Soorten data en de structuur van een dataset

Prof. dr. Herman Callaert

In God we trust
others must have data

Inhoud

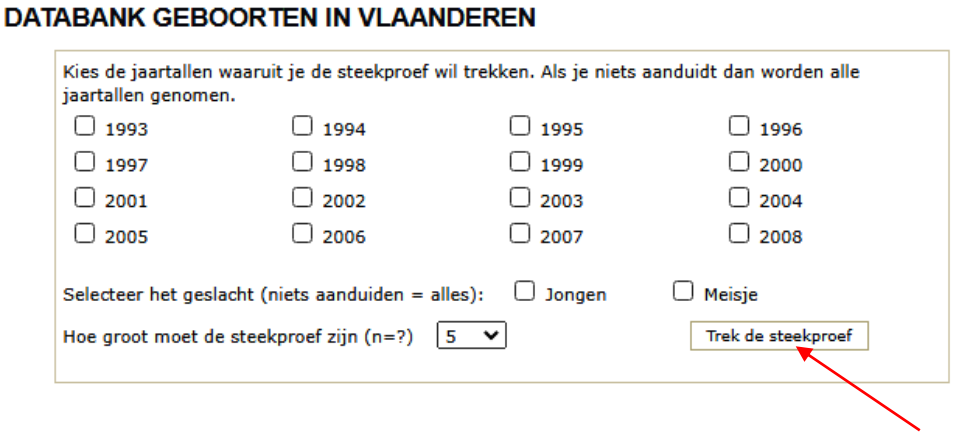
1. De structuur van een dataset

Een dataset (of gegevensverzameling of databank) is niet zomaar een hoop gegevens. Bij een nauwkeurig geformuleerde onderzoeksvraag heb je nauwkeurig opgemeten data nodig met een duidelijke structuur.

Bij de start van je onderzoek werk je met "ruwe" data. "Ruwe" data is de naam voor data die je opschrijft zoals je ze verzameld hebt. Je doet verder nog niets met die data (zoals samenvatten, transformeren, enz..). Als anderen later je onderzoek willen overdoen (of je willen controleren op fraude) dan zullen zij je "ruwe" data opvragen, samen met een beschrijving van de manier waarop je die hebt opgemeten.

1.1. Een concreet voorbeeld

Ga naar https://www.uhasselt.be/lesmateriaal-statistiek en klik op "Databank geboorten" en klik daarna op "Trek de steekproef".



Je hebt nu een dataset met 5 kinderen. Bij elk kind zijn 5 eigenschappen opgemeten: de duur van de zwangerschap, het geboortegewicht, het geslacht, de leeftijd van de moeder en het geboortejaar. De gevonden waarden staan in een rechthoekig schema: dat is de dataset, met daarin de ruwe data.
Zo'n dataset kan er als volgt uitzien:

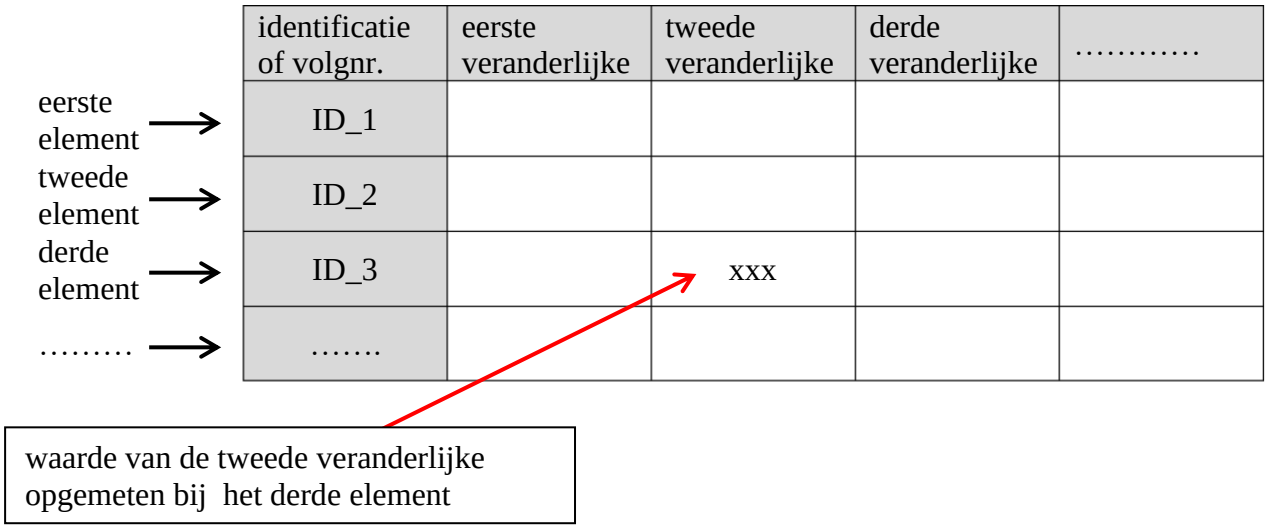
Volgnr	Duur	Gew	Sex	Lft_m	Gebjaar
1	40	3570	0	31	1998
2	38	2700	0	29	1999
3	39	2855	1	31	1996
4	40	3130	1	27	2006
5	41	5000	0	30	2002

