



Tools voor dataverwerking en -analyse


06 oktober2023

dr. Willem De Keyzer – Centre for Applied Data Science HOGENT

Inhoud en programma

- ▶ Databeheer en enkele begrippen
- ▶ Overzicht van verschillende tools en hun voornaamste eigenschappen
- ▶ Demo en zelf beperkte analyses uitvoeren

Programma

09:00-09:15	verwelkoming en kennismaking
09:15-09:45	inleiding databeheer, datastructuur, begrippen gegevensanalyse
09:45-10:30	overzicht en toelichting tools
 10:40-11:40	demo's + oefening Jamovi
11:40-12:00	Q&A + afronding

Doelstellingen

- ▶ Je kent de betekenis van:
 - ▶ Beschrijvende en verklarende statistiek
 - ▶ Getallen voor centrum en spreiding
- ▶ Je hebt noties van de meest gangbare tools voor kwantitatieve gegevensverwerking en –analyse en hun toepassingen
- ▶ Je bent in staat om:
 - ▶ Zelf een databestand aan te maken
 - ▶ Zelf beschrijvende statistiek uit te voeren op data

**Meerwaarde creëren uit data door deze
te verzamelen, te organiseren, te
analyseren en toegankelijk te maken.**

mission statement

We add value to your data.

Expertise

Data governance

- Databases
- Data warehousing
- Open data
- Data protection and privacy

Statistical Data Analysis

- Regression analysis & statistical modeling
- Univariate and multivariate analysis
- Time series analysis

Qualitative analysis

Quality & process

- Measurement system analysis
- Control charts/process capability

Artificial Intelligence

- Machine learning
- Deep learning

Big data

- Internet of Things
- Online/offline analytics
- Industry 4.0

Programming & software development

- Data pipeline
- Model deployment

Data visualisation & storytelling

Research methodology

- Design of experiments
- Study design

Spatial data science

- Topography, remote sensing, 3D
- Geographic Information Sciences
- Cartography

<https://www.hogent.be/onderzoekscentra/centre-for-applied-data-science>

English pages Vacatures Zoeken

HO GENT

opleidingen studeer aan HOGENT onderzoek partners dit is HOGENT nieuws en info

Centre for Applied Data Science.

Praktijkgericht onderzoek van maatschappelijke uitdagingen vereist multidisciplinaire expertise op het gebied van datawetenschappen.

- Het Centre for Applied Data Science (CADS) waarborgt het methodologisch verantwoord verwerven, beheren en analyseren van allerlei soorten data.
- We ondersteunen de creatieve benadering van probleemoplossingen door de nieuwste technologieën rond datawetenschappen in te zetten.

Onze expertise.

Onderzoeksprojecten

- AI-Assisted Master Data Management
- Anthropometric based Estimation of adEoSity
- Applied Geomatics
- Een onderzoek naar activatie van de aanloopstraten en B-locaties in de stedelijke kernen
- HorSine: naar een spin-off rond maatvoering voor paardentuig
- Methodologische aspecten bij de uitvoering van voedselconsumptiepeilingen
- Mobiliteit in een binnenstad en rond een campus
- GDPR-compliant publieke blockchain
- Modelleren bezetting ondergrense parkings en filelengte R40 Stad Gent
- EBSOLLECT: Evidence Based Simulatie Onderwijs en Levenslang Leren door middel van ICT
- Ruimtelijke en financiële simulatie van de Betonstop 2020 - 2040

Dienstverlening en advies

Op zoek naar advies of ondersteuning met betrekking tot onderstaande thema's? Contacteer ons voor een vrijblijvend gesprek, we bekijken wat we kunnen doen voor jou.

- Data Governance
- Statistical Data Analysis
- Qualitative analysis
- Quality & process
- Artificial Intelligence
- Programming & Software development
- Research methodology
- Spatial Data Science

Contact.

Heb je een vraag of wil je met ons samenwerken? Neem dan contact op met cads@hogent.be.

HO GENT

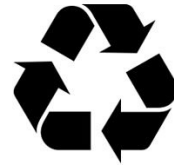
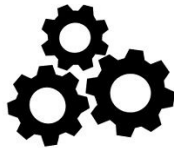
Wat is research data management?

