Simuleren van peilingen met PollSim

Jelke Bethlehem, januari 2021

1. Inleiding

Het programma PollSim

Met het programma *PollSim* kun je peilingen nabootsen (simuleren) die zijn gebaseerd op aselecte steekproeven uit een doelpopulatie van personen. Het programma laat zien hoe belangrijk het is om een steekproef op correcte wijze te trekken. Het toont aan dat je met een aselecte steekproef valide en precieze schattingen kunt maken van allerlei kenmerken van een populatie.

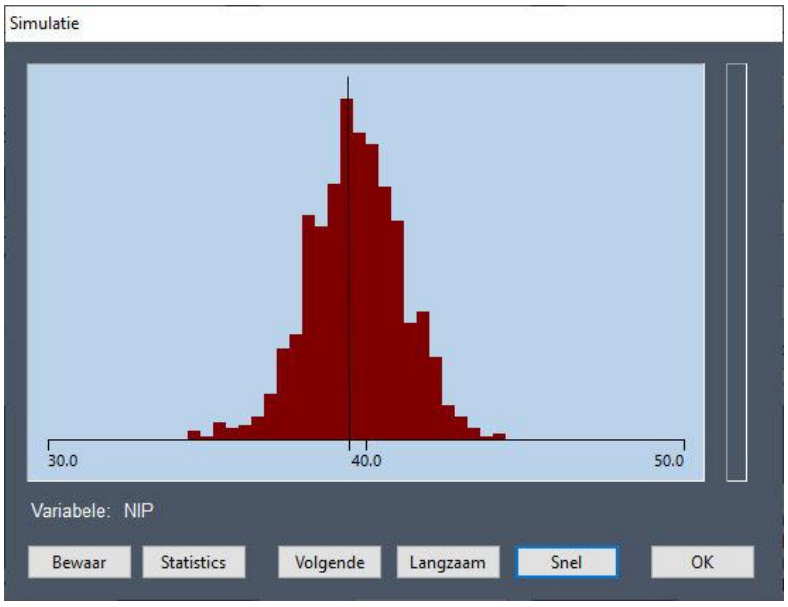
Het programma gaat er vanuit dat je het *gemiddelde* van een variabele in de populatie wilt schatten. Als de variabele slechts de waarden 1 of 0 aanneemt (1 als je een bepaalde eigenschap wel hebt en 0 als je die niet hebt), dan veronderstelt het programma dat je het *percentage* personen in de populatie met die eigenschap wilt schatten. Als, bijvoorbeeld, de doelvariabele meet of iemand wel of niet op een bepaalde politieke partij denkt te gaan stemmen (1=Ja, 0=Nee), dan schat het programma het percentage mensen (in de populatie) dat op die partij denkt te gaan stemmen.

Simulatie met PollSim

Met het programma kun je een *simulatie* uitvoeren. Dat betekent dat je het trekken van de steekproef en het vervolgens berekenen van een schatting voor het populatiegemiddelde of populatiepercentage een groot aantal keren herhaalt. Door een grafiek (histogram) te maken van de verdeling van al die schattingen krijg je een beeld van hoe goed zo'n peiling in staat is een populatiekenmerk te schatten.

Figuur 1.1 toont een voorbeeld van de uitkomsten van een simulatie met *PollSim*. Uit de denkbeeldige gemeente *Samplona* is 1,000 keer een steekproef getrokken van 1.000 personen. Aan de personen in de steekproeven is steeds gevraagd op welke partij ze gaan stemmen bij de komende verkiezingen. Voor elke steekproef is het percentage stemmen op de NIP (*Nieuwe Internet Partij*) berekend. Van al die steekproefpercentages is een histogram gemaakt.

Figuur 1.1. Het resultaat van een simulatie in PollSim.



Elke steekproef levert een schatting op. Die schattingen worden weergegeven als roodbruine blokjes. Waar nodig, zijn de blokjes op elkaar gestapeld. De verticale zwarte lijn duidt het te schatten populatiepercentage (39,5%) aan. In de grafiek is te zien dat de schattingen keurig geconcentreerd liggen om het populatiepercentage. Er is geen sprake van systematische onder- of overschatting. De verdeling is symmetrisch met een top in het midden. Dit is een *normale verdeling*. We kunnen concluderen dat hier sprake is van *valide* (*zuivere*) schattingen.

In hoofdstuk 2 van deze handleiding bespreken we eerst de populatie waarmee we het werken met *PollSim* illustreren. In hoofdstuk 3 leggen we uit hoe je het programma instelt voor een simulatie. En in hoofdstuk 4 leggen we uit hoe je vervolgens de simulatie uitvoert.

In hoofdstuk 5 leggen we uit wat *non-respons* is en hoe je dat kan genereren in je steekproeven. Met simulaties kun je laten zien wat de akelige effecten van non-respons op de uitkomsten van je peiling kunnen zijn. Als je peiling is aangetast door non-respons, kun je proberen de uitkomsten van je peiling te corrigeren. Daarvoor kun je een weging uitvoeren. Hoe je dat doet, beschrijven we in hoofdstuk 6.

Tenslotte leggen we in hoofdstuk 7 uit hoe je eigen populatiebestanden kunt maken voor *PollSim*.

2. Samplona

Het voorbeeld

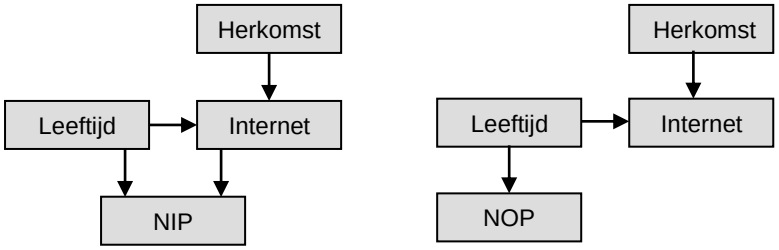
Om de werking van *PollSim* te beschrijven, gebruiken we een concreet voorbeeld. We hebben een doelpopulatie geconstrueerd die uit 30.000 personen bestaat. Dat zijn alle kiesgerechtigde personen in de denkbeeldige gemeente *Samplona*. In het bestand zijn de volgende vijf variabelen beschikbaar:

- De variabele *Leeftijd* met drie categorieën: *Jong*, *Middelbaar* en *Oud*.
- De variabele *Herkomst* met twee categorieën *Nederlandse achtergrond* en *Migratieachtergrond.*
- De variabele *Internet* met twee categorieën *Ja* (heeft internet) en *Nee* (heeft geen internet). De kans op het hebben van internet hangt af van *Leeftijd* en *Herkomst*. Internettoegang neemt af met de leeftijd. Ook hebben personen met een migratieachtergrond veel minder vaak internet dan personen met een Nederlandse achtergrond.
- De variabele *Stemt op de Nationale Ouderen Partij (NOP)* met twee categorieën: *Ja* en *Nee*. De kans om op deze partij te stemmen hangt af van de variabele *Leeftijd*. Vooral ouderen stemmen op de NOP.
- De variabele *Stemt op de Nieuwe Internet Partij (NIP)* met twee categorieën *Ja* en *Nee*. Voor personen met internet neemt de kans op een stem op de NIP af met de leeftijd. Vooral jongeren met internet stemmen op de NIP. Voor personen zonder internet is de kans op een stem op de NIP erg laag ongeacht de leeftijd.

Figuur 2.1 geeft een grafische weergave van de relaties tussen de verschillende variabelen in de populatie Samplona. De relaties zijn wat sterker gemaakt dan ze in een werkelijke situatie zouden zijn. De effecten zijn dus wat meer uitgesproken.

Of iemand stemt op de NOP, hangt dus af van zijn of haar leeftijd. Ouderen stemmen eerder op deze partij dan jongeren. Of een persoon op de NIP stemt, hangt af van zijn of haar leeftijd en of die persoon internet heeft. Mensen met internet zijn eerder geneigd op de NIP te stemmen dan mensen zonder internet. En verder stemmen jongeren eerder op de NIP dan ouderen. Er is ook een relatie tussen herkomst en internetbezit. Internetbezit is erg laag bij personen met een migratieachtergrond en hoog bij personen met een Nederlandse achtergrond.