Op de website staat ook de betekenis van de afkortingen:

Duur =	duur van de zwangerschap (in weken)
Gew =	geboortegewicht (in gram)
Sex =	geslacht met 1=jongen 0=meisje
Lft_m =	leeftijd van de moeder op het moment van de bevalling
Gebjaar =	geboortejaar

1.2. De algemene structuur

In een statistisch onderzoek leidt de onderzoeksvraag je naar een aantal "elementen" (zoals kinderen) waarbij je één of meerdere eigenschappen (zoals zwangerschapsduur) onderzoekt. Die eigenschappen heten "veranderlijken". Bij elk element noteer je de "waarde" van elke veranderlijke. Al die waarden komen terecht in je databank.



1.2.1. Elementen

"Elementen" is de algemene naam voor de objecten in je statistisch onderzoek.
Die objecten kunnen personen zijn (kinderen, soldaten, inwoners, …) of dieren (paarden, muizen, apen, …) of planten (irissen, eiken, tomaten, …), of zaken (gemeenten, auto's, smartphones, …) … Bij het opschrijven van de gegevens komen de elementen terecht op de rijen van het rechthoekige schema. Bij elke rij hoort juist één element.

Als je de studie beperkt tot de dataset die je zopas in het voorbeeld getrokken hebt, dan zijn de elementen in je onderzoek de 5 gevonden kinderen. Elke rij in je schema stelt 1 kind voor. Welk kind dat is zou je kunnen noteren met naam en adres. Maar hier zie je enkel een volgnummer. Het is niet ongewoon dat elementen enkel met een code worden aangegeven. Dat kan te maken hebben met "privacy" of met "medisch geheim".

Het woord "element" vervang je in je onderzoek door een betekenisvolle naam zoals "atleet", "passagier", "smartphone", "huishouden", "respondent" (bij een enquête), "patiënt" (bij een klinische studie), enz.

1.2.2. Veranderlijken

"Veranderlijken" is de algemene naam voor de "eigenschappen" die je onderzoekt.
Per element meet je bepaalde eigenschappen op en de resultaten komen terecht in de kolommen van je databank. Elke kolom draagt een naam die zegt over welke eigenschap het gaat.

Alleen maar een naam zegt soms niet genoeg. Daarom voeg je bij de databank ook een precieze beschrijving van de veranderlijken toe. Bij lengte bijvoorbeeld is het belangrijk dat je weet of die is opgemeten in centimeter of in inches.