- ▶ **Goede omgang met onderzoeksdata**
 - ▶ bevordert onderzoeksorganisatie en –samenwerking
 - ▶ essentieel onderdeel van het onderzoeksproces
 - ▶ wetenschappelijke integriteit (transparantie van methodiek, reproduceerbaarheid, verifieerbaarheid en hergebruik)
- ▶ **Doel: onderzoeksdata vindbaar, toegankelijk en begrijpelijk zijn en blijven**
- ▶ **RDM heeft betrekking op planning, verzameling, organisatie, documentatie, opslag, beveiliging, verwerking, analyse, archivering, ontsluiting en hergebruik van onderzoeksdata**
- ▶ **Datamanagementplan**

<https://www.ugent.be/en/research/datamanagement/overview.htm>

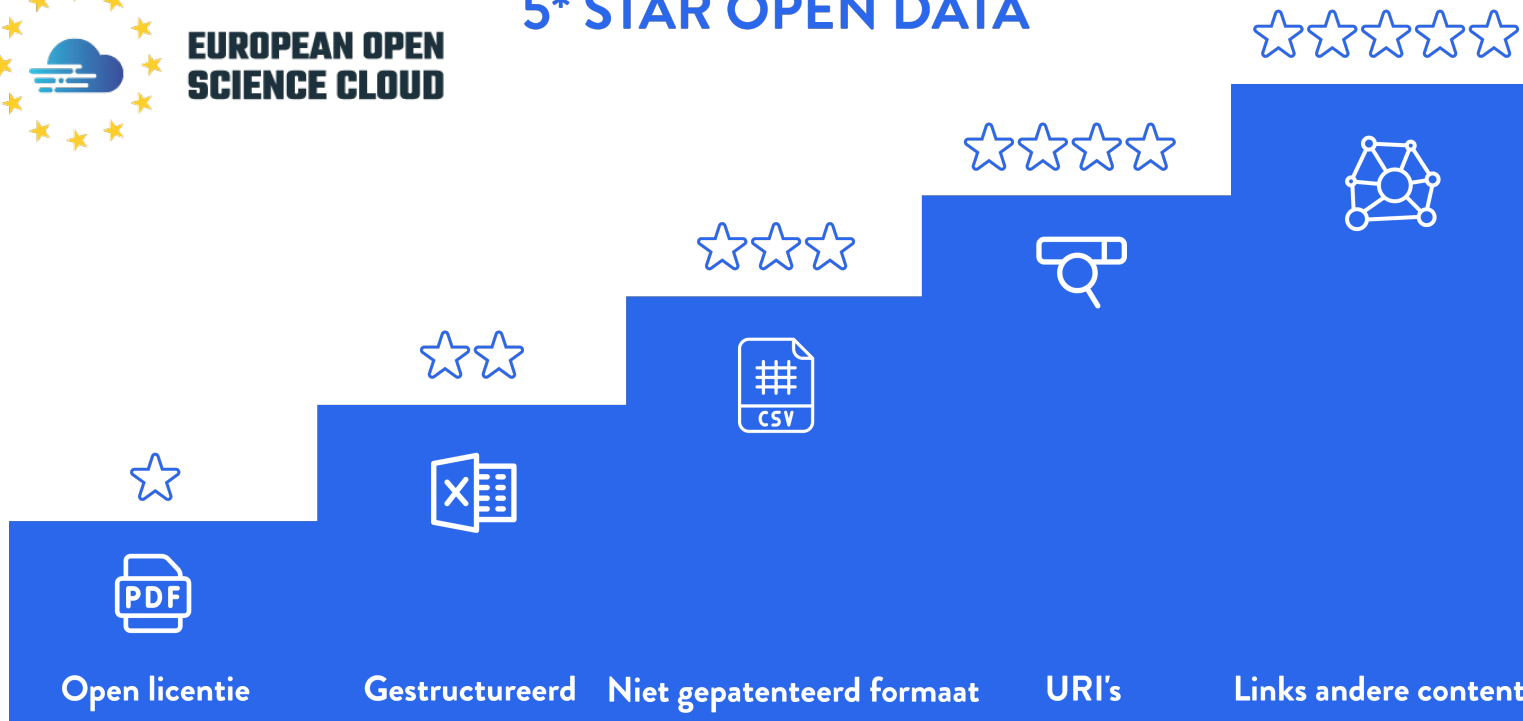
<https://dmponline.be>

