?

b Als je een significantieniveau van 10% neemt, is er dan reden om de nulhypothese te verwerpen?

c Reken na dat de kans dat maximaal 1 leerling van het havo komt ongeveer 0,354 is.

Opgave 4

Stel dat er in een groep van 70 vwo 5 leerlingen er twee met een havodiploma zitten.
Doe dezelfde hypothesetoets als in **Uitleg 2**, alleen nu met deze gegevens. Gebruik een significantieniveau van 10%.

Theorie en voorbeelden

Om te onthouden

Een **populatieproportie** p is het deel van de populatie dat voldoet aan een zeker kenmerk, uitgedrukt als percentage of fractie.

Omdat een populatieproportie, net als een populatiegemiddelde, vaak niet bekend is, bestaat er ook verklarende statistiek die de populatieproportie onderzoekt.
Een steekproef met steekproefomvang n heeft in dat geval als steekproefuitslag de **steekproefproportie** $\widehat{p}$.

De steekproevenverdeling is volgens de centrale limietstelling normaal verdeeld als de steekproefomvang voldoende groot is.

De standaardafwijking van de steekproevenverdeling is: $\sigma = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$.

Het 95%-betrouwbaarheidsinterval voor de populatieproportie p is met $z \approx 2$:
$\left[\widehat{p} - 2 \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}}; \widehat{p} + 2 \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}}\right]$.

Andere betrouwbaarheidsintervallen kunnen met behulp van de z-waarden berekend worden.

Een **hypothesetoets op een populatieproportie** werkt als volgt.
Iemand beweert: de populatieproportie is gelijk aan p.
Iemand anders vertrouwt deze proportie niet en vermoedt (bijvoorbeeld): de populatieproportie is kleiner dan p.

Dit wordt getoetst met een steekproef van grootte n.
Je bepaalt dan de steekproefproportie $\widehat{p}$ en kijkt of deze significant afwijkt van de nulhypothese.

Je kiest een vooraf vastgesteld significantieniveau en dus ook een kritiek gebied. De hypothesetoets kan zowel linkszijdig, rechtszijdig als tweezijdig zijn.

Voorbeeld 1

Een fabrikant van gamefiguren trekt regelmatig een steekproef ter controle van zijn productieproces. Na een grondige renovatie van de fabriek en een periode van testdraaien, controleert de fabrikant 800 aselect gekozen figuren.
Hierbij vertonen er 160 een mankement.

Welk 95%-betrouwbaarheidsinterval voor de proportie defecte gamefiguren van zijn nieuwe productieproces kan de fabrikant op basis van deze controle bepalen en welke conclusie kan hij daarmee over zijn nieuwe productieproces trekken?

Antwoord

Steekproefproportie $\widehat{p}$ is $\frac{160}{800} = 0{,}2$ en steekproefomvang n is 800.

linkergrens: $0{,}2 - 2 \cdot \sqrt{\frac{0{,}2 \cdot (1-0{,}2)}{800}} \approx 0{,}172$

rechtergrens: $0{,}2 + 2 \cdot \sqrt{\frac{0{,}2 \cdot (1-0{,}2)}{800}} \approx 0{,}228$

Het 95%-betrouwbaarheidsinterval is $[0{,}172; 0{,}228]$.

De fabrikant weet nu dat de proportie defecte gamefiguren van zijn nieuwe productieproces in 95% van de steekproeven tussen de 17,2% en de 22,8% van de geproduceerde figuren zal liggen.

Opgave 5

Bekijk **Voorbeeld 1**. Na enige tijd is de fabrikant begonnen om op gezette tijden een steekproef van 50 gamefiguren te nemen. De laatste keer dat hij dat deed, was 14% van de steekproef defect.

a Bepaal de steekproefproportie $\widehat{p}$ en de steekproefomvang n van deze steekproef.

b Bereken het 95%-betrouwbaarheidsinterval voor de schatting van het percentage defecte gamefiguren.

c Is er een kans dat de daadwerkelijke proportie defecte figuren in zijn nieuwe productieproces hoger is dan 25%? Beargumenteer je antwoord.

Opgave 6

De Consumentenbond wil weten of een bepaald type laptop minstens acht uur op de accu kan werken. De bond test 50 aselect getrokken laptops van dat type. Het blijkt dat 41 van de laptops inderdaad minstens 8 uur kan werken op de accu.

a Bereken de steekproefproportie.

b Bereken de standaardafwijking σ van de steekproefproportie. Rond je antwoord af op vier decimalen.

c Bepaal het 90%-betrouwbaarheidsinterval van de populatieproportie.

d Welke uitspraak kun je met 90% betrouwbaarheid doen?

