

Data in onderzoek: Steekproef en sampling uit een populatie.

Wat is dataverzameling voor onderzoek?

Dataverzameling is het systematisch en gestructureerd verzamelen van gegevens, met als doel betrouwbare analyses uit te voeren die het mogelijk maken om de vooropgestelde onderzoeksvragen te beantwoorden.

Steekproef vs. proefopzet

Het belangrijkste verschil tussen een steekproef (sampling) en een experimenteel design (proefopzet) is dat een steekproef bedoeld is om personen, objecten, planten etc. te selecteren die de populatie representeren, terwijl een experimenteel design gericht is op het bewust veranderen of toepassen van een behandeling om oorzaak-gevolgrelaties te onderzoeken.

Basisbegrippen

Populatie

De populatie is de volledige groep personen, objecten of observaties waarover de onderzoeker uitspraken wil doen. Bv. studenten van een universiteit, klanten van een bedrijf, bomen in een bos, bijen in een gebied. De grootte van de populatie noteren we als **N**.

Steekproef

Een steekproef is een kleinere groep die geselecteerd wordt uit de populatie om te onderzoeken. Een steekproef is m.a.w. een deelverzameling van een populatie. De grootte van een steekproef noteren we als **n**.

Parameter

Een parameter is een numeriek kenmerk van de populatie. Dit is de "werkelijke" waarde, deze waarde is vast, maar meestal onbekend.

- Gemiddelde = μ
- Standaardafwijking = σ
- Correlatiecoëfficiënt = r

Statistiek

Een statistiek is een kenmerk van de steekproef. Deze worden vaak gebruikt als schatter van een onbekende populatieparameter.

- Steekproefgemiddelde = $\bar{x}$
- Steekproefstandaardafwijking = s
- Steekproefcorrelatiecoëfficiënt = r

Wat bepaalt steekproefgrootte (sample size, SS)?

Variabiliteit (σ / p): een grotere spreiding → een grotere SS.

Foutenmarge (E): een kleinere gewenste foutenmarge → een grotere SS.

Betrouwbaarheidsniveau (confidence level, CL): een hoger gewenst betrouwbaarheidsniveau → een grotere SS.

Populatiegrootte (N): N heeft in het algemeen weinig invloed op de SS. Enkel wanneer de populatie klein is, kan er invloed zijn.

Bereken steekproefgrootte bij continue data

Deze formule bepaalt de steekproefgrootte om een gemiddelde te schatten met een gekozen foutenmarge en betrouwbaarheidsniveau, uitgaande van een gekende (of geschatte) populatiestandaarddeviatie.

$$n = \left(\frac{Z \cdot \sigma}{E} \right)^2$$

met

- n = steekproefgrootte
- Z = Z-score bij gekozen betrouwbaarheidsniveau
- σ = populatiestandaardafwijking
- E = gewenste foutenmarge (precision)

Voorbeeld: Een fabrikant van pindakaas wil het suikergehalte schatten.

$\sigma = 10$ g

$E = 4$ g

Betrouwbaarheidsniveau = 95% → $Z = 1,96$

$$n = \left(\frac{1,96 \cdot 10}{4} \right)^2 \approx 24$$

⇒ Minimaal 24 potten meten.

- $SS \geq 30$ → grote steekproef → Z-verdeling geldig
- $SS < 30$ → kleine steekproef? t-, F- of chi-square verdeling gebruiken

Bereken steekproefgrootte bij proporties

$$n = \frac{Z^2 \cdot p(1-p)}{E^2}$$

Z = Z-waarde bij gekozen betrouwbaarheidsniveau (bv. 1,96 voor 95%)
 p = geschatte proportie (gebruik 0,5 indien onbekend)
 E = gewenste foutenmarge (bijv. 0,05)

Voorbeeld: Stel je wil het percentage afgekeurde producten schatten.