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}



**EUROPEAN OPEN
SCIENCE CLOUD**

5* STAR OPEN DATA

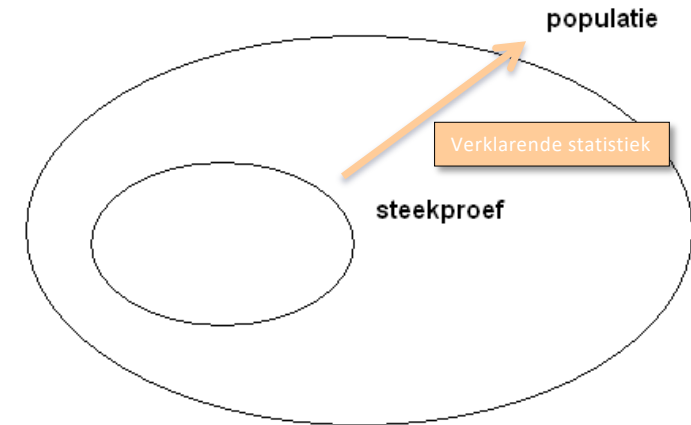


**HO
GENT**

Gegevensanalyse

1. Verzamelen
2. Voorstellen
3. Beschrijven

Beschrijvende statistiek



4. Generaliseren

Verklarende statistiek

Doel ? Uit steekproef betrouwbare besluiten trekken over de populatie
(generaliseren)

Concepten

- ▶ **Populatie**

- ▶ Verzameling van alle te bestuderen objecten in onderzoek
- ▶ Alle inwoners van België, woonhuizen/appartementen in Vlaanderen

- ▶ **Steekproef**

- ▶ Deelverzameling van populatie – aantal eenheden = **omvang** n
- ▶ Selectie van 1500 Belgen, studie van 300 appartementen in Vlaanderen