1.3. Nog een concreet voorbeeld: Californische gezinnen

Een tweede voorbeeld van een dataset halen we uit Californië, waar in 1961 een uitgebreid onderzoek werd gestart bij de geboorte van een kind. Men noteerde toen heel veel kenmerken van dat kind (geslacht, bloedgroep, gewicht, lengte, tijdstip van geboorte, …), samen met kenmerken van de vader en de moeder (leeftijd, gewicht, lengte, rookgedrag, …). Tien jaar later werd elk gezin opnieuw onderzocht. Het tienjarige kind legde toen verschillende psychologische testen af. Ook de klassieke opmetingen (zoals lengte en gewicht) kwamen opnieuw aan bod, zowel bij het kind als bij de ouders. Een heel klein stukje van die databank (een beetje aangepast) ziet er als volgt uit:

ID	SEX	BLGK	LGTK	GEWK	LFTM	LFTV	SIGV
1	J	B	53.3	3810	22	23	0
2	J	AB	55.9	3720	25	25	20
3	M	O	50.8	3180	36	42	0
4	M	O	50.8	2990	25	30	30
5	J	A	50.8	2900	25	32	3
6	M	A	55.9	4350	42	40	0
7	M	AB	49.5	2770	19	33	20
8	M	A	53.3	3670	36	43	6

In dit voorbeeld zijn de elementen “Californische gezinnen die in 1961 een baby kregen”. Elke rij stelt zo’n gezin voor. Deze gezinnen hebben in de databank geen naam gekregen, maar enkel een identificatienummer (afgekort als ID).

In de Californische databank zijn de veranderlijken als volgt omschreven:

- SEX geslacht van het kind (M=meisje, J=jongen)
- BLGK bloedgroep van het kind (O A B AB)
- LGTK lengte (in cm) van het kind bij de geboorte
- GEWK gewicht (in g) van het kind bij de geboorte
- LFTM leeftijd (in jaren) van de moeder bij de geboorte van het kind
- LFTV leeftijd (in jaren) van de vader bij de geboorte van het kind
- SIGV sigaretten (in aantal per dag) gerookt door de vader bij de geboorte van het kind

Voor de volledige dataset ga je naar https://www.uhasselt.be/lesmateriaal-statistiek waar je klikt op *Werkteksten* en dan scrolt naar *4.Methoden en technieken bij een statistisch onderzoek – Soorten data en de structuur van een dataset – California.*

Basisvragen bij een dataset:

- Welke elementen zijn hier onderzocht?
- Welke veranderlijken zijn er bij die elementen opgemeten?

Informatie over de elementen en de veranderlijken hoort bij elke dataset.

2. Soorten data

Welke statistische methoden je in je onderzoek kan gebruiken, hangt voor een deel af van het soort data waarover je beschikt. Ook de manier waarop je data codeert of samenvat, speelt een rol.

In het secundair onderwijs maak je een onderscheid tussen:
- statistische methoden voor categorische veranderlijken
- statistische methoden voor numerieke veranderlijken.

Alternatieve opdelingen en verfijningen vind je in de appendix van deze tekst.

2.1. Categorische veranderlijken

Categorische veranderlijken hebben waarden die in categorieën terechtkomen.
Deze waarden hoeven zich niet te lenen tot "wiskundige bewerkingen".

M&M snoepjes zijn gekleurd. In een bepaald type van zakjes zijn de enig mogelijke kleuren: blauw, bruin, geel, groen, oranje en rood.

Jij hebt zo'n zakje en je bent geïnteresseerd in de kleur van de snoepjes. Dus neem je een snoepje uit het zakje en noteert de kleur: bruin. Een tweede snoepje is rood. Je gaat zo verder tot je de kleur van alle snoepjes (het waren er 50) genoteerd hebt.
In dit onderzoek zijn de "elementen" snoepjes. De "veranderlijke" die je bij elk element opmeet is de kleur. Wat je in je databank opschrijft zijn de "waarden" van de veranderlijke. Hier zijn dat de "waarden" die in de kolom "Kleur" terechtkomen zoals bruin, rood, oranje, enz... Rechts zie je hier een klein deeltje van je dataset.

Kleur
Bruin
Rood
Oranje
Rood
Groen
Geel
Blauw
Groen
……

De waarden die je in de dataset opschrijft, zijn de data waarmee je werkt in een statistisch onderzoek. In dit voorbeeld zijn die data niet geschikt voor wiskundige bewerkingen zoals: "oranje maal groen plus geel". Dat heeft geen zin.

Bij "Kleur" kan je de verschillende waarden in categorieën klasseren. Stel je daarbij de categorieën voor als verschillende doosjes. In het doosje "bruin" kan je alle bruine snoepjes leggen, in het doosje "geel" alle gele snoepjes, enz. De waarden van de veranderlijke "Kleur" komen in verschillende categorieën terecht. Daarom noemen we kleur een categorische veranderlijke.

