# Hoofdstuk 8: Statistische gevolgtrekkingen voor proporties

## 8.1 Een enkele proportie

### Steekproefproportie

We willen vaak weten hoe het met proporties in de populatie zit. Hoeveel procent van de Nederlanders is bijvoorbeeld 18 jaar of ouder? Hoeveel procent van de Nederlandse studenten is tegen de legalisering van drugs?

- De steekproefproportie ($\hat{p}$) wordt berekend door een telling (X) te delen door het aantal deelnemers (n).
- De steekproefproportie wordt gebruikt om de populatieproportie te schatten. Als de populatie minstens 20 keer zo groot is als de steekproef, dan heeft telling X ongeveer een binomiale distributie B(n,p). Als de steekproefgrootte n erg klein is, moeten we significantietoetsen en betrouwbaarheidsintervallen voor p baseren op de binomiale distributie. Als de steekproef groot is, dan is zowel telling X als de steekproefproportie normaalverdeeld.

### Het betrouwbaarheidsinterval voor een grote steekproef

De onbekende populatieproportie p wordt dus geschat aan de hand van de *steekproefproportie* $\hat{p}$=X/n. In deze formule staat X voor het aantal successen.

- Als de steekproefgrootte groot genoeg is, dan is $\hat{p}$ bijna normaalverdeeld met een gemiddelde van p en een standaarddeviatie van $\sqrt{p(1-p)/n}$ Dit betekent dat in 95% van de gevallen $\hat{p}$ binnen 2 $\sqrt{(1-p)/n}$ ligt.
- *De standaardfout* van $\hat{p}$ is de wortel uit $\hat{p}$ (1-$\hat{p}$)/n.
- De *foutenmarge* voor betrouwbaarheidsinterval C is m=z*SE$\hat{p}$. In deze formule is z* de waarde voor de standaard normaalverdeelde curve met gebied C tussen –z* en z*.
- Een *benaderd betrouwbaarheidsinterval* voor p is $\hat{p}$±m. Dit interval moet gebruikt worden voor 90%, 95% of 99% intervallen en wanneer het aantal successen en niet-successen allebei minstens 15 zijn. Voor een voorbeeld zie bladzijde 470.

### Het plus vier betrouwbaarheidsinterval voor een enkele proportie

Uit onderzoek blijkt dat betrouwbaarheidsintervallen die op steekproeven gebaseerd zijn die minder dan 15 deelnemers hebben, vaak niet accuraat zijn. Wanneer dit het geval is, kan een simpele aanpassing aan het betrouwbaarheidsinterval effectief zijn.

- De *plus vier regel* om een populatieproportie te schatten is: p-golf= X+2/n+4. Dit wordt ook wel de *plus vier schatting* genoemd. Het betrouwbaarheidsinterval is gebaseerd op de z-toets die verkregen wordt door de plus vier schatting te standaardiseren. De distributie van de plus vier schatting is bijna normaalverdeeld met gemiddelde p en standaarddeviatie $\sqrt{p(1-p)/(n+4)}$ Om een betrouwbaarheidsinterval te krijgen schatten we p met p-golf.
- Om de *standaardfout* van p-golf te vinden, moet eerst p-golf(1-p-golf)/n+4 berekend worden. Vervolgens moet de wortel uit deze uitkomst getrokken worden.

- De *foutenmarge* voor betrouwbaarheidsinterval C is: m=z*SEp-golf, waarbij z* de waarde voor de standaard normaalverdeelde dichtheidscure is met gebied C tussen –z* en z*.
- Het *benaderde betrouwbaarheidsinterval C van* p is p-golf±m. Dit interval dient gebruikt te worden voor 90%, 95% of 99% intervallen als de steekproef minstens uit 10 deelnemers bestaat.

**Significantietoets voor een enkele proportie**

- Stel: je trekt een SRS van grootte n uit een grote populatie met een onbekende proportie p van successen. Om de nulhypothese te toetsen dat de proportie uit de nulhypothese klopt, maken we gebruik van de volgende berekening:
- Eerst berekenen we $\hat{p}$- $p_0$.
- Vervolgens berekenen we $p_0(1- p_0)/n$. Uit deze uitkomst trekken we de wortel.
- Tot slot delen we de eerste berekening door de tweede berekening. De uitkomst is een z-toets.