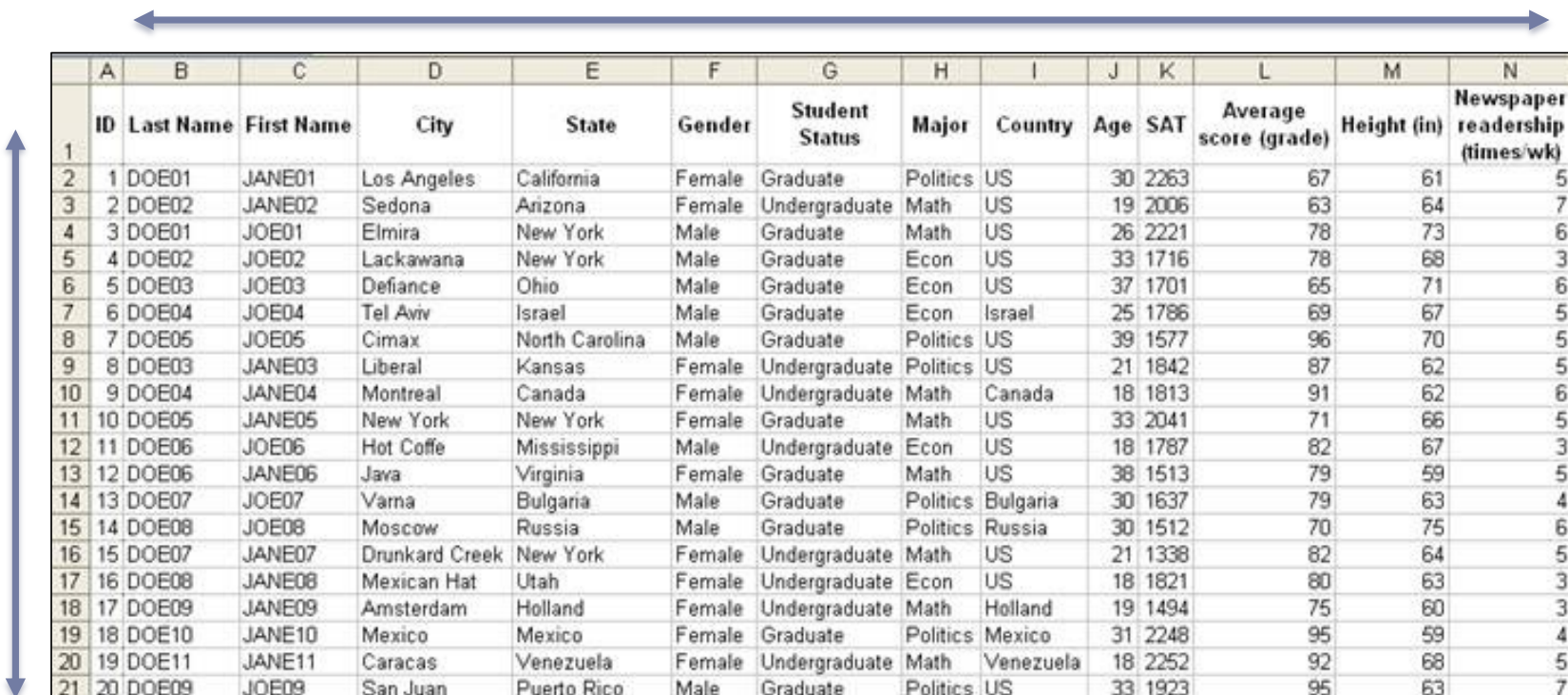
- ▶ **Variabelen**

- ▶ Kenmerken die onderzocht worden
- ▶ Lengte, geboortegewicht, prijs, ...

- ▶ **Verwerking van alle gegevens in een datamatrix**

Datamatrix

variabelen



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)
1														
2	1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
3	2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
4	3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
5	4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
6	5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
7	6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
8	7	DOE05	JOE05	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
9	8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
10	9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
11	10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
12	11	DOE06	JOE06	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
13	12	DOE06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
14	13	DOE07	JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
15	14	DOE08	JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
16	15	DOE07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
17	16	DOE08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
18	17	DOE09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
19	18	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
20	19	DOE11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
21	20	DOE09	JOE09	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7

