PXL-DIGITAL

PXL - IT

Data Advanced Deel 2:
DATA REPRESENTATIE

Lector(en)

Heidi Tans

Sam van Rijn

Julie Vranken

1. Data Representatie: Doel

De student:

- kent het onderscheid tussen verschillende soorten gegevens

- kan voorbeelden om gegevens te verzamelen beoordelen

- weet wat frequentietabellen zijn en kan ze opstellen voor verschillende soorten gegevens

- kan mbv python grafische voorstellingen maken van gegevens en ze interpreteren

- kan bestaande datarepresentaties uit de literatuur met een kritisch oog bekijken en aangeven wat er goed / fout is

- kent de verschillende kengetallen voor locatie, kan ze berekenen mbv python en interpreteren

- kent de verschillende kengetallen voor spreiding, kan ze berekenen mbv python en kan ze interpreteren

- kan werken met datasets in python (inlezen, kolommen / rijen selecteren, bewerkingen uitvoeren, …)

- weet wat een kruistabel is, wanneer te gebruiken en kan deze maken mbv python
- weet wat een scatterplot is, wanneer te gebruiken en kan deze maken mbv python
- Weet wat outliers zijn
- weet wat missing values zijn en hoe ermee om te gaan

2. Data Representatie: Inleiding

De dataset “slaagcijfers.xlsx” (Blackboard) bevat gegevens van 435 studenten:

- Naam student
- Geslacht student: M – V
- Vooropleiding Algemeen: ASO – TSO – BSO
- Vooropleiding IT: veel – matig – geen
- Bis-student: ja - neen
- Score op 3 OLOD’s (afgerond op geheel)
- Studiepunten per OLOD
- Aantal uren besteed OLOD1 (lessenperiode + examenperiode)
- Aantal uren besteed OLOD2 (lessenperiode + examenperiode)
- Gemiddeld aantal uren besteed aan studies (alle OLOD’s) op weekbasis
- Geslaagd: ja – neen
- Procent
- Aantal OLOD’s tweede zit

Merk op:

- lege plaatsen
- “foutieve” / “extreme waarden”
- verschillende waarden: getallen, woorden, codes, ….

➔ Verder onderzoek van de data is aangewezen.

3. Gegevens verzamelen

Gegevens zijn het vertrekpunt van statistisch onderzoek.

Een aantal vragen dringen zich op:

Wat zijn gegevens?

Waar halen we gegevens?

Hoe ‘betrouwbaar’ zijn gegevens?

3.1. Wat zijn gegevens?

Gegevens zijn feiten en cijfers die verzameld, geanalyseerd en samengevat worden voor presentatie en interpretatie.

Er zijn verschillende soorten van gegevens. We onderscheiden

categorische gegevens (ordinaal – nominaal)

numerieke gegevens (discreet – continu)

Afhankelijk van het type gegeven worden er andere analyses uitgevoerd.

Categorische gegevens

Categorische gegevens zijn **labels of namen** die gebruikt worden om eigenschappen van elementen aan te geven. Deze gegevens zijn **niet-numeriek** van aard en hebben een beperkt aantal uitkomstencategorieën (de categorieën kunnen met een getal aangegeven worden, maar dit is enkel een code die geen verdere betekenis heeft).

Deze categorische gegevens kunnen verder onderverdeeld worden in

nominale gegevens d.w.z. **niet - geordend**:

er is geen orde in de verschillende waarden

voorbeeld: Geslacht

ordinale gegevens d.w.z. **geordend**:

er is een orde in de verschillende waarden

voorbeeld: Vooropleiding IT

Numerieke gegevens

Numerieke gegevens geven een **hoeveelheid of grootte** aan. Zulke gegevens worden verkregen door tellen, meten, …

Deze numerieke gegevens kunnen verder onderverdeeld worden in discrete gegevens en continue gegevens.

discrete gegevens

Men spreekt van discrete gegevens als de observaties van die aard zijn dat zij worden uitgedrukt door getallen die niet willekeurig dicht bij elkaar kunnen liggen en waarvan slechts een beperkt aantal verschillende uitkomsten voorkomen in de dataset. In de praktijk heeft men meestal te maken met gehele getallen met een beperkt aantal uitkomstenmogelijkheden.

voorbeeld: Studiepunten per OLOD

continue gegevens

Als de mogelijke numerieke waarden in een zeker bereik willekeurig dicht bij elkaar kunnen liggen, hebben we continue gegevens. In principe kunnen de gegevens elke numerieke waarde aannemen.

Voor de statistische verwerking van continue gegevens zal men deze data dikwijls in groepjes samenvatten (binning: zie later).

Voorbeeld: Gemiddeld aantal uren besteed aan studies (alle OLOD's) op weekbasis

Afhankelijk van het soort gegeven worden andere analyses uitgevoerd.

3.2. Waar halen we deze gegevens (gegevensbronnen)?

We kunnen gegevens verkrijgen bij organisaties die gespecialiseerd zijn in het verzamelen en beheren van gegevens. Maar soms zijn de gegevens die nodig zijn voor een bepaalde toepassing al aanwezig (vb databases met gegevens over de werknemers van een bedrijf, gegevens over koopgedrag van klanten, gegevens met betrekking tot studenten, …).