Hieronder zie je enkele voorbeelden van categorische veranderlijken.

naam van de categorische veranderlijke	mogelijke **waarden** van de categorische veranderlijke
bloedgroep	O A B AB
merk smartphone	Samsung, Apple, Huawei, LG, ander
Vlaamse provincie	Oost-Vl., West-Vl., Vl.-Brabant, Limburg, Antwerpen
merk frisdrank	Coca-Cola, Pepsi, Fanta, Sprite, 7UP, Schweppes, ander
handvoorkeur	linkshandig, rechtshandig, ambidexter
kennis van statistiek	uitstekend, behoorlijk, matig, slecht
mate van akkoord	volledig akkoord, eerder wel, eerder niet, helemaal niet
lust koriander	ja, neen

2.2. Numerieke veranderlijken

Numerieke veranderlijken hebben waarden die numeriek (= getallen) zijn.
De waarden zijn getallen en daarop zijn “wiskundige bewerkingen” mogelijk.

Bij de naam van een numerieke veranderlijke hoort ook de eenheid waarin de opmeting gebeurt. In de Californische databank is “LGTK” de lengte van het kind bij de geboorte. Die lengte is opgemeten in centimeter. Ook “LFTV” is een numerieke veranderlijke. Die geeft aan hoe oud de vader was toen de baby geboren werd. Die leeftijd is uitgedrukt in jaren.

Wiskundige bewerkingen op numerieke data leiden tot getallen die op een zinvolle manier kunnen geïnterpreteerd worden. Maar ook hier is wat gezond verstand nuttig. Dat Vlaamse gezinnen gemiddeld 1.5 kinderen hebben neem je best niet te letterlijk.

Hieronder zie je enkele voorbeelden van numerieke veranderlijken.

naam van de numerieke veranderlijke	mogelijke **waarden** van de numerieke veranderlijke
lengte van een baby (in cm)	51.5 48.3 50.7 ….
leeftijd van rusthuisbewoners (in jaren)	83 90 79 ….
aantal huisdieren (honden en katten)	0 1 2 3 ….
tijd "100m vrouwen op Olympische spelen" (in s)	10.83 10.78 10.62 ….
afstand van thuis naar school (in km - afgerond)	0 1 2 3 ….
breedte van het kelkblad van irissen (in mm)	28, 33, 24 ….
lichaamstemperatuur van medeleerlingen (in °C)	36.4 35.8 37.1 ….

2.3. Wijze van opmeten

De manier waarop veranderlijken worden opgemeten, kan mee bepalen tot welke soort zij behoren.

Een **naam** van een veranderlijke (zoals leeftijd) zegt niet alles.

Als het over de leeftijd van mensen gaat dan denk je aan numerieke data die terechtkomen in een interval (bijvoorbeeld tussen 0 en 130). Bij de veranderlijke "LFTM" in de Californische databank kan je vermoeden dat een interval van 10 tot 60 al ruimschoots voldoende is, want het gaat daar over de leeftijd van de moeder bij de geboorte van haar kind. Als de leeftijd genoteerd is in jaren dan behandel je *leeftijd* als een *numerieke veranderlijke*.

In sommige studies splits je leeftijd op in groepen. Je werkt dan bijvoorbeeld met de categorieën "kinderen", "jongeren" en "volwassenen". In zo'n studie behandel je *leeftijd* als een *categorische veranderlijke*.

De **waarden** van een veranderlijke (zoals getallen) zeggen niet alles.

Als je in de databank een kolom ziet staan met getallen zoals 18, 111, 154, …. dan weet je nog niet of dat waarden zijn van een *numerieke veranderlijke*. Inderdaad, de getallen die je hier ziet zijn de rugnummers van Wout van Aert, Greg van Avermaet en Thomas de Gendt, Belgische renners in de Ronde van Frankrijk 2020. Op rugnummers voer je geen wiskundige bewerkingen uit, dat zijn geen getallen om mee te rekenen. Zij dienen als identificatie van de wielrenner. Je hebt hier te maken met een *categorische veranderlijke*.