eenheden

Codering

geboortegeslacht

0 = Man
1 = Vrouw

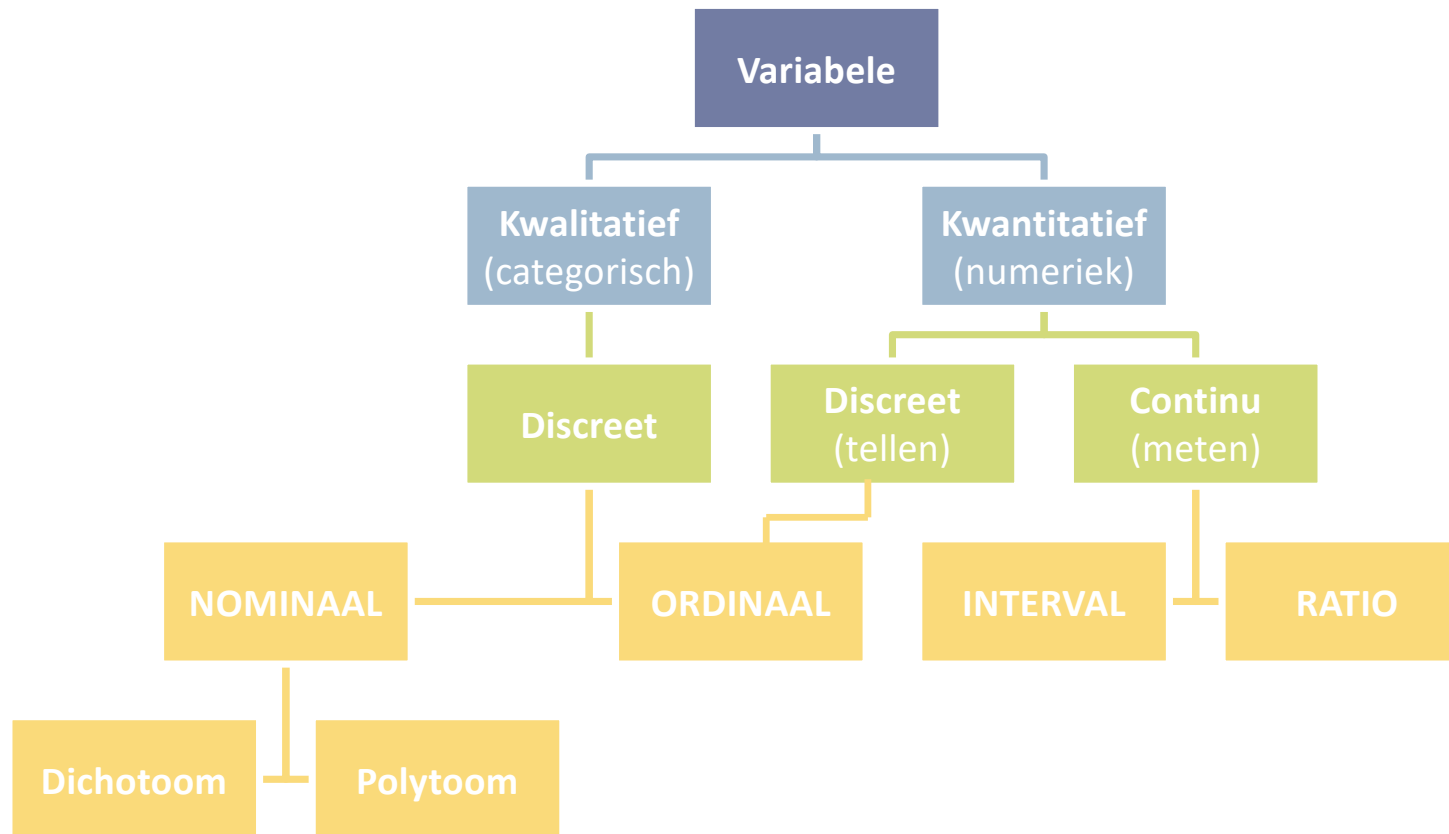
huidige situatie

0 = Student
1 = Werkend
2 = Geen van beide

roker?

0 = Neen
1 = Ja

Meetniveau



Getallen voor centrum en spreiding

▶ Centrummaten

- ▶ Modus
- ▶ Mediaan
- ▶ Gemiddelde

▶ Spreidingsmaten

- ▶ Spreidingsbreedte
- ▶ Standaarddeviatie
 - ▶ Steekproefstandaarddeviatie
 - ▶ Populatiestandaarddeviatie

Modus

- ▶ Meest voorkomende waarde
- ▶ vb. 0 0 0 1 1 2
- ▶ Modus = 0

Mediaan

- ▶ Middelste waarde na rangschikking.
- ▶ vb. 60 50 70 eerst rangschikken -> 50 60 70
- ▶ Mediaan is 60.
- ▶ vb. 50 60 70 80
- ▶ Mediaan is gemiddelde van 60 en 70, dus 65

Gemiddelde

- ▶ Populatiegemiddelde μ
- ▶ Steekproefgemiddelde \bar{x}
- ▶ Alle getallen optellen en delen door het totaal aantal getallen
- ▶ vb. 50, 60 en 70
- ▶ Gemiddelde is $(50 + 60 + 70)/3 = 60$

Spreidingsbreedte

- ▶ Ook range genoemd
- ▶ Hoogste getal minus laagste getal
- ▶ vb. 250 280 300
- ▶ Spreidingsbreedte is $300 - 250 = 50$