Figuur 2.1. Relaties tussen de variabelen in de populatie Samplona



Het populatiebestand

Figuur 2.2 bevat een tabel met de verdeling van een aantal variabelen in het populatiebestand van Samplona. In deze tabel met populatiepercentages is bijvoorbeeld te zien dat 39,5% van de kiesgerechtigde inwoners van Samplona zegt op de NIP te gaan stemmen. En 25,4% kiest voor de NOP. We kunnen deze percentages uitrekenen om dat we de populatie helemaal kennen. We hebben hem immers zelf geconstrueerd. Normaal is dit niet het geval. Dan doen we juist een peiling omdat we de populatie niet kennen. Met de uitkomsten van peilingen proberen we schattingen te maken van dit soort kenmerken van een populatie. En vraag is nu hoe goed die schattingen zijn.

Figuur 2.2. Populatiepercentages van een aantal variabelen in Samplona

Variabele	Ja	Nee	Totaal
Migratieachtergrond	15,1%	84,9%	100,0%
Heeft internet	63,4%	36,6%	100,0%
Stemt op de NIP	39,5%	60,5%	100,0%
Stemt op de NOP	25,4%	74,6%	100,0%

We richten ons bij het beschrijven van het programma PollSim op het schatten van het percentage NIP-stemmers in de populatie. Dus we proberen het populatiepercentage van 39,5% voor de NIP te schatten. We laten zien dat grotere steekproeven preciezere schatters opleveren. We laten ook zien hoe non-respons de schatting kan aantasten. En we beschrijven hoe je door het uitvoeren van een weging kunt proberen voor de effecten van non-respons te corrigeren.

3. Het instellen een simulatie

Om een simulatie uit te voeren, moet je eerste een aantal zaken specificeren. Dat doe je op het hoofdscherm van PollSim. Zie figuur 3.1 voor een voorbeeld. Het scherm is verdeel in een aantal panelen. Elk paneel behandelt een ander aspect van de simulatie. Het scherm bevat ook een aantal knoppen. De functies van die panelen en knoppen gaan we hieronder stap voor stap beschrijven. We beginnen met een overzicht van de stappen die je voor een simulatie moet doorlopen:

(1) Lees een populatiebestand in. Dat doe je met de knop **Lees populatie**.

(2) Geef op wat voor steekproeven je wilt trekken. Dat doe je op het paneel *Steekproef*. Vul in hoe groot de steekproeven moeten zijn. Selecteer ook de doelvariabele. Dit is de variabele waarvoor jet het populatiegemiddelde of het populatiepercentage wilt schatten

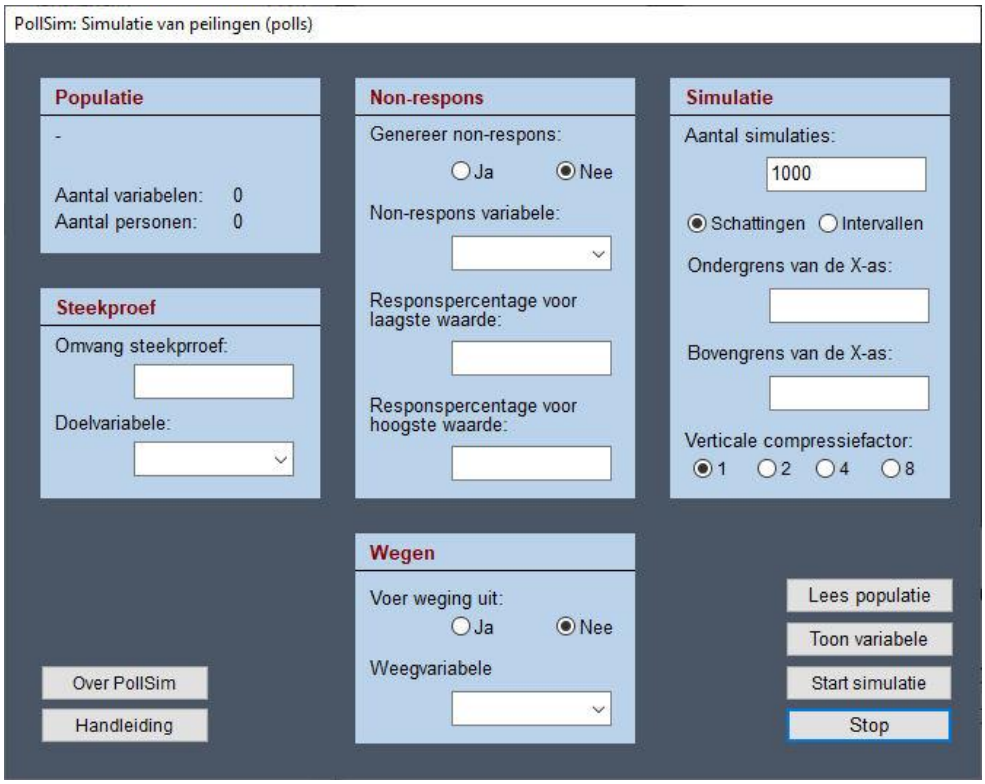
(3) Indien je dat wenst, kun je non-respons genereren. Dat doe je op het paneel *Non-respons*. Bij het opstarten van het programma staat het genereren van non-respons uit. Je krijgt dan dus steekproeven met volledige respons.

(4) Indien je dat wenst, kun je de uitkomsten wegen. Dat doe je op het paneel *Wegen*. Dit is vooral van belang voor steekproeven die zijn aangetast door non-respons. Bij het opstarten van het programma staat het wegen van uitkomsten uit.

(5) Geef de simulatie-parameters op. Dat doe je op het paneel *Simulatie*. Bij het opstarten van het programma is het aantal te trekken steekproef alvast op 1.000 gezet. Uiteraard kun je dit aanpassen.

(6) Start de simulatie met de knop **Start simulatie**.

Figuur 3.1. Het hoofdscherm van PollSim



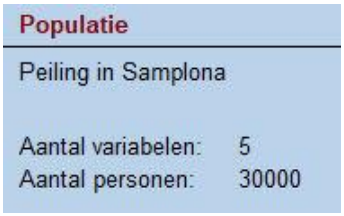
In dit hoofdstuk beschrijven we hoe je een recht-toe-recht-aan simulatie kunt doen, zonder non-respons te genereren en zonder een weging uit te voeren. In hoofdstuk 4 gaan we dieper in op de uitkomsten van een simulatie. Hoofdstuk 5 gaat het over non-respons. We laten met behulp van een simulatie zien hoe non-respons tot verkeerde schattingen kan leiden. In hoofdstuk 6 gaat het over wegen. We laten zien dat wegen soms verkeerde schattingen kan corrigeren.

Lees populatie

Je begint een simulatie altijd met het inlezen van een populatiebestand, Door te klikken op de knop **Lees populatie** krijg je een lijstje te zien van beschikbare bestanden. Dat zijn bestanden met de extensie *pop*. Meer over de structuur van populatiebestanden kun je vinden in hoofdstuk 7. Met deze informatie kun je ook je eigen populatiebestanden maken.

Selecteer het populatiebestand van je keuze. Nadat het programma het geselecteerde bestand heeft ingelezen, verschijnt in het paneel *Populatie* met enige informatie over het bestand. Zie figuur 3.2 voor een voorbeeld.

Figuur 3.2. Het paneel Populatie



Het paneel bevat een korte omschrijving van het bestand ('Peiling in Samplona'). Verder vermeldt het paneel het aantal variabelen in het bestand (hier: 5) en het aantal personen in de doelpopulatie (hier: 30.000).