Voorbeeld

https://statbel.fgov.be/nl/open-data

https://archive.ics.uci.edu/ml/index.php

www.kaggle.com

Wanneer geen gegevens beschikbaar zijn via instanties of bestaande bronnen, kunnen ze ook door onderzoek bekomen worden.

Bij dit soort van onderzoek is o.a. een onderscheid te maken tussen experimenteren & waarnemen

Experimenteren

In experimenteel onderzoek worden de variabelen van belang vastgesteld. Daarna wordt bestudeerd hoe bepaalde factoren de variabelen in het onderzoek beïnvloeden.

Voorbeeld

Invloed van Study Buddy op resultaat van student

Variabele van belang = score OLOD

Factor die variabele beïnvloedt = Study buddy

Om gegevens te verkrijgen over de invloed van een Study buddy op de score op een OLOD worden een aantal proefpersonen geselecteerd. De gegevens omtrent de score van deze proefpersonen worden genoteerd voor en na het gebruik maken van een Study buddy (bvb eerste zit en tweede zit).

Waarnemen

Bij waarnemend onderzoek wordt niet geprobeerd om variabelen te “manipuleren”.

De enquête is de meest gebruikte soort van waarnemend onderzoek.

Bij een enquête worden een aantal onderzoeksvragen gesteld en voorgelegd aan de proefpersonen.

Een opmerking bij waarnemend gegevens verzamelen is dat deze manier vaak veel tijd en geld kost.

Voorbeeld

Bevraging EVA – OLOD

3.3. Hoe betrouwbaar zijn deze gegevens (meetfouten)?

Data analisten dienen zich altijd bewust te zijn van mogelijke foute gegevens in datasets. Het gebruik van foute gegevens kan in bepaalde gevallen erger zijn dan helemaal geen gegevens te hebben en geen conclusies te trekken. Speciale procedures kunnen gebruikt worden om de interne consistentie van de gegevens te controleren. Het blindelings gebruiken van gegevens die voorhanden zijn, kan gevaarlijk zijn.

Slecht voorbeeld van een steekproef

Stel dat we een groep proefpersonen een enquête zouden sturen met daarin de vraag "Vult U graag enquêtes in?" en we kijken naar de teruggestuurde formulieren, dan zouden we waarschijnlijk de conclusie kunnen trekken dat de overgrote meerderheid graag enquêtes invult!! Zo past de steekproef zichzelf aan. 't Is eigenlijk net zo dom als een steekproef per e-mail houden en de vraag "Heeft U een computer?" stellen. De conclusie zal ongetwijfeld zijn dat 100% van de mensen een computer heeft.

Dit zijn natuurlijk wel heel voor de hand liggende voorbeelden, maar soms is het fout zijn van een steekproef slechter te zien. Zo wilden twee leerlingen op de middelbare school onderzoeken hoeveel er gerookt werd onder scholieren. Ze gingen aan het begin van de pauze bij de buitendeur staan en vroegen de eerste 50 leerlingen die naar buiten kwamen: Rook je?" Helaas komen natuurlijk in de pauze de rokers het eerst naar buiten.....

Lees zelf eens onderstaande artikels:

https://www.demorgen.be/buitenland/een-op-de-vijf-vlaamse-moslims-heeft-begrip-voor-is-en-haar-manier-van-actievoeren-b8338104/

http://www.demorgen.be/opinie/de-islam-enquete-is-een-voorbeeld-van-hoe-het-niet-moet-ba5df0e6/Ravqg/

Foutieve gegevens

- 22 jaar werkervaring van een 20 – jarige medewerkster van het bedrijf
- Bekijk observatie 78 in slaagcijfers.xlsx

Foutieve conclusies

http://peilingpraktijken.nl/weblog/2015/01/542/

Hoe zijn deze cijfers berekend?

Stel: slechts 3 jongeren en 3 vakanties en jongere 1 heeft pech op vakantie 1, jongere 2 heeft pech op vakantie 2 en jongere 3 heeft pech op vakantie 3.

	VAKANTIE 1	VAKANTIE 2	VAKANTIE 3
JONGERE 1	X		
JONGERE 2		X	
JONGERE 3			X

➔ 100% kans om pech te hebben op vakantie als men “verkeerd” rekent…

Er zijn in het totaal 3 jongeren en in het totaal 3 keer “pech” ➔ 3 / 3 * 100 = 100%