Steekproefstandaarddeviatie

- ▶ Formule:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

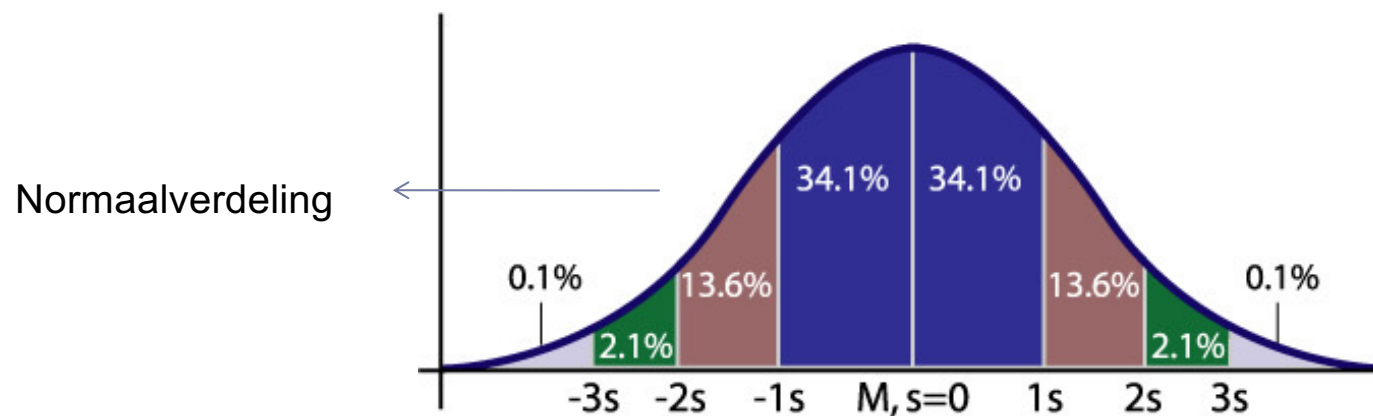
- ▶ Delen door $n - 1$

Voorbeeld berekening steekproefstandaarddeviatie

- ▶ 50 60 70
- ▶ Gemiddelde is 60
- ▶ Verschil met gemiddelde -10, 0 en 10
- ▶ Kwadraat verschillen 100, 0 en 100
- ▶ Optellen kwadraten $100 + 0 + 100 = 200$
- ▶ Delen door $n - 1 = 3 - 1 = 2$
- ▶ Dan is $200/2 = 100$
- ▶ Tot slot de wortel van 100 en we vinden $s = 10$

Verdeling en sigmagebieden

- ▶ 68% van alle gegevens bevinden zich tussen het gemiddelde en 1 maal plus en min de standaarddeviatie
- ▶ 95% van alle gegevens bevinden zich tussen het gemiddelde en 2 maal plus en min de standaarddeviatie

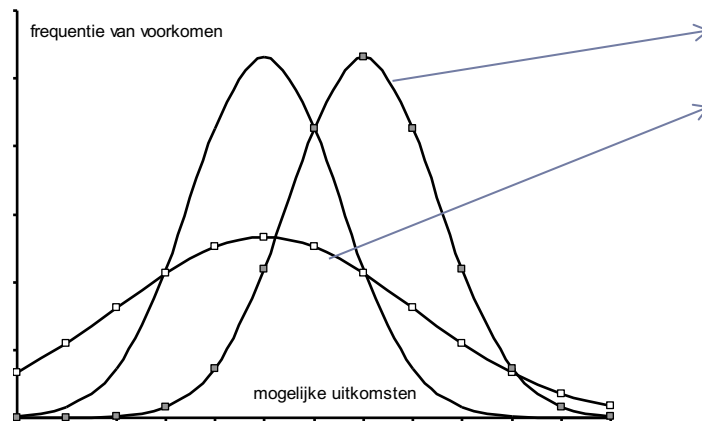


Waarom is de verdeling belangrijk?

- ▶ Bij uitvoering van statistische toetsen wordt afhankelijk van de toets uitgegaan van een normaalverdeling (2 sd rond gemiddelde = 95%).
- ▶ Wanneer de verdeling niet normaal is (scheef is), wordt bij statistische toetsen geen gebruik gemaakt van het gemiddelde als centrummaat.