Steekproef

In het paneel *Steekproef* geef je aan wat voor steekproeven je wilt trekken. Het programma trekt altijd eenvoudige aselecte steekproeven zonder teruglegging en met gelijke kansen. Dus er wordt geloot en iedere persoon in de doelpopulatie heeft dezelfde kans om in de steekproef te komen. In het eerste veld (*Omvang steekproef*) geef je aan hoe groot die steekproeven moeten zijn. Je moet minimaal twee personen trekken. De omvang van de steekproef mag niet groter zijn dan de omvang van de doelpopulatie.

Figuur 3.3 bevat een voorbeeld van een paneel *Steekproef*. Hierin is de omvang van de steekproef op 1.000 gezet. In dit voorbeeld trekken we steekproeven uit de doelpopulatie Samplona. Die bestaat uit 30.000 personen. Dus de omvang van de steekproef moet minsten 2 zijn en mag niet groter zijn de 30.000. Als je een te grote waarde (of een te kleine waarde) opgeeft, dan kan het programma geen simulatie uitvoeren.

Figuur 3.3. Het paneel Steekproef

In het tweede veld van het paneel *Steekproef* geef je aan voor welke variabele in de doelpopulatie je schattingen wilt berekenen. Daarvoor klik je op het pijltje rechts in het veld en vervolgens selecteer je de variabele van je keuze. In het voorbeeld in figuur 3.3 is gekozen voor de variabele NIP (stemt wel of niet op de Nieuwe Internet Partij). Je had ook een andere variabele kunnen kiezen. De volgende variabelen waren beschikbaar: Leeftijd, Herkomst, Internet, NIP en NOP.

Voor de gekozen variabele schat het programma in principe steeds het populatiegemiddelde. Als de variabele echter slechts de waarden 1 of 0 aanneemt (1 als je een bepaalde eigenschap wel hebt en 0 als je die eigenschap niet hebt), dan veronderstelt het programma dat je het *percentage* personen in de doelpopulatie met die eigenschap wilt schatten.

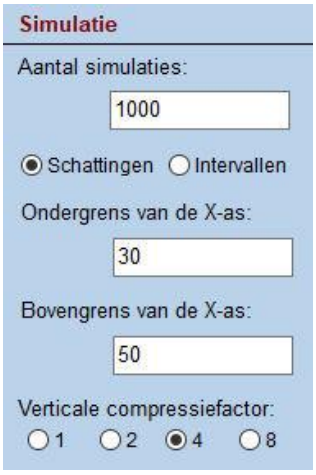
In het voorbeeld is de variabele NIP gekozen. Deze variabele neemt alleen de waarden 1 (stemt op de NIP) en 0 (stemt niet op de NIP) aan. Dus schat het programma steeds het percentage stemmen op de NIP.

Simulatie

Voordat je steekproeven kunt gaan trekken, moet je eerste een aantal instellingen goed zetten. Dat doe je in het paneel *Simulatie*. Zie figuur 3.4 voor een voorbeeld.

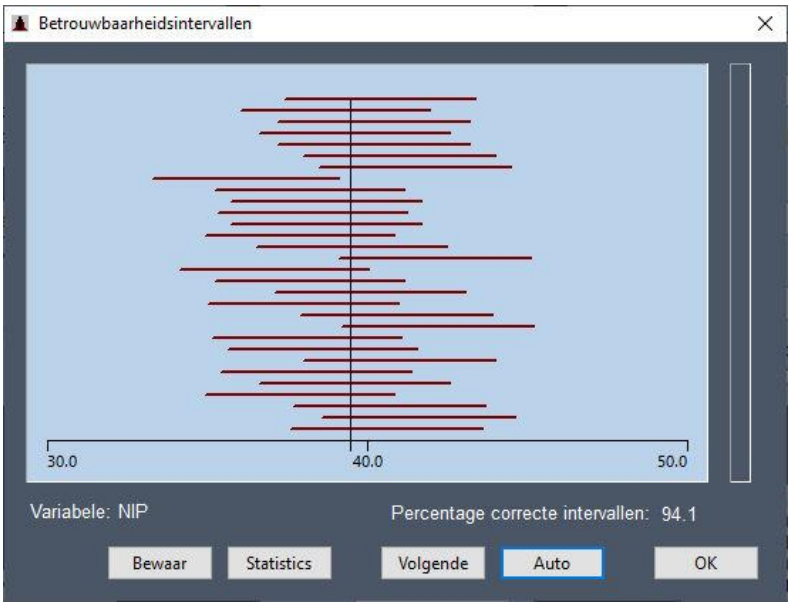
In het eerste veld (*Aantal simulaties*) geef je aan hoeveel steekproeven het programma moet trekken voor een simulatie. Bij het opstarten van het programma staat het aantal simulaties op 1.000. PollSim trekt dus 1.000 steekproeven voor een simulatie. Die 1.000 is meteen ook het maximum aantal simulaties. Het aantal simulaties moet minimaal 2 zijn.

Figuur 3.4. Het paneel Simulatie



Bij de tweede instelling in het panel *Simulatie* kun je aangeven wat voor soort uitvoer de simulatie moet produceren. Je kunt kiezen uit *Schattingen* of *Intervallen*. Kies je voor *Schattingen*, dan maakt het programma een histogram van de verdeling van de schattingen. Figuur 1.1 bevat hiervan een voorbeeld. Kies je voor *Intervallen*, dan maakt het programma voor elke steekproef een *95%-betrouwbaarheidsinterval* voor het populatiegemiddelde of populatiepercentage. Figuur 3.5 bevat hiervan een voorbeeld. De verticale zwarte lijn geeft het te schatten gemiddelde/percentage in de doelpopulatie weer. Elk horizontaal rood lijnstuk stelt een betrouwbaarheidsinterval voor.

Figuur 3.5. Betrouwbaarheidsintervallen (95%)



De interpretatie van een 95%-betrouwbaarheidsinterval is als volgt: Het omvat met een grote waarschijnlijkheid van 95% de werkelijke waarde in de populatie. Of anders gezegd: 19 op de 20 keer bevat het betrouwbaarheidsinterval de werkelijke waarde. Als je kijkt naar het voorbeeld in figuur 3.5, dan zie je daarin de laatste 30 berekende betrouwbaarheidsintervallen van de

simulatie. 29 intervallen bevatten de waarde in de populatie (het rode lijnstuk kruist de zwarte lijn). Dat komt neer op 97%. Dat ligt dus dicht in de buurt van 95%.

Voor deze simulatie zijn in totaal 1.000 betrouwbaarheidsintervallen berekend. In figuur 3.5 kun je aflezen dat 94,1% van die 1.000 intervallen de te schatten waarde bevatten. Dit percentage ligt ook dicht bij de 95%. Het betrouwbaarheidsinterval doet dus wat het belooft.

Met het tweede en derde veld kun je het *waardebereik* van de schattingen instellen. Dat doe je door een ondergrens en een bovengrens voor de X-as in te voeren. Als je niets doet, dan neemt PollSim voor de ondergrens van de X-as de laagst voorkomende waarde van de variabele. Voor de bovengrens gebruikt het programma de hoogst voorkomende waarde. Het zo verkregen interval kan te breed zijn, vooral voor grotere steekproeven. Om het histogram beter te kunnen interpreteren kun je de ondergrens verhogen en de bovengrens verlagen. Een voorbeeld hiervan is te zien in figuur 3.4 en 3.5, waarin de ondergrens gezet is op 30.0 (procent) en de bovengrens op 50.0 (procent).

Als de berekende schattingen voor een variabele dicht bij elkaar in de buurt liggen, kan het gebeuren dat de staven van het histogram te lang worden en daardoor worden afgekapt. Als je dat niet wilt, kun je histogram in verticale zin comprimeren. Daarvoor moet je een *compressiefactor* kiezen. Dat kun je onderaan in het paneel *Simulatie* doen. Een compressiefactor van 1 betekent geen compressie. Bij een compressiefactor van 2 wordt het histogram half zo hoog. De grootste compressie krijg je met een compressiefactor van 8.