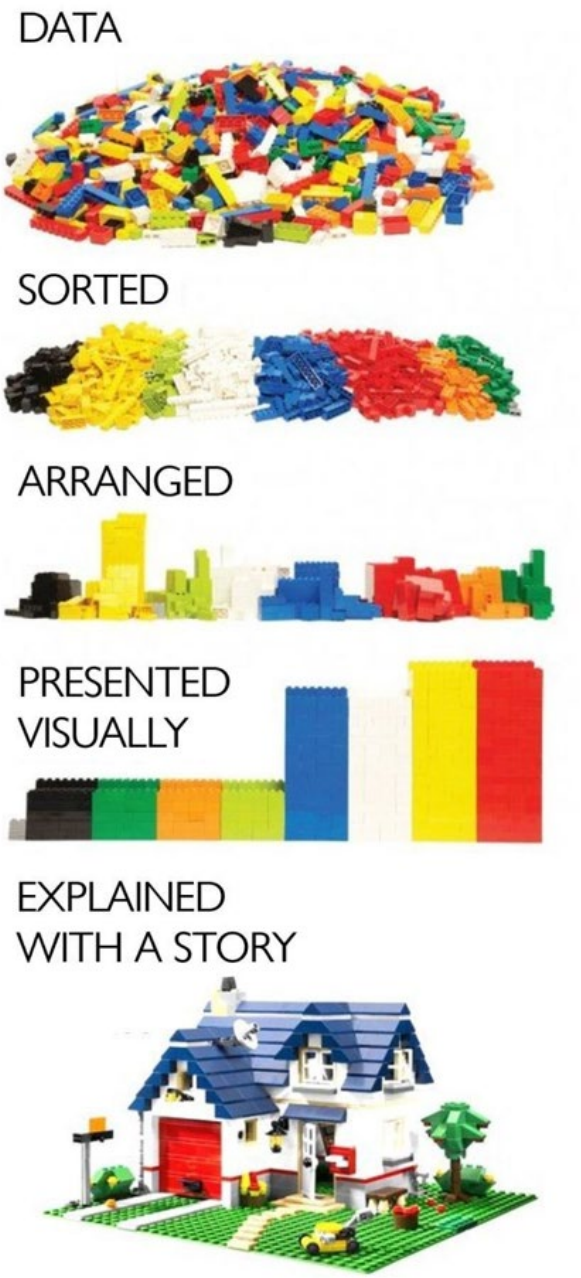
Nog een foutieve conclusie:

Aantal geslaagden over de jaren heen is idem in de opleiding Slavistiek en IT: We vergelijken het slaagcijfer van de opleiding Slavistiek en de opleiding IT. In beide opleidingen zijn er evenveel geslaagden in het eerste jaar: slaagcijfer is gelijkaardig maar…. In de opleiding slavistiek zitten maar 10% van het aantal studenten in de opleiding IT. Let op met het interpreteren van absolute cijfers!

4. Gegevens voorstellen

De meeste statistische informatie in kranten, tijdschriften, rapporten en andere publicaties bestaat uit gegevens die zo zijn samengevat en gepresenteerd dat de lezer ze eenvoudig kan begrijpen. Dergelijke samenvattingen van gegevens, zowel in tabellen, grafieken als getallen, vallen onder wat men noemt data representatie.

In deze paragraaf wordt dieper ingegaan op het presenteren van gegevens in tabellen en de grafische voorstelling ervan.



4.1. Frequentietabel categorische gegevens

Definitie

De absolute frequentie f_i van waarneming x_i is het aantal keer dat deze waarneming voorkomt in de gegevens. De som van de absolute frequentie is gelijk aan het totaal aantal observaties (**n**) : $\sum_{i=1}^{k} f_i = n$

Dataset: slaagcijfers_verkort

Vooropleiding Algemeen	f_i
ASO	1
TSO	15
BSO	4
	20

In bovenstaande dataset komen 15 studenten uit TSO. Dit is zeer veel (in het achterhoofd onthoud je natuurlijk dat dit 15 studenten van de 20 zijn).

Moesten we niet 20 maar 200 studenten geobserveerd hebben, dan is dit weinig.

Vandaar dat het begrip relatieve frequentie in het leven geroepen is.

Definitie

De relatieve frequentie φ_i wordt als volgt berekend: $\varphi_i = \frac{f_i}{n}$ met $\sum_{i=1}^{k} \varphi_i = 1$.

Voorbeeld (slaagcijfers_verkort.xlsx)

- **Vooropleiding Algemeen**
- **Tweede zit**

Vooropleiding Algemeen	f_i	φ_i
ASO	1	0.05
TSO	15	0.75
BSO	4	0.2
	20	**1**

Tweede zit	f_i	φ_i
zwaar	5	0.25
matig	4	0.2
licht	5	0.25
geen	6	0.3
	20	**1**

4.2. Frequentietabel numerieke discrete gegevens

De verschillende gegevens rangschikken we in **stijgende** volgorde zodat we een gerangschikte frequentietabel bekomen.

We vullen de frequentietabel uit vorige paragraaf aan met 2 extra kolommen: de cumulatieve frequenties.

Definitie

Veronderstel dat $x_1, \ldots, x_k$ de k verschillende waarnemingen zijn (gerangschikt van klein naar groot); $f_1, \ldots, f_k$ de corresponderende frequenties en n het totaal aantal gegevens.