3. Voorbeelden

3.1. Fisher's Iris data

R. A. Fisher (1890–1962) was een Engelse statisticus die veel werkte met gegevens uit landbouw, biologie en genetica. In één van zijn studies probeerde hij om op basis van blaadjes van irissen de soorten irissen van elkaar te onderscheiden.
De dataset die Fisher gebruikte, publiceerde hij in Annals of Eugenics 7, 179-188 (1936).

Het onderzoek gaat over 3 soorten irissen: Iris Setosa, Iris Versicolor, en Iris Verginica. Van elke soort zijn 50 bloemen (= 50 "elementen") onderzocht waarbij 5 kenmerken (= 5 "veranderlijken") zijn genoteerd: het soort iris, de lengte en de breedte van een bloemblad en de lengte en de breedte van een kelkblad. In totaal zijn er in de databank 150 elementen terechtgekomen.

Hieronder vind je een stukje van de beroemde "Fisher's Iris data".

Volgnr	Naam	L_Bbl	B_Bbl	L_Kbl	B_Kbl
1	3	51	15	63	28
2	1	14	2	50	33
3	2	38	11	55	24
4	3	50	20	57	25
5	2	39	14	52	27
6	2	44	14	66	30
7	1	13	2	47	32
8	1	15	2	52	34

In de eerste kolom staat het volgnummer van de opgemeten bloem.
De namen van de veranderlijken hebben de volgende betekenis:

- Naam = een cijfercode voor het soort iris waarbij
  - 1 = Iris Setosa
  - 2 = Iris Versicolor
  - 3 = Iris Verginica
- L_Bbl = lengte van het bloemblad (in mm)
- B_Bbl = breedte van het bloemblad (in mm)
- L_Kbl = lengte van het kelkblad (in mm)
- B_Kbl = breedte van het kelkblad (in mm)

Bemerk dat de naam van het soort iris genoteerd is met een cijfer (1, 2 of 3). Dit zijn geen cijfers waarop je wiskundige bewerkingen uitvoert. Die cijfers zijn hier waarden van een *categorische veranderlijke*.

Voor de volledige dataset ga je naar https://www.uhasselt.be/lesmateriaal-statistiek waar je klikt op *Werkteksten* en dan scrolt naar *4.Methoden en technieken bij een statistisch onderzoek – Soorten data en de structuur van een dataset – Irissen.*

3.2. De Titanic

Op 14 april 1912 botste de Titanic tegen een ijsberg en zonk. Deze ramp is uitvoerig gedocumenteerd en er is ook een film over gemaakt. Op het internet vind je over de Titanic een massa informatie met onder meer gegevens over de opvarenden. De rijkere passagiers hadden een duur "eerste klas" ticket gekocht om in luxueuze kajuiten de overtocht te maken. Men beweert dat de proportie overlevenden veel groter was bij de "eerste klas" passagiers dan bij de anderen. Is dat waar? Kan je dat uit de dataset halen? Er zit ook nog veel andere informatie in die dataset waarvan je hieronder een klein deeltje ziet.

Volgnr	Afloop	Haven	Ticket	Leeftijd	Geslacht
1	D	S	3	42	M
2	D	S	4	21	M
3	D	S	3	14	M
4	D	S	3	16	M
5	L	S	1	39	V
6	L	S	3	16	V
7	L	S	3	25	M
8	D	C	2	30	M
9	D	C	2	28	V
10	L	S	3	20	M

In de eerste kolom staat het volgnummer van de passagier zoals die in de databank is opgenomen. De namen van de veranderlijken hebben de volgende betekenis:

- Afloop : beschrijft hoe de ramp is afgelopen met
  - de opvarende is verdronken (code D = "dood")
  - de opvarende is gered (code L = "levend")
- Haven : de haven waar de opvarende aan boord ging met
  - B = Belfast
  - C = Cherbourg
  - Q = Queenstown
  - S = Southampton
- Ticket : soort ticket van de opvarende met
  - 1 = eerste klas
  - 2 = tweede klas
  - 3 = derde klas
  - 4 = bemanningslid
- Leeftijd : leeftijd in jaren (0 = kind jonger dan 1 jaar)
- Geslacht : M = man, V = vrouw

Welk verhaal vertelt de dataset over de passagier met volgnummer 4?

