Wat zijn data?

1 Wat zijn data?

Een onderzoek start met het bepalen van welke kenmerken je precies wilt onderzoeken. Zo een kenmerk noemen we een *variabele*. Vervolgens verzamel je *data* door metingen van dat kenmerk uit te voeren.

Definitie 1. Een **variabele** is een kenmerk dat tijdens een onderzoek gemeten of geobserveerd wordt.

De **data** of **gegevens** zijn de resultaten van het meten of observeren van een variabele. We noteren deze vaak als:

$$x_1, x_2, x_3, \ldots, x_n$$

waarbij elke x_i een individuele meting is. De index i loopt van 1 tot en met n, waarbij n het totaal aantal metingen voor die variabele voorstelt.

Als je een variabele meet bij n verschillende eenheden, dan vormen alle x_i samen een **dataset** of **gegevensverzameling**.

Voorbeeld 1. We kunnen als variabele de leeftijd van studenten Humane Wetenschappen in de eerste bachelor onderzoeken. We bevragen de leeftijd van acht studenten en verkrijgen de volgende dataset:

$$x_1 = 18, \quad x_2 = 19, \quad x_3 = 17, \quad x_4 = 20, \quad x_5 = 18, \quad x_6 = 19, \quad x_7 = 18, \quad x_8 = 18$$

Dit is een dataset van $n = 8$ waarnemingen voor de variabele *leeftijd*.

2 Welke soorten data zijn er?

In een onderzoek verzamel je data, maar niet alle data zijn van hetzelfde type. Sommige gegevens zijn getallen waarmee je kunt rekenen, andere zijn categorieën die iets zeggen over voorkeuren, groepen of rangordes. Het is belangrijk om het verschil te kennen tussen deze soorten data, zodat je ze correct kunt analyseren.

Definitie 2. Data kunnen we onderverdelen in twee grote groepen:

- **Kwantitatieve data** zijn numerieke gegevens waarmee je zinvol kunt rekenen. Ze zijn
  - **continu** als elke waarde binnen een interval (of meerdere intervallen) mogelijk is;
  - **discreet** als in elk interval van eindige lengte slechts een eindig aantal waarden mogelijk zijn.
- **Kwalitatieve data** zijn gegevens die categorieën uitdrukken. Rekenen met deze data is meestal niet zinvol. Ze zijn
  - **ordinaal** als de categorieën een natuurlijke rangorde hebben;
  - **nominaal** indien niet.

Voorbeeld 2.

CC BY-NC-SA 4.0

2 WELKE SOORTEN DATA ZIJN ER?

- **Continu kwantitatief:** Lengte van studenten (in cm): 172.3, 168.9, 174.1, 170.0, 169.7
- **Discreet kwantitatief:** Aantal huisdieren: 2, 0, 0, 3, 1
- **Ordinaal kwalitatief:** Tevredenheid over een cursus: goed, matig, uitstekend, slecht, goed
- **Nominaal kwalitatief:** Lievelingskleur: blauw, groen, rood, blauw, geel

Opmerking 1. Kwantitatieve data laten toe om verschillen zinvol te interpreteren. Denk bijvoorbeeld aan een lengteverschil van 3 cm tussen twee personen. Dit is een betekenisvol verschil dat iets zegt over hun fysieke eigenschappen. Voor kwalitatieve data is dit niet het geval.

Opmerking 2. Kwalitatieve data kunnen ook bestaan uit getallen. Denk maar aan postcodes: hoewel ze numeriek lijken, drukken ze een locatie uit en is het niet zinvol om ermee te rekenen.

Oefening 1. Wat is het datatype van de volgende variabelen?

1. Merk van smartphone

Continu kwantitatief	Discreet kwantitatief	Ordinaal kwalitatief	Nominaal kwalitatief

2. Beoordeling van een film (slecht, matig, goed, uitstekend)

Continu kwantitatief	Discreet kwantitatief	Ordinaal kwalitatief	Nominaal kwalitatief

3. Temperatuur in °C

Continu kwantitatief	Discreet kwantitatief	Ordinaal kwalitatief	Nominaal kwalitatief

4. Geboorteprovincie

Continu kwantitatief	Discreet kwantitatief	Ordinaal kwalitatief	Nominaal kwalitatief

5. Schoenmaat

Continu kwantitatief	Discreet kwantitatief	Ordinaal kwalitatief	Nominaal kwalitatief

6. Rugnummer op een voetbaltruitje

Continu kwantitatief	Discreet kwantitatief	Ordinaal kwalitatief	Nominaal kwalitatief

Oefening 2. Voor welke van de onderstaande variabelen is het zinvol om een gemiddelde te berekenen?

Meerkeuze (selecteer alle correcte antwoorden):

(a) Lievelingskleur

(b) Lengte van personen

(c) Temperatuur in graden Celsius

(d) Postcode

(e) Aantal kinderen per gezin

CC BY-NC-SA 4.0

2 WELKE SOORTEN DATA ZIJN ER?

(f) Rugnummers op voetbaltruitjes

CC BY-NC-SA 4.0