Cumulatieve absolute frequentie: $cf_m = \sum_{j=1}^{m} f_j$

Cumulatieve relatieve frequentie: $c\phi_m = \sum_{j=1}^{m} \phi_j = \sum_{j=1}^{m} \frac{f_j}{n}$

Voorbeeld: examenresultaten van 20 studenten (score op 20)

17	7	15	5	7	5	15	16	16	9
11	5	14	7	5	10	5	7	2	2

Score OLOD	f_i	φ_i	cf_i	$c\varphi_i$
2	2	0.1	2	0.1
3	0	0	2	0.1
4	0	0	2	0.1
5	5	0.25	7	0.35
6	0	0	7	0.35
7	4	0.2	11	0.55
8	0	0	11	0.55
9	1	0.05	12	0.6
10	1	0.05	13	0.65
11	1	0.05	14	0.7
12	0	0	14	0.7
13	0	0	14	0.7
14	1	0.05	15	0.75
15	2	0.1	17	0.85
16	2	0.1	19	0.95
17	1	0.05	20	1
	20	**1**		

4.3. Frequentietabel numerieke continue gegevens

Wanneer we continue gegevens hebben of een groot aantal verschillende waarden in discrete gegevens dan heeft het begrip frequentietabel, zoals gehanteerd in vorige paragrafen weinig zin.

Immers alle of haast alle gegevens zullen verschillend zijn zodat alle frequenties f_i bijna gelijk zijn aan 1.

Met andere woorden: de frequentietabel zal geen vereenvoudigde weergave zijn van de echte (= ruwe) gegevens.

Voorbeeld examenresultaten van 20 studenten (score op 20 VOOR afronding)

Zonder klassen

Score OLOD1 (voor afronding)	f_i	φ_i	cf_i	$c\varphi_i$
1.89	1	0.05	1	0.05
2.06	1	0.05	2	0.1
	…	…	…	…
	…	…	…	…
16.86	1	0.05	19	0.95
16.14	1	0.05	20	1
	20	**1**		

Met klassen

Continue waarnemingen alsook discrete met veel verschillende uitkomsten kunnen wel overzichtelijk voorgesteld worden door gebruik te maken van **frequentietabellen met klassenindeling.**

Klassen zijn zelf geconstrueerde intervallen zodat elke waarneming tot één en slechts één klasse behoort. Aangezien er in de praktijk geen strikte afspraken bestaan voor de keuze van deze klassen, volgen we hier één bepaalde methode, nl. deze waarbij het aantal klassen niet minder is dan 5 en niet meer is dan 15, waarbij de klassen allen even breed zijn en waarbij de intervallen links gesloten en rechts open zijn.

Een mogelijke werkwijze voor het opstellen van een frequentietabel met klassenindeling:

- Zoek het grootste en kleinste waarnemingsgetal (in ons voorbeeld is dit 1.89 en 16.86)
- Bereken het verschil tussen deze extreme waarden (16.86 – 1.89 = 14.97).
- Deel dit verschil door 5 en door 15 en kies een klassenbreedte b tussen deze uitkomsten (0.998≤ klassenbreedte ≤ 2.994 ; kies b = 2). ➔ 8 klassen

Score (voor afronding)	f_i	φ_i	cf_i	$c\varphi_i$
[1.89 ; 3.89 [	2	0.1	2	0.1
[3.89 ; 5.89 [	5	0.25	7	0.35
	…	…	…	…
	…	…	…	…
[13.89 ; 15.89 [	2	0.1	17	0.85
[15.89 ; 17.89 [	3	0.15	20	1
	20	**1**		

Definities:

- **Klassengrenzen**: zijn de kleinste en grootste grens van een klasse, in die zin dat de onderste grens in die klasse wel en de bovenste grens niet kan bereikt worden.
  Zo bevat de klasse [15.89; 17.89[alle getallen die groter of gelijk zijn aan 15.89 en strikt kleiner dan 17.89

- **Klassenbreedte**: is het verschil tussen de grootste en kleinste klassengrens van een klasse.

- **Klassenmidden**: is de helft van de som van de grootste en kleinste klassengrens van een klasse.
  Zo is het klassenmidden van de klasse [15.89 ; 17.89[gelijk aan 16.89

- **Klassenfrequentie**: de klassenfrequentie van de i – de klasse is het aantal waarnemingen dat tot deze klasse behoort.

4.4. Grafische voorstelling

Grafieken zijn een efficiënte manier om gegevens voor te stellen. Veelgebruikte types zijn:

- cirkeldiagram (pieplot)
- staafdiagram (barplot)
- histogram
- boxplot (zie ook pg 85)
- spreidingsdiagram of scatterplot

Wij beperken ons tot bovenstaande types en tonen in deze paragraaf hoe deze grafische voorstellingen gemaakt kunnen worden m.b.v. Python.

voorbeeld (slaagcijfers_verkort): Cirkeldiagram van "vooropleiding"

```
import pandas as pd
```

```
#Zorg ervoor dat je notebook en jouw excel in dezelfde map staan
slaagcijfers_kort = pd.read_excel("slaagcijfers_verkort.xlsx")
```

```python
#Cirkeldiagram via "pandas"
#Selecteer kolom "vooropleiding"
vooropleiding = slaagcijfers_kort['vooropleiding']
print(vooropleiding)
```