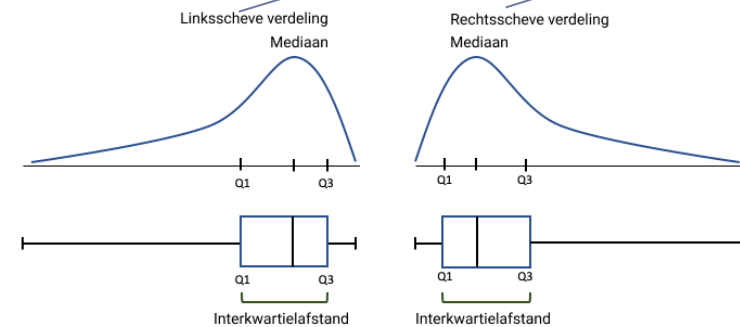
Normaal verdeling nagaan

► Vorm frequentieverdeling: normaal of scheve verdeling



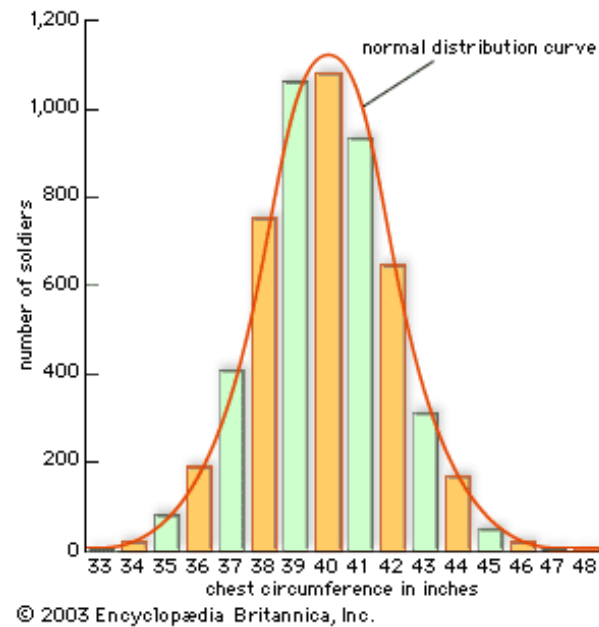
≠ kurtosis

≠ skewness



Normaal verdeling nagaan

► Vorm histogram maken



Normaal verdeling nagaan

▶ Kengetallen

- ▶ Mediaan versus gemiddelde
- ▶ Kurtosis en skewness (± 1 = normaal verdeeld, niet hoger dan 2)

▶ Toetsen

- ▶ Shapiro Wilk's toets
- ▶ Kolmogorov Smirnov toets

Toetsen of de frequentieverdeling afwijkt van de normale verdeling

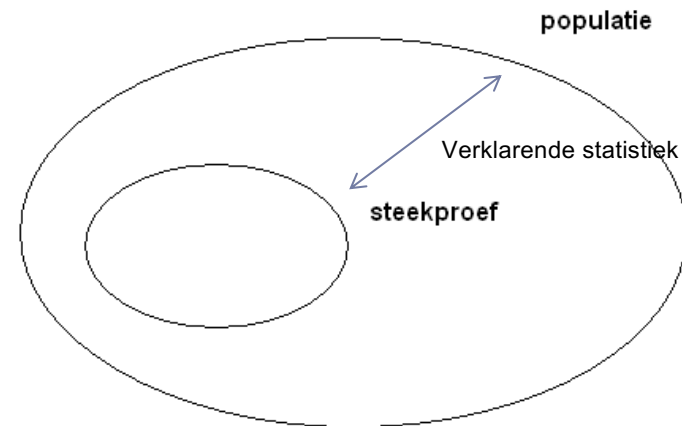
Hypothese toetsen (ifv generaliseren)

- ▶ *Nulhypothese*: er is geen effect, samenhang of verschil
- ▶ *Alternatieve hypothese*: er is wel een effect, samenhang of verschil
- ▶ *Kans*: wanneer op basis van een steekproef uitspraken gedaan worden over de populatie bestaat er een kans dat men zich vergist. Een verschil of samenhang die gevonden wordt in de steekproef kan immers berusten op toeval.
- ▶ *Significantie*: verschillen of samenhang zijn significant wanneer de kans op toeval dat het gevonden verschil of samenhang beneden een bepaald niveau ligt (bv. 5% of 1% bij grote steekproeven).

Hypothese toetsen

► *Significantie:*

- ~ grootte verschil
- ~ grootte steekproef
- ~ spreiding
- ~ gekozen significantieniveau



- Indien de steekproef de populatie benadert (heel groot is), wat is dan de kans dat een gemeten waarde verschilt van de werkelijke waarde?