Als alle instellingen zijn voltooid, kun je de simulatie (het trekken van steekproeven) starten. Dat doe je met de knop **Start simulatie**. Het uitvoeren van de simulatie en het beschrijven van de uitkomsten doen we in hoofdstuk 4.

4. Het uitvoeren van een simulatie

Het genereren van de steekproeven

In het voorgaande hoofdstuk hebben we de verschillende instellingen van PollSim voor een simulatie beschreven. Nu leggen we uit hoe je de simulatie daadwerkelijk uitvoert en hoe de uitkomsten eruit zien. We gaan daarbij uit van de volgende instellingen:

- De doelpopulatie is Samplona. Het populatiebestand bevat gegevens van 5 variabelen over 30.000 personen.
- Het aantal simulaties is 1,000. Het programma trekt dus 1,000 keer een steekproef.
- We gaan steekproeven van omvang 500 trekken. We schatten het populatiepercentage voor de variabele NIP (Nieuwe Internet Partij). Het werkelijke, te schatten, populatiepercentage is 39,5%.
- We genereren geen non-respons en we voeren geen weging uit.
- We hebben gekozen voor een histogram en niet voor betrouwbaarheidsintervallen.
- De ondergrens van de X-as is gezet op 30.0% en de bovengrens op 70.0%.
- De compressiefactor is op 8 gezet. Dus het histogram wordt maximaal gecomprimeerd.

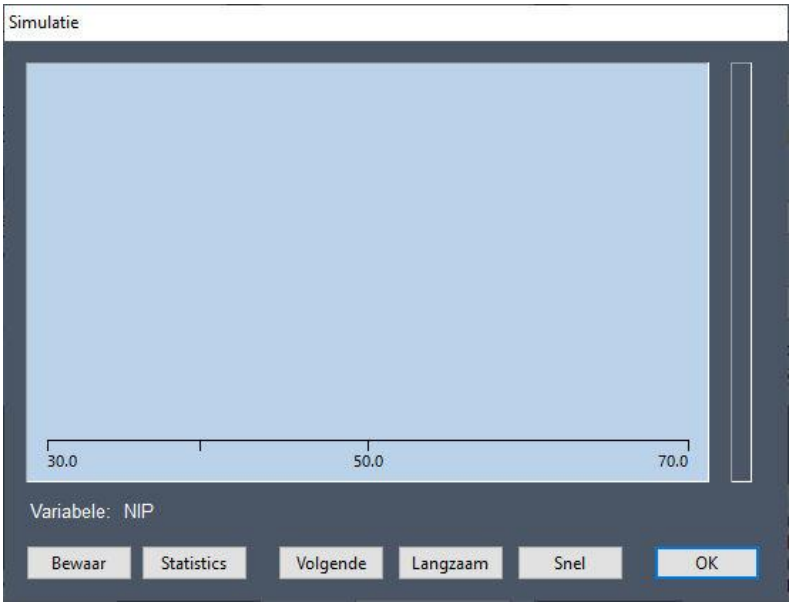
Nadat de simulatie is gestart door klikken op de knop **Start simulatie**, verschijnt het simulatiescherm. Dit is weergegeven in figuur 4.1. In eerste instantie is het scherm leeg. Er is nog geen histogram. Eerst moet je het programma steekproeven laten trekken. Dat kan op drie verschillende manieren.

1) Door klikken op de knop **Volgende** trek je één steekproef. Er wordt een schatting uitgerekend (het percentage stemmers op de NIP in de steekproef). En er verschijnt een blokje op de juiste plaats in het histogram-scherm. Elke keer dat je op de **Volgende** klikt, trekt het programma een nieuwe steekproef, berekent een schatting en voegt het bijbehorende blokje toe aan het histogram. Zo bouw je dus stap voor stap de verdeling van de schattingen op.

2) Door klikken op de knop **Langzaam** trek je in één keer alle steekproeven van de simulatie. Dit gebeurt wel langzaam, zodat je goed kunt zien wat er gebeurt.

3) Door klikken op de knop **Snel** trek je ook in één keer alle steekproeven, maar nu gaat het heel snel. Deze optie is vooral bedoeld voor als je alleen geïnteresseerd in de uitkomsten van de simulatie en niet in het proces van trekken van de steekproeven.

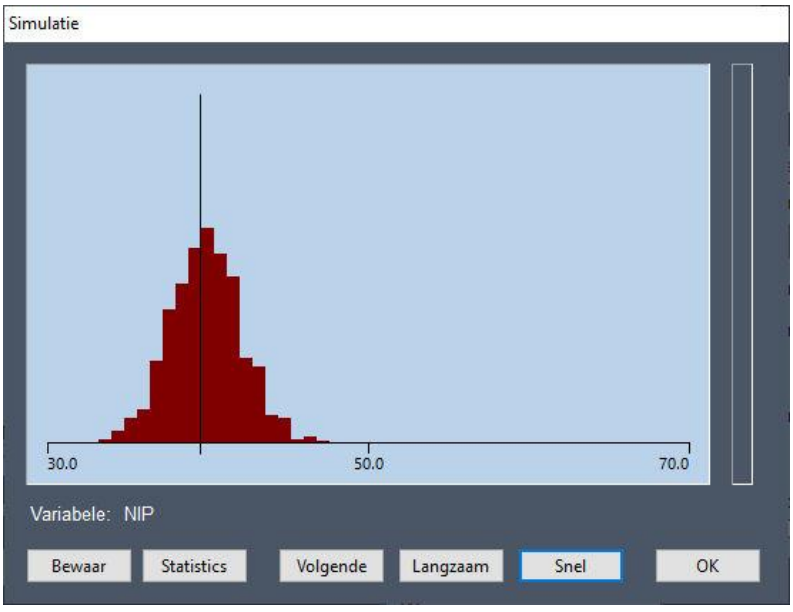
Figuur 4.1. Start van de simulatie



De resultaten

Nadat alle steekproeven zijn getrokken en de simulatie is voltooid, is het scherm gevuld met een histogram van de verdeling van de schattingen. Het ziet er dan bijvoorbeeld zo uit zoals in figuur 4.2.

Figuur 4.2. Einde van de simulatie

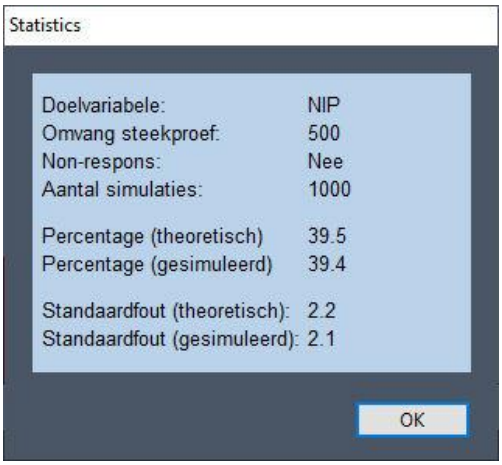


De verticale zwarte lijn geeft de te schatten populatiewaarde aan (39,5%). De rode blokjes zijn de schattingen. Een aantal zaken vallen op in de uitkomsten van deze simulatie. In de eerste

plaats liggen de schattingen keurig geconcentreerd rondom de te schatten waarde. Er is geen sprake van systematische over- of onderschatting. Verder is de verdeling symmetrisch met een top in het midden. Je kunt concluderen dat er bij benadering sprake is van een *normale verdeling*. Dit betekent bijvoorbeeld dat je betrouwbaarheidsintervallen kunt uitrekenen en die geven een goed beeld van de onzekerheidsmarges (ruis) van de schattingen.

Het simulatiescherm bevat nog enkele andere knoppen. In de eerste plaats kun je de knop **Statistics** gebruiken om een numeriek overzicht van de resultaten van de simulatie te maken. Voor de hierboven beschreven simulatie ziet dit overzicht eruit als in figuur 4.3.