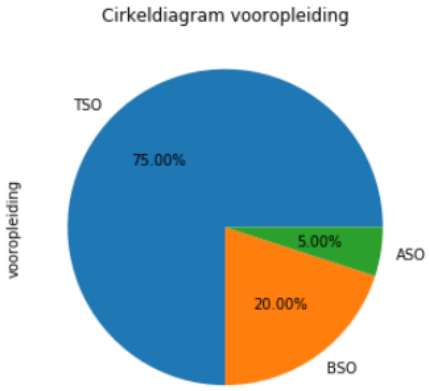
```
0     ASO
1     TSO
2     TSO
3     TSO
4     TSO
5     TSO
6     TSO
7     TSO
8     TSO
9     TSO
10    TSO
11    TSO
12    TSO
13    TSO
14    TSO
15    TSO
16    BSO
17    BSO
18    BSO
19    BSO
```

```python
#via methode value_counts() tellen we de absolute aantallen
data_pie_plot = vooropleiding.value_counts()
print(data_pie_plot)
```

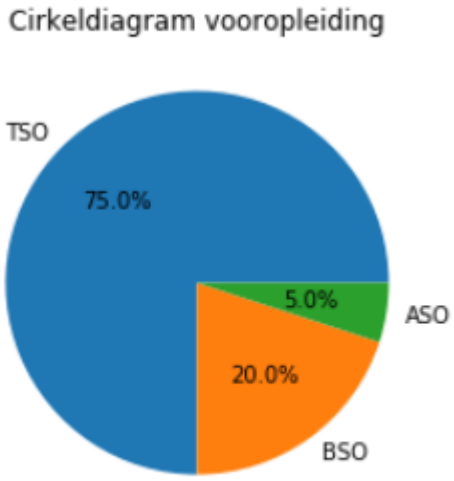
```
TSO    15
BSO     4
ASO     1
Name: vooropleiding, dtype: int64
```

```
#via plot.pie maken we een cirkeldiagram van het categorisch gegeven "vooropleiding"
pie_plot = data_pie_plot.plot.pie(figsize=(5,5), autopct='%3.2f%%', title = 'Cirkeldiagram vooropleiding')
print(pie_plot)
```

```
AxesSubplot(0.135,0.125;0.755x0.755)
```




```
#Alternatief: cirkeldiagram maken via Matplotlib
import matplotlib.pyplot as plt
labels = 'TSO' , 'BSO' , 'ASO'
plt.pie(data_pie_plot, labels = labels, autopct='%1.1f%%')
plt.title('Cirkeldiagram vooropleiding')
plt.show()
```

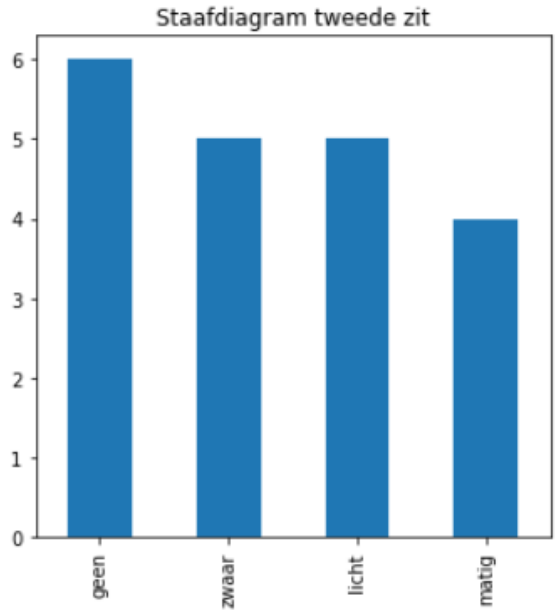


voorbeeld (slaagcijfers_verkort): Staafdiagram van “tweede zit”

```
# Staafdiagram van gegeven tweede zit
tweede_zit = slaagcijfers_kort['tweede zit']
data_bar_plot = tweede_zit.value_counts()
print(data_bar_plot)
```

```
geen     6
zwaar    5
licht    5
matig    4
Name: tweede zit, dtype: int64
```

```
bar_plot = data_bar_plot.plot.bar(figsize=(5,5),title = 'Staafdiagram tweede zit')
```

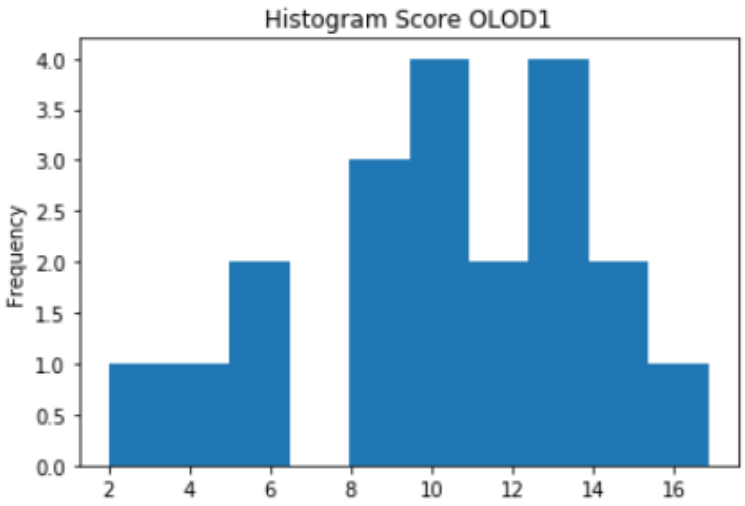


voorbeeld (slaagcijfers_verkort): Histogram van “score OLOD1 (voor afronding)”

```
# Histogram: score OLOD1 (voor afronding)
score_OLOD1 = slaagcijfers_kort['score OLOD1 (voor afronding)']
```

```
score_OLOD1.plot(kind = 'hist', title = 'Histogram Score OLOD1')
```