Voor de volledige dataset ga je naar https://www.uhasselt.be/lesmateriaal-statistiek waar je klikt op *Werkteksten* en dan scrolt naar *4.Methoden en technieken bij een statistisch onderzoek – Soorten data en de structuur van een dataset – Titanic*.

4. Appendix

Bijna alle disciplines gebruiken statistiek. Zij doen dat dikwijls vanuit hun eigen invalshoek en met een eigen terminologie. De wereld van de sociologie is nogal verschillend van de wereld van de fysica of van de geneeskunde of de archeologie of de psychologie of …

Bij teksten uit die verschillende disciplines kom je wel eens andere namen of indelingen tegen bij "soorten data" en "soorten statistische methoden". Het is handig dat je dan weet waarover het gaat.

4.1. Vier meetschalen

Soms gebruikt men de volgende 4 meetschalen om een onderscheid tussen data te maken.

In verschillende tekstboeken beperkt men zich tot 3 meetschalen: nominaal, ordinaal en interval. Bij deze auteurs dekt het woord "interval" zowel de derde als de vierde meetschaal.

4.1.1. De nominale schaal

In het system van de 4 meetschalen staat de nominale schaal op de laagste trap.
Een nominale veranderlijke heeft waarden die alleen maar dienen ter identificatie.

De veranderlijke "bloedgroep" heeft als waarden: O, A, B, en AB. Die waarden zeggen welke bloedgroep je hebt. Bewerkingen op bloedgroepen, de afstand tussen bloedgroepen of een zinvolle volgorde van bloedgroepen komt hier niet ter sprake.
"Naam" is in het Latijn "*nomen*". Op de *nominale* schaal heb je alleen maar een "naam" (of een "label" of een "ID-identificatie").

4.1.2. De ordinale schaal

De ordinale schaal staat een trap hoger. Naast de identificatie komt hier een zinvolle ordening bij.

Nadat de leerkracht je heeft uitgelegd welke soort veranderlijken er allemaal bestaan, kan je je mening geven over de snelheid waarmee die uitleg gebeurde. Je kan daarbij kiezen tussen 5 mogelijkheden:

- veel te traag
- te traag
- juist goed
- te snel
- veel te snel

De veranderlijke "je mening over de snelheid van uitleg" heeft hier 5 mogelijke waarden. Die waarden zijn in woorden uitgedrukt. Zij hebben een zinvolle volgorde maar je kan daar niet mee rekenen. Je weet ook niet of de afstand tussen "veel te traag" en "te traag" even groot is als de afstand tussen "te traag" en "juist goed".

4.1.3. De interval-schaal

De interval-schaal heeft de eigenschappen van de ordinale schaal met bovendien een zinvol afstandsbegrip.
Om de afstand te kunnen berekenen moeten de opgemeten waarden getallen zijn. Twee opmetingen bepalen dan een interval en de lengte van dat interval is de afstand tussen die twee opmetingen.

Temperatuur kan je meten in graden Celsius en in graden Fahrenheit met $^\circ C = \frac{5}{9}\left(^\circ F - 32\right)$. Van drie voorwerpen A, B en C zie je hieronder de temperatuur en de temperatuurverschillen (hoog min laag):

Temperatuur			
	A	B	C
°*C*	10	20	40
°*F*	50	68	104

Temperatuurverschillen			
	B–A	C–B	C–A
°*C*	10	20	30
°*F*	18	36	54

Bewerkingen op **intervallen** (bewerkingen op verschillen, die opgetekend staan in de rechtse tabel) is hier zinvol. Het temperatuurverschil tussen A en B is dubbel zo groot als tussen B en C en de som van deze verschillen is gelijk aan het verschil tussen A en C. Deze bewering is juist zowel in graden Celsius als in graden Fahrenheit.

Uitspraken gebaseerd op de **waarden zelf** (in de linkse tabel) kunnen zo maar niet. Dat B "dubbel zo warm" is als A is blijkbaar juist in graden Celsius maar het is fout in graden Fahrenheit !!!