Voorbeeld 2

Een groenteboer verkoopt een zak met 10 peren; 2 daarvan blijken van binnen helemaal rot. De groenteboer beweert dat slechts 4% van zijn peren bij aankoop al rot is.
Hoe kun je deze bewering toetsen met een hypothesetoets en welke conclusie kun je trekken met de uitslag van deze hypothesetoets als het significantieniveau 5% is?

Antwoord

Dit is een rechtszijdige hypothesetoets met als hypothesen:
H_0: $p = 0{,}04$
H_1: $p > 0{,}04$

Toevalsvariabele R is het aantal rotte peren in de steekproef van 10 peren.
Te berekenen: $\text{P}(R \geq 2)$ als iedere peer een kans van 0,04 heeft om rot te zijn.
$\text{P}(R \geq 2) = 1 - \text{P}(R \leq 1) = 1 - \text{P}(R = 0 \text{ of } R = 1) \approx 0{,}0582$ ofwel 5,8% .

Deze kans is groter dan het significantieniveau van 5% en de bewering van de groenteboer hoeft niet verworpen te worden.

Opgave 7

Gebruik de gegevens uit **Voorbeeld 2**.

a Leg uit waarom is $\text{P}(R \geq 2)$ in deze situatie gelijk is aan $1 - \text{P}(R \leq 1)$.

b Bereken zelf de kans op minimaal 2 rotte peren als je er 10 koopt terwijl iedere peer een kans van 4% heeft om rot te zijn.

Opgave 8

Gebruik de gegevens uit **Voorbeeld 2**. Bedenk een manier om het kritieke gebied te formuleren voor de hypothesetoets.

Verwerken

Opgave 9

Bij een aselecte steekproef in een provincie blijken onder 1500 geënquêteerden er 833 tegen de aanleg van een provinciale weg te zijn.

a Bereken de steekproefproportie. Rond je antwoord af op drie decimalen.

b Bereken de standaardafwijking van deze steekproevenverdeling. Rond je antwoord af op vier decimalen.

c Tussen welke percentages ligt binnen deze provincie het aantal tegenstanders van deze weg met een betrouwbaarheid van 95% in een nieuwe steekproef?

Opgave 10

Een fabrikant in lampen verkoopt een doos met 100 lampen; 2 daarvan blijken kapot te zijn. De fabrikant beweert dat slechts 0,5% van zijn lampen bij aankoop kapot zijn.

Toets de bewering van de fabrikant met een significantieniveau van 5%.

Opgave 11

Uit een enquête in opdracht van de Stichting tegen Kanker van maart/april 2007 onder 1988 Belgen bleek 61% voorstander te zijn van het rookvrij maken van cafés. In oktober 2006 was dat nog maar 55% van een even grote groep.

a Bepaal bij het onderzoek van oktober 2006 het 95%-betrouwbaarheidsinterval. Geef je antwoord in procenten. Rond af op één decimaal.

b Bepaal bij het onderzoek van maart/april 2007 het betrouwbaarheidsinterval bij een betrouwbaarheid van 95%. Geef je antwoord in procenten. Rond af op één decimaal.

c Welke conclusie kun je trekken?

Opgave 12

Minke heeft een dobbelsteen, waarvan ze zegt dat die onzuiver is. Ze beweert dat je gemiddeld in slechts 8% van de keren 6 ogen gooit.
Bart gaat deze bewering toetsen. Hij gooit 30 keer met de dobbelsteen en gooit 4 keer 6 ogen.

a Stel de hypothesetoets op.

b Wat is de conclusie van Bart bij een significantieniveau van 10%?

Opgave 13

Voorafgaand aan de verkiezingen worden opiniepeilingen gehouden. Met behulp van een aselecte steekproef wordt aan 2000 Nederlanders gevraagd naar de partij van hun voorkeur. Eén partij gaat in een opiniepeiling van 30 naar 31 zetels (van de 150 zetels in de Tweede Kamer).
Onderzoek of de partij met een betrouwbaarheid van 90% zeker kan zijn van een zetel winst.

Opgave 14

Onderzoekers hebben door middel van een steekproef een bepaald kenmerk onderzocht. Zij concludeerden dat 8% van de steekproef het kenmerk heeft. Het 95%-betrouwbaarheidsinterval is [7,97; 8,03].

a Toon aan dat de standaardafwijking van de steekproef 0,015 is.

b Bereken de grootte van de steekproef.

Toepassen

Opgave 15: Lampen

Een fabrikant van lampen beweert dat 5% van de lampen die verkocht worden kapot zijn.
De controledienst gaat deze bewering toetsen.
Bij een steekproef van 500 lampen vond de controledienst 35 kapotte lampen.