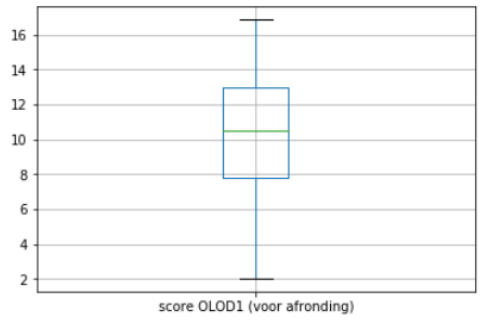
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3a8a50128>
```



voorbeeld (slaagcijfers_verkort): Boxplot van “score OLOD1 (voor afronding)”

```
#Boxplot
slaagcijfers_kort[slaagcijfers_kort['score OLOD1 (voor afronding)'].notnull()].boxplot('score OLOD1 (voor afronding)')
```

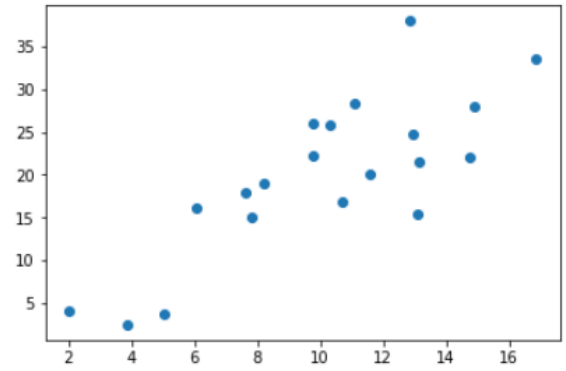
```
<matplotlib.axes._subplots.AxesSubplot at 0x1deb2050208>
```



voorbeeld (slaagcijfers_verkort): Spreidingsdiagram (Scatterplot) van "score OLOD1 (voor afronding)" t.o.v. "uren gestudeerd"

```
#Scatterplot
plt.scatter(slaagcijfers_kort['score OLOD1 (voor afronding)'], slaagcijfers_kort['uren gestudeerd'])
```