Figuur 4.3. Het paneel Statistics



Het eerste blokje tekst bevat informatie over de opzet van de simulatie. De doelvariabele is NIP. Dus het programma heeft het percentage stemmers op de NIP geschat. Dit is gebeurd met steekproeven van omvang 500. Er is geen non-respons gegenereerd. En er zijn 1.000 steekproeven getrokken.

Het tweede blokje tekst meldt dat het te schatten percentage in de populatie (het theoretische gemiddelde) gelijk is aan 39,5%. Het gesimuleerde gemiddelde is het gemiddelde van alle 1.000 schattingen in de simulatie. In dit geval is het gesimuleerde percentage gelijk aan 39,4%. Dat is dus praktisch gelijk aan het theoretische percentage. Je kunt de conclusie trekken dat de steekproeven goed in staat waren om het populatiepercentage te schatten.

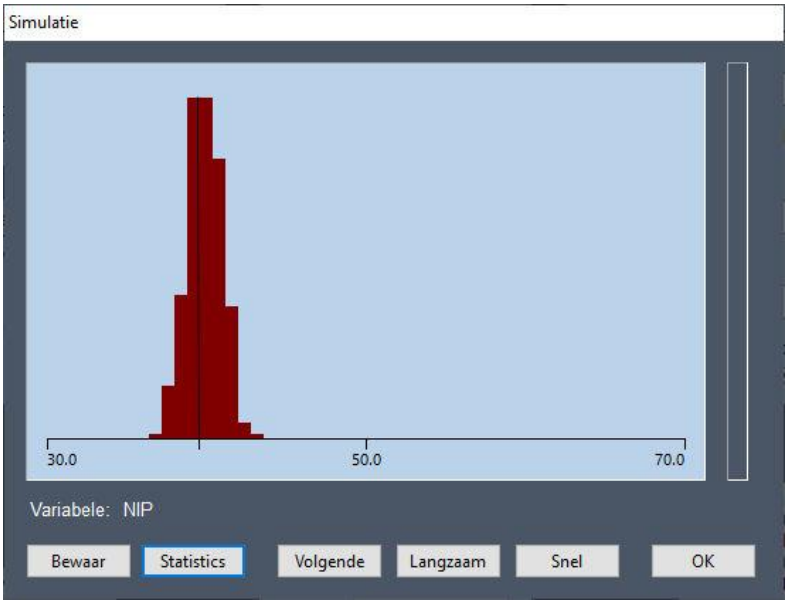
De laatste twee regels in figuur 4.3 geven informatie over de standaardfout. De *standaardfout* is een maat voor de precisie van de schattingen. Naarmate de schattingen dichter bij elkaar in de buurt liggen is de precisie groter en de standaardfout kleiner. Volgens de steekproeftheorie zou de standaardfout gelijk moeten zijn aan 2,2. De schatting van de standaardfout op basis van de 1.000 steekproeven kwam uit op 2,1. De theoretische en de gesimuleerde standaardfout zijn vrijwel gelijk aan elkaar. De simulatie werkte dus naar behoren.

Linksonder op het simulatiescherm kun je nog de knop **Bewaar** vinden. Klikken op die knop zorgt ervoor dat het programma het histogram opslaat als een grafisch bestand. Het is een bitmap-bestand met extensie *bmp*. Je kunt dit plaatje eventueel importeren in een ander bestand.

Het effect van de steekproefomvang

Je kunt (theoretisch) aantonen de schattingen voor een steekproef zitten naarmate de omvang van de steekproef groter is. Dit kun je ook laten zien met het programma PollSim. We gaan uit van de simulatie zoals die hiervoor is beschreven. Zie figuren 4.2 en 4.3. Vervolgens passen we de omvang van de steekproef aan. Die zetten we op 2.000 in plaats van 500. De steekproeven zijn dus vier keer zo groot. We doen weer een simulatie op basis van 1.000 steekproeven. Het resultaat is te zien in figuur 4.4

Figuur 4.4. Simulatie met een steekproefomvang van 2.000.



Er is weer sprake van een keurige symmetrische verdeling rondom het te schatten populatiepercentage. Alleen is nu de spreiding van de schattingen veel kleiner. De schattingen liggen veel dichter in de buurt van de populatiewaarde. De schatttig zijn zuiver en hebben een grotere precisie. Uit de *Statistics* blijkt dat de standaardafwijking is nu 1,1 in plaats van 2,2. De schattingen zijn dus twee keer zo precies.

5. Non-respons

Het probleem van de non-respons

Als je een aselecte steekproef trekt, kun je valide (zuivere) schattingen maken van populatiekenmerken zoals het populatiegemiddelde en het populatiepercentage. Bovendien kun je betrouwbaarheidsintervallen uitrekenen. Die geven aan hoe groot het verschil tussen schatting en werkelijkheid maximaal kan zijn. Een aselecte steekproef levert dus valide uitkomsten op die je kunt generaliseren van steekproef naar doelpopulatie.

Helaas is het in de praktijk niet allemaal rozengeur en maneschijn. Er kunnen zich altijd problemen voordoen die de correctheid van de uitkomsten aantasten. Een van de belangrijkste problemen is non-respons. *Non-respons* is het verschijnsel dat er personen in de steekproef zitten die de gewenste antwoorden op de vragen niet willen of kunnen geven.

Non-respons kan de uitkomsten van een peiling ernstig aantasten. Daarom moet je proberen om non-respons in het veld zoveel mogelijk tegen te gaan. Maar helaas, hoe je ook je best doet, er blijft altijd non-respons over. En als het dan niet mogelijk is om non-respons te vermijden, dan zal je andere maatregelen moeten nemen om de vertekening in de uitkomsten weg te werken of ze ten minste te verminderen.

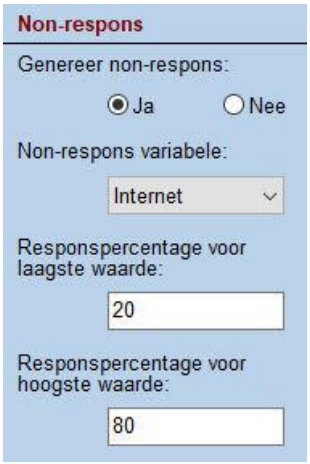
Non-respons in PollSim

Om de effecten van non-respons te bestuderen, kun je PollSim non-respons laten genereren in de getrokken steekproeven. Je kunt dan zien hoe non-respons de schattingen aantast. Dat vereist het aanpassen van een aantal instellingen. Dat doe je in het paneel *Non-respons*. Zie figuur 5.1. voor een voorbeeld.

De eerste stap is dat je op het keuzerondje *Ja* klikt. Daarmee geef je het programma opdracht om non-respons te genereren.

De tweede stap is dat je een *non-respons-variabele* moet kiezen. Die variabele selecteer je in het lijstje nadat je het hebt opengeklapt. De kans op non-respons van een persoon hangt af van de waarde van de non-respons-variabele. In het voorbeeld is de variabele *Internet* gekozen. Dat betekent dat de kans op non-respons afhangt van het wel of niet hebben van Internet.