Als de populatie niet minstens 20 keer zo groot als de steekproef is, dan dient deze procedure niet gebruikt te worden. Als een steekproef groot is, dan heeft de bijbehorende significantietoets een hoge power. Dit zorgt ervoor dat zelfs een klein effect vastgesteld kan worden. Als een steekproef erg klein is, dan kunnen belangrijke verschillen over het hoofd gezien worden.

**Betrouwbaarheidsintervallen geven aanvullende informatie**

Een betrouwbaarheidsinterval geeft altijd meer informatie dan de uitkomst van een significantietoets. We gebruiken in de praktijk zelden significantietoetsen voor een enkele proportie, omdat het in de echte wereld zelden voorkomt dat er een precieze $p_0$ bestaat die we willen toetsen. Uit data van vroegere grote steekproeven kan soms de waarde van $p_0$ afgeleid worden.

**Een steekproefgrootte kiezen**

Als we aan de hand van een vaststaande foutenmarge een bijbehorende steekproefgrootte moeten kiezen, gebruiken we de volgende formule:

- $N = (z^*/m)^2 p^*(1-p^*)$.

De foutenmarge hangt af van z*, $\hat{p}$ en n. Omdat we de waarde van $\hat{p}$ niet kennen totdat we de data verzameld hebben, moeten we raden wat deze waarde is om de waarde in onze berekeningen te kunnen gebruiken. Deze geraden waarde noemen we p*. De waarde kan op twee manieren gevonden worden:

1. Gebruik een steekproefschatting die voortvloeit uit eerdere, soortgelijke onderzoeken.
2. Gebruik p*=0.5. Omdat de foutenmarge het grootst is als $\hat{p}$ 0.5 is, geeft deze keuze een steekproefgrootte die iets groter is dan wat we daadwerkelijk nodig hebben.

Als we p* gekozen hebben en een foutenmarge hebben vastgesteld, kunnen we de benodigde steekproefgrootte berekenen met de volgende formule:

- $N = 1/4(z^*/m)^2$

In deze formule is z* de kritische waarde voor betrouwbaarheid C en p* is de geraden waarde voor de proportie van successen in de toekomstige steekproef. De foutenmarge zal kleiner of gelijk aan m zijn als p* 0.5 gekozen wordt. De waarde van de verkregen n is niet erg gevoelig voor de keuze van p*, als deze maar dichtbij de 0.5 ligt. Als de waarde van p kleiner dan 0.3 of groter dan 0.7 is, dan kan het gebruik van p*=0.5 leiden tot het gebruik van een steekproefgrootte die veel groter uitvalt dan gewenst is.

### 8.2 Twee proporties vergelijken

**Populatieproporties en steekproevenproporties**
In de praktijk willen we vaak twee proporties (die gepaard gaan met verschillende groepen) vergelijken. De twee groepen die we vergelijken noemen we 'populatie 1' en 'populatie 2'. De twee populatieproporties noemen we $p_1$ en $p_2$. De data bestaan uit twee afzonderlijke random geselecteerde steekproeven met grootte $n_1$ voor de eerste populatie en grootte $n_2$ voor de tweede populatie. De proportie successen in elke steekproef schat de corresponderende populatieproportie.

- De steekproefproportie voor de eerste steekproef is $\hat{p}_1 = X_1/n_1$.
- De steekproefproportie van de tweede steekproef is $\hat{p}_2 = X_2/n_2$.
- Om de twee populaties te vergelijken, gebruiken we het verschil tussen de twee steekproefproporties: $D=\hat{p}_1 - \hat{p}_2$. D staat voor 'difference'.

Als de twee steekproeven groot zijn, dan is de steekproevendistributie van D normaalverdeeld. Proporties worden door middel van z-toetsen met elkaar vergeleken. De eerste stap is het vaststellen van het gemiddelde en de standaarddeviatie van D:

- $\mu_D = p_1 - p_2$.
- $\sigma^2_{D=} p_1(1-p_1)/ n_1 + p_2(1-p_2)/ n_2$.
- De standaarddeviatie kan gevonden worden door de wortel uit de variantie van D te trekken.