```
<matplotlib.collections.PathCollection at 0x1deb2192a58>
```



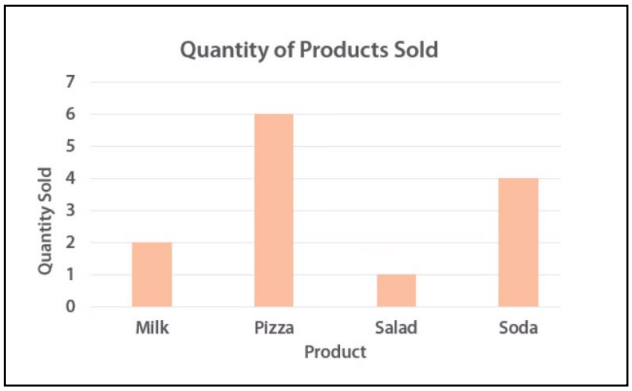
4.5 Verwarrende / foutieve representaties

Zelfde data, andere representatie

Tabel

ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1

Grafisch



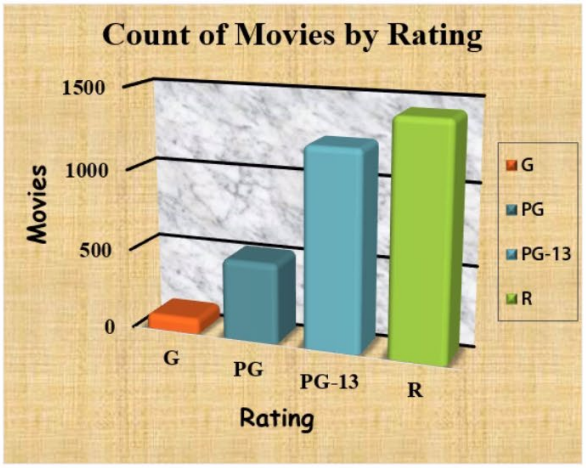
Tabel:

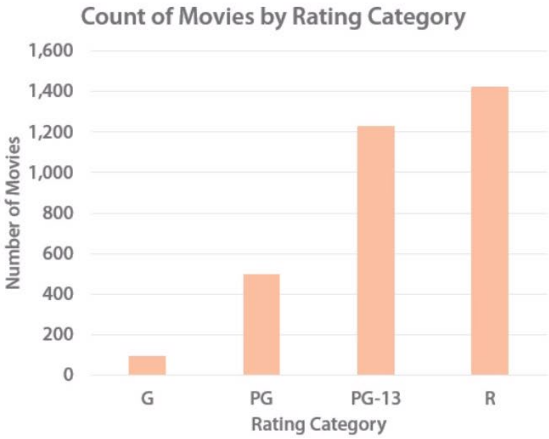
- Meer gegevens beschikbaar: datum – klant …
- Globaal beeld: niet onmiddellijk duidelijk

Grafisch:

- Verlies aan gegevens: datum – klant …
- In één oogopslag is de hoeveelheid in verkoop duidelijk

3D **2D**





3D:

- Oogt mooier
- Minder duidelijk om waardes af te lezen

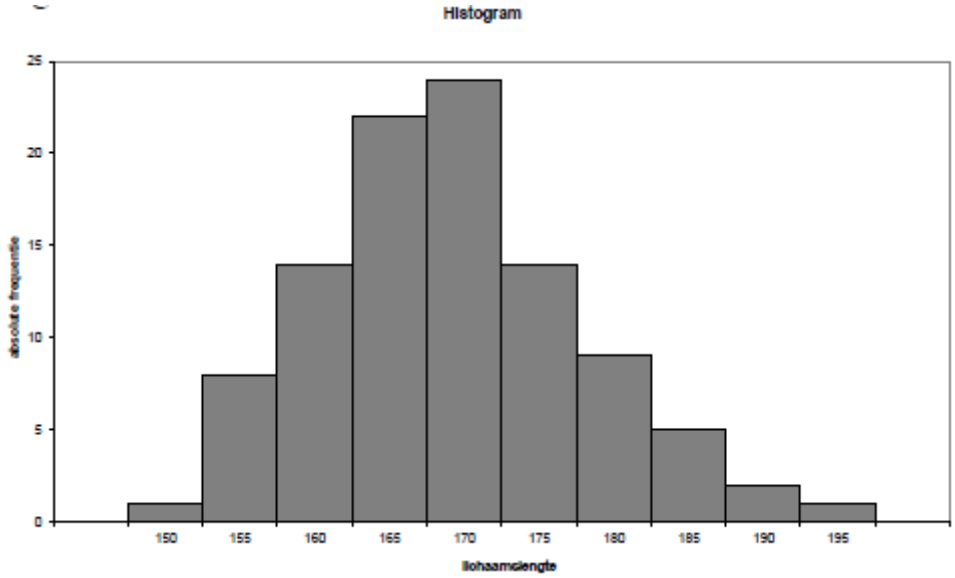
2D:

- Gegevens duidelijk in één oogopslag

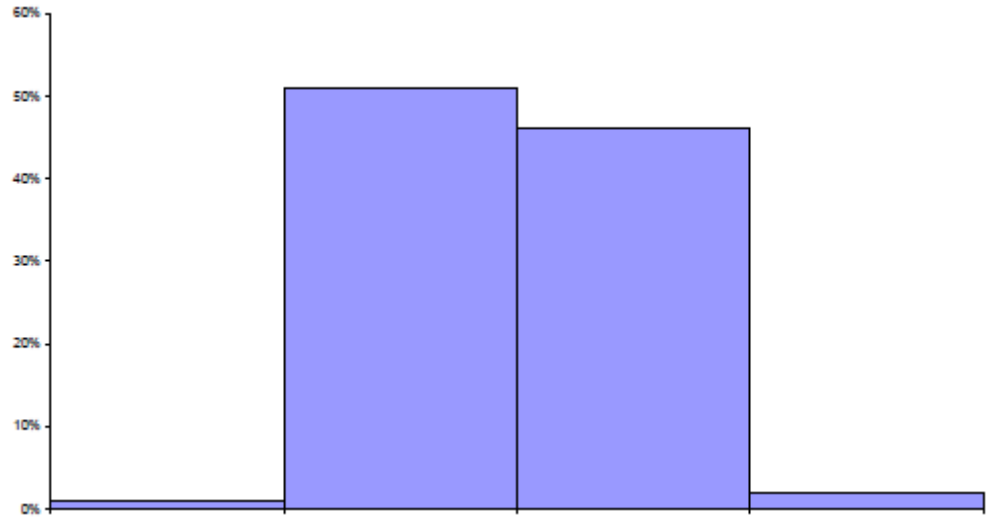
Keuze klassenbreedte