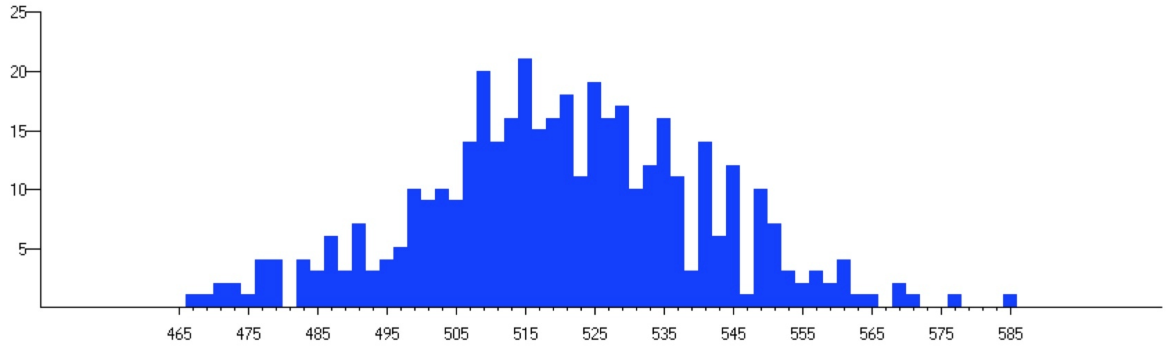
Om een conclusie te kunnen trekken moet je de kans $\text{P}(X \geq 35)$ uitrekenen, waarbij X het aantal kapotte lampen is. Vanwege de centrale limietstelling kun je deze kans benaderen met behulp van een normale verdeling met gemiddelde $500 \cdot 0{,}05 = 25$ en standaardafwijking $\sqrt{500 \cdot 0{,}05(1-0{,}05)}$.
Vanwege het afronden moet je nu $\text{P}(X \geq 34{,}5)$ berekenen.

a Waarom zou je bij de normale benadering als grens 34,5 gebruiken?

b Toets de bewering van de controledienst. Welke conclusie kun je trekken met de uitslag van deze hypothesetoets als het significantieniveau 5% is?

Opgave 16: Ja of nee?

Bekijk het histogram van de steekproevenverdeling van 400 steekproeven uit een ja/nee-verdeelde populatie, ieder met een omvang van 2000.



Figuur 1

a Bepaal op basis van de kleinste en grootste steekproefuitslag en met behulp van de vuistregels van de normale verdeling de gemiddelde proportie en de standaardafwijking van deze 400 steekproeven.

b Bepaal met de vuistregels van de normale verdeling en met je net berekende gemiddelde en standaardafwijking wat de grenzen van het 95%-gebied zijn van deze steekproevenverdeling.

c Leg uit wat de betekenis van dit 95%-gebied en van de gemiddelde proportie is.

Uit dezelfde populatie waaruit de 400 steekproeven getrokken zijn, wordt opnieuw een steekproef van 2000 stuks getrokken. Bekijk de verdeling van deze steekproef: de steekproefuitslag is 512 op de 2000.

Figuur 2

d Bepaal het 95%-betrouwbaarheidsinterval van deze steekproef en leg uit wat de betekenis ervan is, ook in samenhang met de gemiddelde proportie uit de steekproevenverdeling van 400 andere steekproeven.

Testen

Opgave 17

Helmond is een stad van ongeveer 90000 inwoners en de stad heeft een stadspanel met 1300 deelnemers. Op een verzoek van de gemeente een vragenlijst in 2012 over de lokale omroep in te vullen hebben 839 mensen gereageerd.
Uit deze meting blijkt dat 37% de lokale omroep niet kent.

a Welke uitspraak kun je doen met 95%-betrouwbaarheid over de onbekendheid van de lokale omroep in Helmond?

b Welke uitspraak kun je met 95% betrouwbaarheid doen over de bekendheid van de lokale omroep in Helmond?

c Welke kanttekeningen kun je bij dit onderzoek maken?

Opgave 18

Bij een marktonderzoek wordt gekeken naar de belangstelling voor elektrische auto's onder particulieren in Nederland. In een aselecte steekproef worden 1660 mensen benaderd en hiervan zeggen 917 particulieren dat ze de overstap naar een elektrische auto serieus overwegen.
De onderzoekers willen het percentage belangstellenden met een betrouwbaarheid van 95% vaststellen met een marge van 0,5%.

Hoe groot moet hun steekproef dan zijn?

Practicum

Met de volgende practica kun je zien hoe je betrouwbaarheidsintervallen berekent met de **grafische rekenmachine**. Doe alleen het onderdeel dat betrekking heeft op het betrouwbaarheidsinterval bij proporties.

- **Kansverdelingen met de TI84**
- **Kansverdelingen met de TInspire**
- **Kansverdelingen met de Casio fx-CG50**
- **Kansverdelingen met de HP-prime**
- **Kansverdelingen met de NumWorks**

Maar je kunt ook heel goed betrouwbaarheidsintervallen berekenen met behulp van **Excel**. Bekijk daartoe het practicum:

- **Steekproeven en uitspraken**