Geschatte proportie $p = 0,5$

Betrouwbaarheidsniveau 95% → $Z = 1,96$

Foutenmarge $E = 0,05$

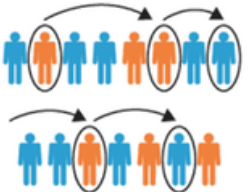
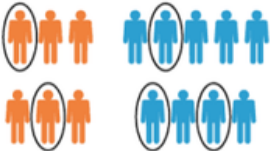
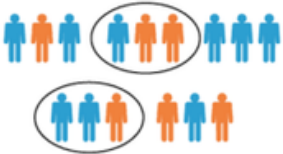
$$n = \frac{(1,96)^2 \cdot 0,5 \cdot (1 - 0,5)}{(0,05)^2} = \frac{3,8416 \cdot 0,25}{0,0025} = \frac{0,9604}{0,0025} \approx 384,16$$

⇒ Minimaal $n \approx 385$ respondenten of eenheden. Kleine populaties kunnen een correctie voor eindige populatie nodig hebben:

$$n_{\text{adjusted}} = \frac{n}{1 + \frac{n-1}{N}}$$

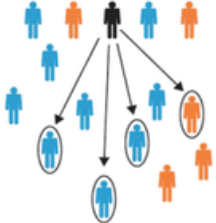
Samplingsmethodes: Probability sampling.

Probability sampling zijn steekproefmethodes waarbij elke eenheid in de populatie een bekende kans heeft om geselecteerd te worden, zodat de steekproef representatief is en statistische inferentie naar de gehele populatie mogelijk is. Al kan een specifieke steekproef door toeval toch minder representatief uitvallen

Sampling methode	Beschrijving	Illustratie
Aselecte random steekproef	Je neemt random samples, zonder patroon. Hier heeft elk object (plant, dier, mens, staal) dezelfde kans om geselecteerd te worden. Vb. Je hebt 1.000 maïs-planten. Je selecteert willekeurig 50 planten.	
Systematische steekproef	Je kiest elke k-de object volgens een vast interval. Vb. elke 20e plant in een rij. Risico als er een patroon de populatie zit, enkel voor homogene populaties	
Gestratificeerde steekproef	Je verdeelt de populatie in homogene deelgroepen (strata) en neemt daarna binnen elk stratum een random steekproef. Vb. perceel bestaat uit zand en klei, je neemt steekproeven in beide bodemtypes. Of vb. mannen vs. vrouwen, leeftijdscategorieën. Je neemt afzonderlijk steekproeven bij mannen en vrouwen.	
Geclusterde steekproef	Je verdeelt de populatie in clusters op basis van een bepaalde variabele en selecteert willekeurig enkele clusters. Vb. Een bos is verdeeld in 20 proefvlakken. Je kiest 5 proefvlakken en meet daar alle bomen.	

Samplingsmethodes: Non-probability sampling.

Non-probability sampling zijn steekproefmethodes waarbij niet elke eenheid een bekende kans heeft om geselecteerd te worden, waardoor de steekproef niet altijd representatief is en resultaten mogelijk bias bevatten

Sampling methode	Beschrijving	Illustratie
Convenience Sampling (Gemakssteekproef)	Je neemt een sample dat makkelijk toegankelijk is, i.e. de eenheden die dichtbij/eenvoudig te bereiken zijn. Nadeel: kan een vertekend beeld geven. Vb. Je hebt 1.000 maïsplanten. Je selecteert de eerste 50 planten	
Voluntary Response Sampling (Vrijwillige respons)	Deelnemers melden zich vrijwillig aan en worden niet gekozen door de onderzoeker. Nadeel: kan bias bevatten, mensen die een sterke mening hebben doen mee.	
Purposive / Judgmental Sampling (Doelgerichte steekproef)	Je neemt een sample op basis van expertise/kennis. Nadeel: kwetsbaar voor bias en subjectieve selectie. Vb. Je wilt het effect van een zeldzame bemestingsmethode meten. Je kiest de velden waar de methode correct is toegepast.	
Snowball Sampling (Ketting- of sneeuwbal-steekproef)	Gebruikt wanneer deelnemers moeilijk te vinden of zeldzaam zijn. Begin met een paar deelnemers die andere deelnemers aandragen. Nadeel: kan leiden tot clusters en bias, omdat deelnemers vaak uit dezelfde groep komen.	