**Het plus vier betrouwbaarheidsinterval voor D**
Een kleine aanpassing van steekproefproporties kan de accuraatheid van betrouwbaarheidsintervallen sterk verbeteren. De plus vier schattingen van twee populatieproporties zijn:

- $p_1$-golf= $X_1 + 1/ n_1 + 2$ en $p_2$-golf= $X_2 + 1/ n_2 + 2$.
- Het geschatte verschil tussen de populaties is: D-golf = $p_1$-golf-$p_2$-golf.
- De standaarddeviatie van D-golf wordt gevonden door eerst $p_1(1- p_1)/(n_1+2)+ p_2(1-p_2)/(n_2+2)$ uit te rekenen. Vervolgens moet de wortel uit de uitkomst getrokken worden.
- De *standaardfout* van D-golf wordt gevonden door eerst $p_1$ golf$(1-p_1$ golf$)/(n_1+2)+ p_2$ golf$(1-p_2$ golf$)/(n_2+2)$ uit te rekenen. Daarna moet de wortel uit deze uitkomst getrokken worden.
- De *foutenmarge* is: $m=z^*SE_{D\text{-golf}}$. In deze formule is z* de waarde van de normaalverdeelde curve met gebied C tussen –z* en z*.
- Een *benaderd betrouwbaarheidsinterval C* voor $p_1-p_2$= D-golf±m. Deze formule dient gebruikt te worden voor betrouwbaarheidsintervallen van 90%, 95% en 99% en als beide steekproeven minimaal uit 5 observaties bestaan.

**Betrouwbaarheidsintervallen voor D bij grote steekproeven**

Om een betrouwbaarheidsinterval voor $p_1$-$p_2$ te berekenen, gebruiken we niet de standaarddeviatie van de populatie (want deze is onbekend), maar de standaarddeviatie van de steekproeven. Dit resulteert in de standaardfout.

- De *standaardfout van D ($SE_D$)* wordt gevonden door eerst $\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2$ uit te rekenen. Uit deze uitkomst moet vervolgens de wortel getrokken worden.
- De *foutenmarge* voor het betrouwbaarheidsinterval is: $m= z^*SE_D$. In deze formule is z* de waarde van de normaalverdeelde curve met gebied C tussen –z* en z*. Een *benaderd betrouwbaarheidsinterval C* voor $p_1$-$p_2$= D±m. Deze formule dient gebruikt te worden voor betrouwbaarheidsintervallen van 90%, 95% en 99% en als het aantal successen en niet-successen in elke steekproef minstens 10 is.

**Significantietoets voor D**

We geven de voorkeur aan het berekenen van betrouwbaarheidsintervallen voor D, maar in sommige gevallen worden ook significantietoetsen voor D uitgevoerd. De nulhypothese is dan dat de twee populatieproporties hetzelfde zijn. We standaardiseren $D=\hat{p}_1 - \hat{p}_2$ als volgt:

- $\sigma_{D=} p_1(1-p_1)/ n_1+ p_2(1-p_2)/ n_2$. Vervolgens moet de wortel uit de deze uitkomst getrokken worden. Als de steekproeven groot zijn, dan zal het gestandaardiseerde verschil ongeveer een gemiddelde van 0 en een standaarddeviatie van 1 hebben: N(0,1).
- We schatten de waarde van p door middel van de algemene proportie van successen in de twee steekproeven: $\hat{p}_= X_1+X_2/n_{1+}n_1$. De schatter van p wordt de *gepoolde schatter* genoemd, omdat deze de informatie van beide steekproeven combineert. Om deze gepoolde schatter te vinden moet allereerst $SE_{DP} =\hat{p}_1(1-\hat{p}_1)(1/n_1+1/n_2)$ berekend worden. Vervolgens moet de wortel uit het resultaat getrokken worden.
- Bij een significantietoets voor het toetsen van proporties, geldt: $H_0$:$p_1=p_2$.
- De z-toets wordt gevonden aan de hand van de formule $z=\hat{p}_1 -\hat{p}_2 / SE_{DP}$. Vervolgens moet deze z-toets opgezocht worden in de z-tabel om een p-waarde te vinden en deze te gebruiken om de nulhypothese te behouden of af te wijzen.
- Het *relatieve risico (RR)* is een ratio van beide steekproefproporties. Als onze steekproefproporties $\hat{p}_1$ en $\hat{p}_2$ zijn, dan wordt RR gevonden door de steekproefproporties door elkaar te delen: $RR=\hat{p}_1/\hat{p}_2$.