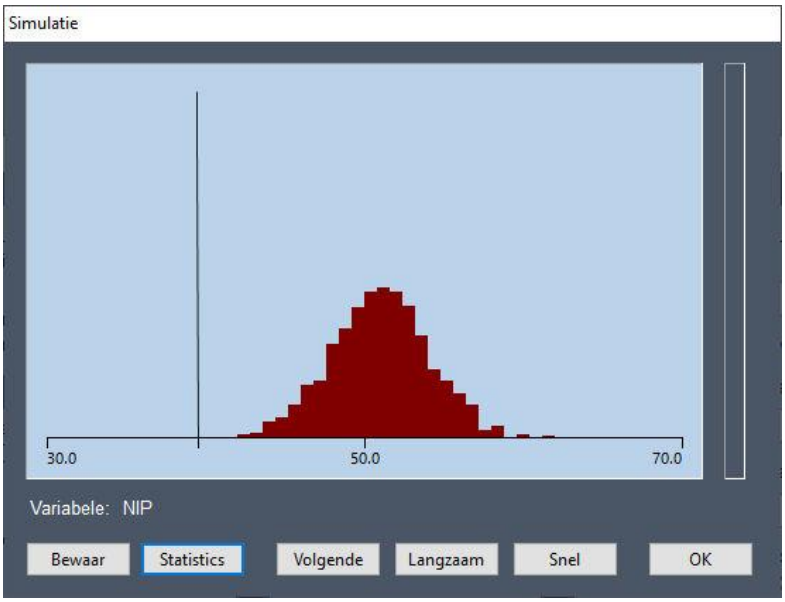
Figuur 5.1. Het panel Non-respons



Met de twee invoervelden geef je aan hoe de non-respons afhangt van de waarde van de non-respons-variabele. In het eerste veld zet je het percentage respons voor de laagste waarde van de non-respons-variabele. In het voorbeeld kan de non-respons-variabele maar twee verschillende waarden aannemen: 0 (heeft geen internet) en 1 (heeft wel internet). De laagste waarde is dus 0. Voor personen met die waarde is het percentage respons gelijk aan 20%. Dus de kans op respons is klein.

In het tweede veld zet je het percentage respons voor de hoogste waarde van de non-respons-variabele. In het voorbeeld is die hoogste waarde gelijk aan 1. Voor personen met internet genereert het programma dus 80% respons. Voor de instellingen in figuur 5.1 is de kans op respons dus hoog (80%) voor personen met internet en laag voor personen zonder internet (20%).

Figuur 5.2. Het resultaat van een simulatie met non-respons

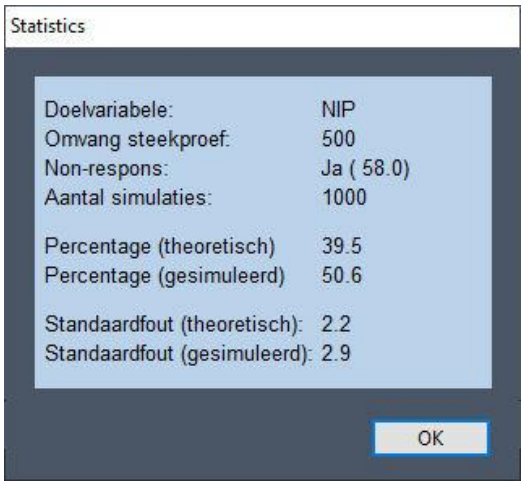


Je start de simulatie (met non-respons) op de gebruikelijke manier door klikken op de knop **Start simulatie**. Figuur 5.2 toont een voorbeeld van de uitkomst van de simulatie als je de

instellingen van figuur 5.1 gebruikt. Het gaat weer om het schatten van het percentage stemmers op de NIP met steekproeven van omvang 500.

De verticale zwarte lijn in figuur 5.2 geeft het te schatten percentage in de populatie (39,5%) aan. Het is duidelijk dat alle schattingen er behoorlijk naast zitten. Door het optreden van non-respons zijn al die schattingen veel te hoog uitgevallen. De hele verdeling van de schattingen is naar rechts geschoven. Er is geen sprake meer van een valide peiling. De schattingen zijn niet meer zuiver. Je kunt nog wat meer inzicht in de vertekening krijgen door naar de *Statistics* te kijken. Klik hiervoor op de knop **Statistics**. Figuur 5.3 bevat een voorbeeld. De derde regel van dit overzicht bevat (tussen haakjes) het percentage respons. Dit percentage was hier dus 58,0%.

Figuur 5.3. Statistics voor een simulatie met non-respons

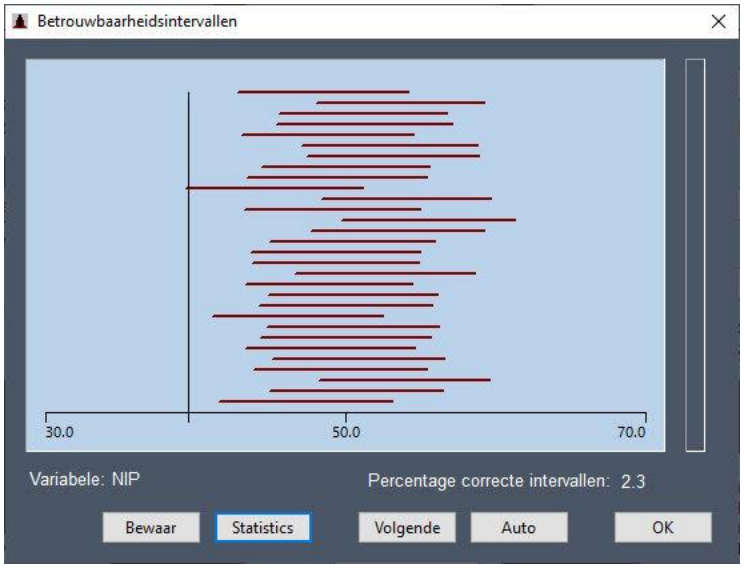


Non-respons en betrouwbaarheidsintervallen

Als er non-respons optreedt, kunnen de schattingen een vertekening hebben. Dat is duidelijk te zien in het paneel *Statistics*. Het theoretische te schatten populatiepercentage, is gelijk aan 39,5%. De steekproeven in de simulatie komen echter gemiddeld uit op 50,6%. Dat betekent dat er een vertekening is van 50,6 – 39,5 = 11,1 procentpunten.

Merk ook op dat de standaardfout van de simulatie (2.9) groter is dan die van de theoretische standaardfout (2.2). Ook dit wordt veroorzaakt door non-respons. Er zijn immers minder waarnemingen beschikbaar gekomen. Dit leidt tot een grotere standaardafwijking groter.

Figuur 5.4. Simulatie van betrouwbaarheidsintervallen met non-respons



Non-respons maakt ook dat je het betrouwbaarheidsinterval niet meer kunt gebruiken als indicatie voor de precisie van je schattingen. Heb je geen non-respons dan ligt de werkelijke populatiewaarde met grote waarschijnlijkheid (95%) in het 95%-betrouwbaarheidsinterval. Die uitspraak gaat niet meer op als je te maken krijgt met non-respons. Dat kun je goed zien in figuur 5.4. Het gaat weer om het schatten van het percentage NIP-stemmers. Er is non-respons gegenereerd met de instellingen van figuur 5.1. In plaats van een histogram, zijn de laatste 30 betrouwbaarheidsintervallen getekend. Slechts één van die 30 intervallen bevat de te schatten waarde. Alle andere 29 intervallen zitten ernaast. De betrouwbaarheid van de intervallen is niet meer gelijk aan 95%, maar slechts 2,3%.

6. Wegen

Over wegen

Het optreden van non-respons kan leiden tot een selectieve respons, en dus kunnen schattingen vertekend zijn. Ze zijn dan niet meer valide. Als je te maken krijgt met non-respons, dan moet je daaraan iets doen. Er is een correctie nodig. De meest gebruikte correctietechniek is *wegen*. Hierbij ken je aan elke respondent een gewicht toe. Als een respondent deel uitmaakt van een ondervertegenwoordigde groep, dan moet het gewicht groter dan 1 zijn. En als de respondent in een oververtegenwoordigde groep zit, dan moet het gewicht kleiner dan 1 zijn.

Om gewichten te kunnen berekenen, heb je *hulpvariabelen* nodig. Die hulpvariabelen moet je hebben gemeten in de peiling en bovendien moet de verdeling van die hulpvariabele in de doelpopulatie beschikbaar zijn. Wegen is alleen effectief voor het reduceren van de vertekening als de hulpvariabelen aan twee voorwaarden voldoen:

1) De hulpvariabelen moeten gecorreleerd zijn met de doelvariabele. Als er geen verband is tussen doelvariabele en hulpvariabelen, dan helpt wegen met deze variabelen niet.
2) De hulpvariabelen moeten gecorreleerd zijn met het responsgedrag. Als er geen verband tussen hulpvariabelen en responsgedrag is, dan helpt wegen met deze variabelen niet.