Overzicht

Beschrijvende statistiek






Verklarende, inferentiële, toetsende, inductieve statistiek

		(MISSCHIEN) NIET NORMAAL VERDEELDE GEGEVENS			NORMAAL VERDEELDE GEGEVENS		
		Nominaal meetniveau	Ordinaal meetniveau	Interval- of rationiveau	Nominaal meetniveau	Ordinaal meetniveau	Interval- of rationiveau
DATA REDUCEREN	Trend (+grafieken)	modus	mediaan		modus	mediaan	gemiddelde
	Variatie (+grafieken)		spreidings-breedte	spreidings-breedte percentielen			standaard-deviatie, z-score
RELATIE AANTONEN	Verband (+grafieken)	Chi ² toets voor associaties, Phi-coëfficiënt	Spearman's rangcorrelatie, Kendall's tau				Product-moment correlatie van Pearson
	Verschil (+grafieken)	Chi ² verschillen-toets, Mc Nemar	Mann-Whitney U-test, Wilcoxon rangteken-test				t-toetsen ANOVA

Software tools



Overzicht tools

	interface	leercurve	mogelijkheden	kostprijs	reproduceerbaarheid
	GUI	++	-	+	-
	GUI + syntax	+	++	++	+
	GUI + script	+	++	++	+
	code	--	+++	gratis	+++
	GUI + code	+++	+	gratis	+

Graphical User Interface

Excel

- ▶ Handig voor data invoer
- ▶ Gegeensvalidatie mogelijk
- ▶ Vrij universeel bestandsformaat

- ▶ Gegevensanalyse (Analysis ToolPak)
- ▶ Draaitabellen
- ▶ (Draai-)Grafieken



SPSS

- ▶ Populair in medische en sociale wetenschappen
- ▶ Gebruik van syntaxen maakt analyses reproduceerbaar
- ▶ Relatief eenvoudig voor beschrijvende statistiek
- ▶ Uitgebreide set van analysetools
- ▶ Minder flexibel
- ▶ Wat gedateerde look and feel en data handling
- ▶ Veel bestanden: data (.sav), output (.spo), syntax (.sps)
=> screenmanagement !



JMP

- ▶ Meer gekend in industriële sector
- ▶ Veel online trainingsmateriaal (proprietary)
- ▶ Dynamische interface, vooral bij data exploratie!
- ▶ Interactieve graph builder
- ▶ Scripting is mogelijk
- ▶ Elke analyse komt in een apart venster => screenmanagement !



R en Python

- ▶ Open-source programmeertalen met een grote community
- ▶ R voornamelijk gebruikt voor statistische analyse
- ▶ Python ook voor algemene programmeerdoeleinden
- ▶ State-of-the-art programmeertaal gericht op datawetenschappen
- ▶ Heel grote flexibiliteit
- ▶ Beide vereisen een grote tijdsinvestering om te leren
- ▶ Worden beide intensief gebruikt in academia



Jamovi

- ▶ Local install of cloud toepassing
- ▶ Uitbreiding met R bibliotheken mogelijk
- ▶ Laagdrempelig maar betrouwbaar
- ▶ Overzichtelijk en intuïtief
- ▶ Snel, efficiënt en mooi rapport
- ▶ Rapport als template bewaren => data vernieuwen en rapport updaten



Online materiaal

- ▶ Excel
 - ▶ Microsoft tutorials, Datacamp, Udemy, LinkedIn Learning, Coursera
- ▶ SPSS
 - ▶ Documentation (<https://www.ibm.com/docs/en/spss-statistics/28.0.0>)
- ▶ JMP
 - ▶ JMP learning academy (https://www.jmp.com/en_ca/learning-library.html)
- ▶ R
 - ▶ R for data science (<https://r4ds.had.co.nz>)
- ▶ Python
 - ▶ Python for beginners (<https://www.python.org/about/gettingstarted>)
- ▶ Jamovi
 - ▶ Youtube tutorials Bart Poulson (<https://datalab.cc/jamovi>)

Contact

- ▶ dr. Willem De Keyzer
coördinator Centre for Applied Data Science
willem.dekeyzer@hogent.be
cads@hogent.be
<https://www.hogent.be/onderzoekscentra>

Campus Schoonmeersen – Gebouw C – lokaal
GSCHC.2.120
Valentin Vaerwyckweg 1
BE-9000 Gent
T +32 9 243 23 95
M +32 476 80 40 15
<https://www.hogent.be/onderzoekscentra/cads>