4.1.4. De ratio-schaal

De ratio-schaal heeft de eigenschappen van de interval-schaal en bovendien zijn hier bewerkingen op de waarden zelf zinvol.

De *ratio*-schaal is de schaal der *verhoudingen*. Als je twee keer zo groot bent als je kleine zusje dan kan je dat meten in meter, centimeter of zelfs millimeter, maar je blijft twee keer zo groot.

Deze schaal is de meest gebruikte schaal wanneer de waarden van je veranderlijke getallen zijn. Denk maar aan lengte, gewicht, aantallen, bloeddruk, tijd, …

4.2. Kwantitatief en kwalitatief

Ook de woorden kwantitatief en kwalitatief worden gebruikt om data te classificeren.

Kwantitatief drukt een hoeveelheid uit. Kwantitatieve data geven je informatie over: hoeveel, hoe lang, hoe zwaar, hoe warm, hoe dikwijls, … Opmetingen op de interval-schaal en de ratio-schaal zijn kwantitatief.

Kwalitatief drukt een kwaliteit uit, niet een hoeveelheid. De waarden van een kwalitatieve veranderlijke zijn meestal woorden of codes zoals "1 = jongen en 2 = meisje" of "groen, geel, blauw" of "goed, beter, best" enz.

4.3. Discreet en continu

In statistiek spreek je over discreet numerieke data en continu numerieke data om te verwijzen naar:
- statistische methoden waarbij je de data als discreet behandelt
- statistische methoden waarbij je de data als continu behandelt.

Getallen die je in een dataset ontmoet zijn altijd discreet. Dat hangt samen met de onnauwkeurigheid van meetinstrumenten en met het feit dat je bij het opschrijven toch ergens moet stoppen.

Bij de discreet/continu classificatie denk je in statistiek aan een "onderliggend model voor de werkelijkheid" samen met "goede methoden die zinvolle resultaten leveren".

Bij numerieke veranderlijken met een beperkt aantal waarden kan je methoden voor **discreet numerieke veranderlijken** gebruiken. Denk bijvoorbeeld aan het aantal snoepjes in een M&M zakje, het aantal leerlingen in je klas of het aantal huisdieren (honden en katten) in een huishouden.

Bij lengte, tijd, oppervlakte, bloeddruk… denk je aan uitkomsten die (theoretisch) een continuüm van waarden kunnen aannemen. Hier kan je methoden voor **continu numerieke veranderlijken** gebruiken. Deze methoden werken dikwijls ook goed bij discreet numerieke veranderlijken die heel veel verschillende waarden kunnen aannemen.

4.4. Categorisch systeem

Kleuren, bloedgroepen, rugnummers…. zijn categorische veranderlijken. De elementen in je dataset (zoals pasgeboren kinderen) komen voor de bestudeerde eigenschap (zoals bloedgroep) in categorieën terecht. Die categorieën hebben een "naam" of een "label" zoals O, A, B en AB.
Als elk element juist in één en slechts één categorie terechtkomt, dan heb je **een categorisch systeem**.

Een categorisch systeem met slechts twee categorieën is een **dichotomie.** In de databank met de geboorten heb je voor het geslacht alleen de waarden "Jongen" (code = 1) en "Meisje" (code = 0).

Statistiek heeft eigen methoden om veranderlijken te bestuderen die een categorisch systeem vormen. Die methoden kan je ook gebruiken bij discreet numerieke veranderlijken met een beperkt aantal waarden.

Een kind heeft 20 melktanden. Als je bij kinderen het aantal gewisselde melktanden noteert, dan komt ieder kind terecht in één en slechts één van de 21 categorieën: 0, 1, 2, …19, 20. Je kan hier technieken voor categorische veranderlijken gebruiken terwijl de waarden getallen zijn waarop je toch zinvolle wiskundige bewerkingen kan uitvoeren.

Statistische technieken voor "categorische veranderlijken" kan je gebruiken bij:
- een nominale veranderlijke
- een ordinale veranderlijke
- een discreet numerieke veranderlijke met een beperkt aantal uitkomsten
- een continu numerieke veranderlijke die gegroepeerd is in een beperkt aantal categorieën.