De achterliggende gedachte bij wegen is dat je moet proberen de respons van de peiling representatief te maken met betrekking tot de hulpvariabelen. Daarvoor bereken je gewichten. De waarden van de gewichten moeten zo zijn dat de gewogen verdeling van elke hulpvariabele in de respons gelijk is aan de corresponderende verdeling in de doelpopulatie. Als het mogelijk is de respons representatief te maken met betrekking tot een reeks hulpvariabelen, en al die hulpvariabelen zijn gecorreleerd met de doelvariabelen en het responsgedrag, dan zal de gewogen respons (bij benadering) ook representatief zijn met betrekking tot de doelvariabelen. Daarom zullen schattingen die gebaseerd zijn op de gewogen respons beter zijn dan schattingen gebaseerd op de ongewogen respons.

Berekenen van gewichten

Figuur 6.1 bevat een voorbeeld van het berekeningen van gewichten. De gewichten zijn gebaseerd op de hulpvariabele *Leeftijd*. We veronderstellen dat de verdeling in de populatie van deze variabele bekend is. Zie de tweede kolom in de tabel in figuur 6.1. De derde kolom bevat van de verdeling van de variabele *Leeftijd* in de steekproef. Die verdeling is gebaseerd op 1.000 simulaties van een steekproef van 500 personen. We hebben weer non-respons gegenereerd zoals aangegeven in figuur 5.1.

Figuur 6.1. Het berekenen van gewichten met de variabele Leeftijd

Leeftijd	Populatie	Steekproef	Gewicht
Jong	39,8%	46,1%	0,863
Middelbaar	35,3%	34,5%	1,023
Oud	24,9%	19,4%	1,284
Totaal	100,0%	100,0%	

Duidelijk is te zien dat de jongeren met 46,1% oververtegenwoordigd zijn in de steekproeven. Het percentage jongeren in de populatie is met 39,8% veel kleiner. Om hiervoor te corrigeren berekenen we een gewicht voor jongeren. Dat krijgen we door het percentage jongeren in de populatie te delen door het percentage jongeren in de steekproef. Het resultaat is een gewicht van 39,8 / 46,1 = 0,863. Het gewicht voor de jongeren is dus kleiner dan 1. Dat is logisch want er zaten teveel jongeren in de steekproeven.

De gewichten voor de personen van middelbare leeftijd en die voor de ouderen kun je op dezelfde manier berekenen. Merk op dat er te weinig ouderen in de steekproeven zitten. Daarom is hun gewicht (1,284) groter dan 1.

In het populatiebestand Samplona zit nog een andere potentiële weegvariabele. Dat is de variabele *Internet*. Die heeft twee mogelijke waarden: *Ja* (heeft internet) en *Nee* (heeft geen internet). Ook deze hulpvariabele zit in het populatiebestand. Dus je zou hem kunnen gebruiken om te wegen. In figuur 6.2 vergelijken we verdeling in de populatie van deze variabele met de (gemiddelde) verdeling in de gesimuleerde steekproeven (met non-respons zoals in figuur 5.1).

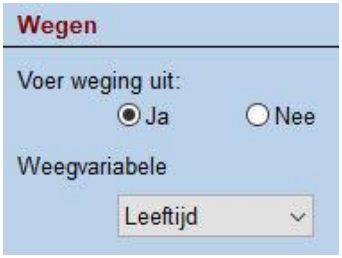
Figuur 6.2. Het berekenen van gewichten met de variabele Leeftijd

Internet	Populatie	Steekproef	Gewicht
Heeft internet	63,4%	87,4%	0,725
Heeft geen internet	36,6%	12,6%	2,905
Totaal	100,0%	100,0%	

Personen met internet zijn behoorlijk oververtegenwoordigd in de steekproeven. In de populatie heeft 63,4% internet terwijl 87,4% in de gesimuleerde steekproeven internet heeft. We moeten dus een stevige correctie uitvoeren. Personen met internet krijgen een gewicht 0,725. Dat is veel kleiner dan 1, want er zitten teveel internetbezitters in de steekproeven.

Je kunt PollSim een weging laten uitvoeren op gesimuleerde steekproeven die zijn aangetast door non-respons. Daarvoor moet je de instellingen van het programma aanpassen. Dat doe je in het paneel *Wegen*. Zie ook figuur 6.3. Eerst moet je klikken op het keuzerondje *Ja* om wegen aan te zetten. Als je wegen weer uit wilt zetten, dan klik je op het keuzerondje *Nee*.

Figuur 6.3. Het panel Wegen



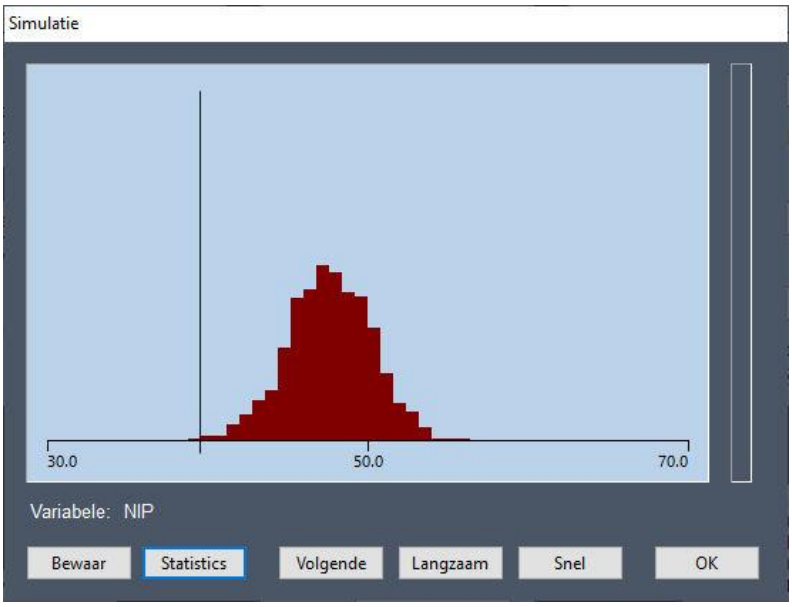
Om te kunnen wegen, heb je hulpvariabelen nodig. Liefst zoveel mogelijk. PollSim beperkt zich tot wegen met één hulpvariabele. En die variabele moet in het populatiebestand zitten om de verdeling in de populatie te kunnen bepalen. In figuur 6.3 is gekozen voor de hulpvariabele *Leeftijd*.

Merk op dat je alleen categorische variabelen voor wegen kunt gebruiken. De waarden van zo'n variabele moeten opeenvolgende nummers zijn (1, 2, …) zijn. Het nummer van de eerste categorie moet 1 zijn. Het programma kan maximaal 10 categorieën aan. Een voorbeeld is de hulpvariabele *Leeftijd* met drie categorieën 1 (jong), 2 (middelbaar) en 3 (oud).

Simulatie met non-respons en wegen

Figuur 6.2. bevat een voorbeeld van de uitkomsten van een simulatie waarin is gewogen met de variabele *Leeftijd*. Als je deze uitkomsten vergelijkt met de ongewogen uitkomsten in figuur 5.2, dan zie je dat de verdeling weliswaar wat is opgeschoven in de richting van het populatiepercentage, maar er is nog steeds sprake van een vertekening. De schattingen vallen nog steeds te hoog uit.

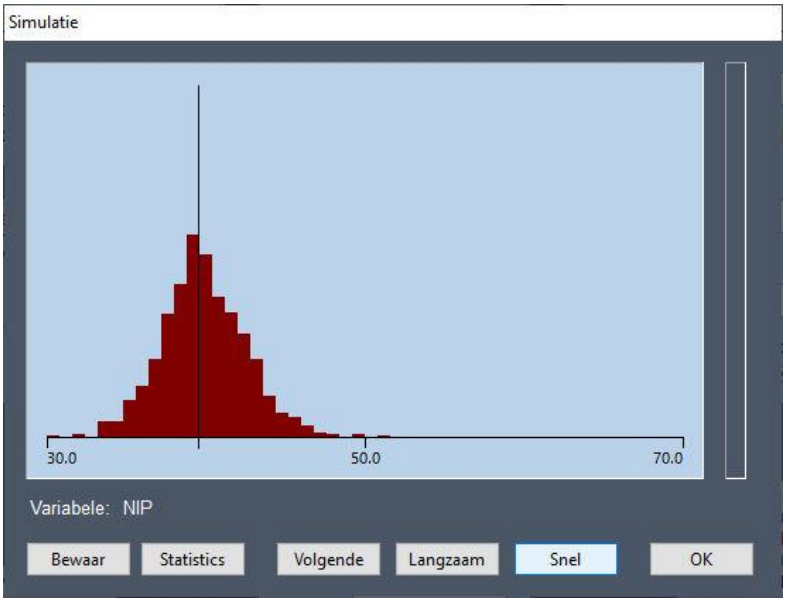
Figuur 6.4. Wegen met de variabele Leeftijd



Als je de *Statistics* bekijkt voor de simulatie in figuur 6.4, dan blijkt dat de schattingen gemiddeld gelijk zijn aan 47,2%. Dat is veel hoger dan het populatiepercentage (39,5%). En dat is maar een klein beetje lager dan het ongewogen percentage (50,6%).

Je kunt ervoor kiezen om te wegen met de variabele *Internet* in plaats van *Leeftijd*. Figuur 6.5 laat de uitkomsten zien van een simulatie waarin eerst non-respons is gegenereerd en vervolgens is gewogen met Internet.

Figuur 6.5. Wegen met de variabele Internet



Deze weging is succesvol. De gewogen verdeling is symmetrisch om de te schatten populatiewaarde. De vertekening is verdwenen. De schattingen zijn valide (zuiver). In de tabel in figuur 6.6 kun je resultaten van de wegingen nog eens met elkaar vergelijken en met de ongewogen simulatie.

Figuur 6.6. Het effect van wegen

Weging	Schatting	Populatie	Vertekening
Geen weging	50,6%	39.5%	11,1%
Wegen met *Leeftijd*	47,2%	39,5%	7,7%
Wegen met *Internet*	39,6%	39,5%	0,1%
Totaal	100,0%	100,0%	

Het is duidelijk dat non-respons zonder weging leidt tot uitkomsten met een substantiële vertekening van 11,1 procentpunten. Wegen met *Leeftijd* vermindert de vertekening wel wat (van 11,1 naar 7,7 procentpunten) maar er blijft toch nog veel vertekening over. Wegen met *Internet* is wel effectief. Deze weging is in staat de vertekening te elimineren.

Waarom werkt wegen met *Leeftijd* niet en wegen met *Internet* wel? Dat komt omdat het non-respons-mechanisme zo werkt dat de kans op respons afhangt van het wel of niet hebben van internet. Immers, personen zonder internet hebben een kleine kans op respons en mensen met internet hebben een grote kans op respons. Als je weegt met *Internet*, dan corrigeer je voor de selectiviteit van dit mechanisme.

Wegen met *Leeftijd* werkt veel minder goed. Er is wel enig verband tussen respons/non-respons en *Leeftijd* (meer jongeren hebben internet en minder ouderen) maar dit verband is niet zo sterk. Daarom is wegen niet zo effectief.

7. Overige zaken

Over PollSim

Op het hoofdscherm van het programma PollSim kun je twee knoppen vinden die we nog niet hebben genoemd. De eerste knop is de knop **Over PollSim**. Door te klikken op deze knop krijg

je enige informatie over het programma. Het versienummer is nuttig om te vermelden als je problemen met het programma wilt rapporteren. Wil je contact opnemen, dan kun je hier ook de naam en het e-mailadres van de maker van het programma vinden.

Figuur 7.1. Over PollSim

Toon variabele

Met de knop **Toon variabele** kun je enige informatie krijgen over de verdeling van een variabele in de populatie. Na klikken op de knop verschijnt een lijstje met alle beschikbare variabelen. Selecteer daarin de variabele waarover je meer informatie wilt hebben. Vervolgens krijg je een deels grafisch en deels numeriek overzicht. Figuur 7.2 bevat een voorbeeld.

Figure 7.2.. De verdeling in de populatie van de variabele Leeftijd.

Boven in het overzicht staat een simpel histogram waarin de mogelijke waarden van de variabele te zien is. In figuur 7.2 is bijvoorbeeld te zien dat de variabele Leeftijd drie mogelijk waarde kan aannemen: 1 (voor jongeren), 2 (voor personen van middelbare leeftijd) en 3 (voor ouderen). Het histogram geeft ook een globaal beeld van hoe vaak de verschillende waarden voor komen.

Onderin het overzicht staat een aantal numerieke indicatoren: de minimale en de maximale waarde, de gemiddelde waarde en de standaardafwijking.

Populatiebestanden

Het programma PollSim gebruikt populatiebestanden. Dat zijn bestanden die alle waarden in de populatie bevatten voor een reeks variabelen. Voor een simulatie trekken we de steekproeven uit deze bestanden. Populatiebestanden moeten een bepaalde structuur hebben. Die structuur beschrijven we hieronder. De naam van een populatiebestand moet altijd de extensie *pop* hebben.

Bij het programma zit een voorbeeld van een populatiebestand. Dit is het bestand *Samplona.pop*. Het bevat informatie over het stemgedrag van alle 30.000 stemgerechtigden in de gemeente Samplona. De volgende variabelen zitten in dit bestand:

- *Leeftijd* in drie categorieën: 1 (Jong), 2 (Middelbaar) en 3 (Oud).
- *Herkomst* in twee categorieën 1 (Nederlandse achtergrond) en 0 (Migratieachtergrond).
- *Internet* in twee categorieën: 1 (heeft internet) en 0 (heeft geen internet).
- *NIP* (Stemt op de Nieuwe Internet Partij) in twee categorieën: 1 (Ja) en 0 (Nee).
- *NOP* (Stemt op de Nationale Ouderen Partij) in twee categorieën: 1 (Ja) en 0 (Nee).

Je kunt eventueel zelf een populatiebestand maken. Daarbij gelden de volgende beperkingen: (1) er mogen niet meer dan 30.000 personen in het bestand zitten, (2) het aantal variabelen is beperkt tot 10 en (3) een weegvariabele mag niet meer dan 10 categorieën hebben.

Figuur 7.3 bevat een voorbeeld van het eerste stuk van een populatiebestand. Het is het bestand *Samplona.pop*.

Figuur 7.3. Het populatiebestand Samplona

```
Peiling in Samplona
30000 5
Leeftijd
Herkomst|D
Internet|D
NIP|D
NOP|D
1 1 1 1 0
1 1 1 1 0
1 1 1 1 0
1 1 1 1 0
3 1 0 0 1
2 1 0 0 1
1 1 1 1 0
3 1 1 1 0
3 1 0 0 0
2 0 0 1 0
...
```

De eerste regel van het bestand moet een korte omschrijving van het bestand bevatten. Dat is in dit geval de tekst ‘Peiling in Samplona’). In de tweede regel geef je aan hoe groot de populatie is (hier: 30.000) en hoeveel variabelen er in het bestand zitten (hier: 5).

De volgende regels moeten de namen van de variabelen bevatten. Dat zijn hier dus vijf namen. Om aan te geven dat een variabele alleen de waarden 0 en 1 aanneemt, en dus dat het programma percentages moet schatten, voeg je |D aan de naam toe. D staat voor dummy. In het voorbeeld zijn vier van de vijf variabelen dummy-variabelen (0/1-variabelen).

Na de namen van de variabelen komen de gegevens. Elke persoon in de populatie begint op een nieuwe regel. De waarden van de variabelen moet je van elkaar scheiden door één of meer

spaties. Het voorbeeld-bestand in figuur 7.3 bevat vijf waarden per persoon, corresponderend met de vijf variabelen.