Wanneer de klassenbreedte bij het maken van een frequentietabel anders gekozen wordt, dan ziet het histogram er uiteraard anders uit.

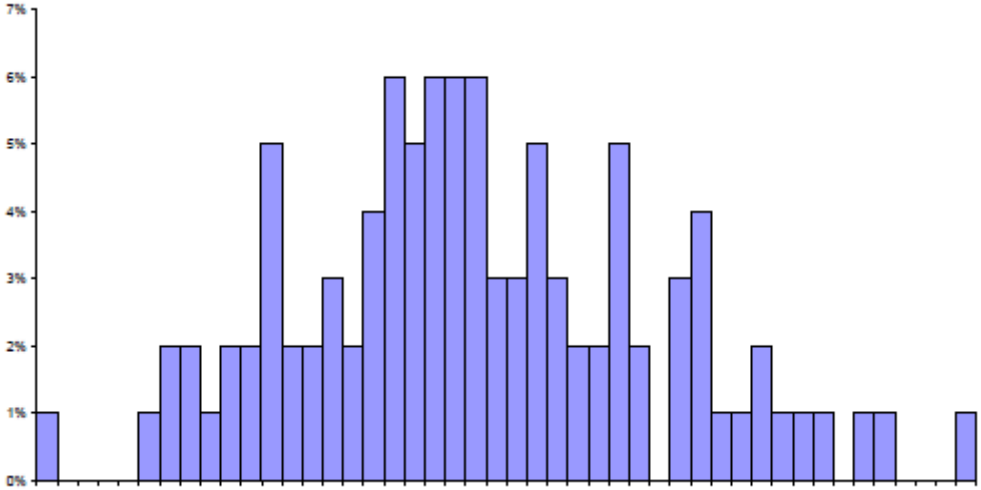
Histogram van de lichaamslengte van 100 2de jaars studenten TIN



Als het aantal intervallen te klein is, dan gaat het algemene patroon van de verdeling verloren.

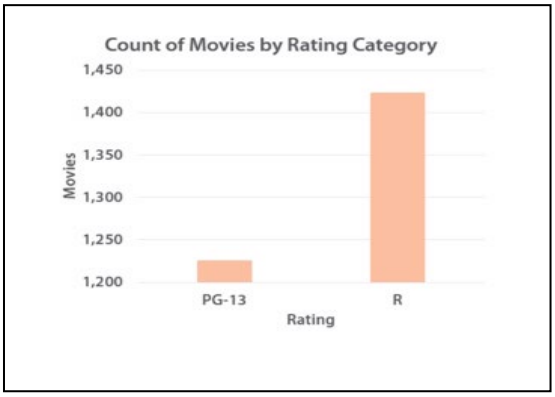


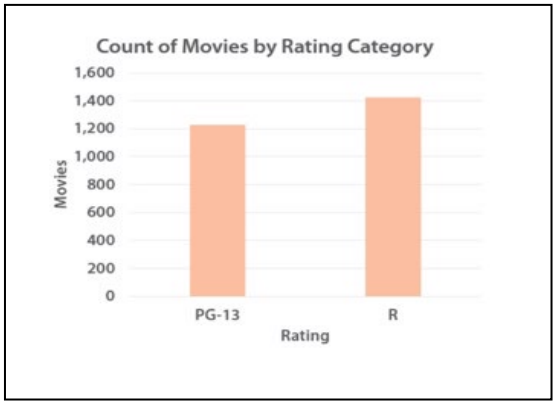
Als het aantal intervallen te groot is, dan komen alle (irrelevante) details (veroorzaakt door het toeval dat in elke dataset aanwezig is) op de voorgrond. Het histogram wordt minder overzichtelijk.



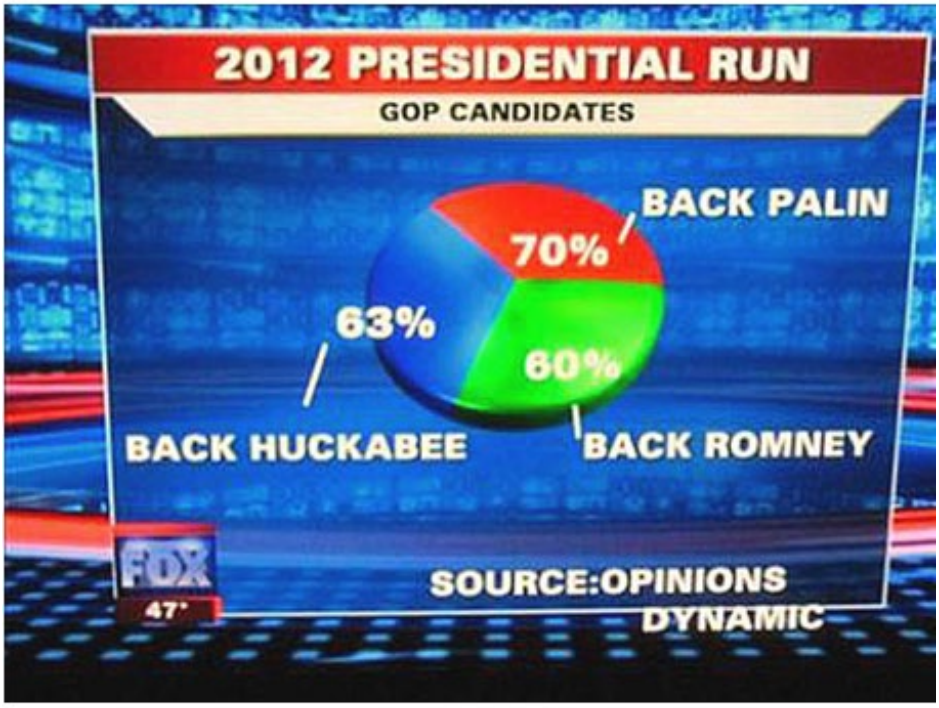
Het is aan te raden de invloed van verschillende klassenbreedten op het uitzicht van het histogram na te gaan.

Waar het fout kan gaan…



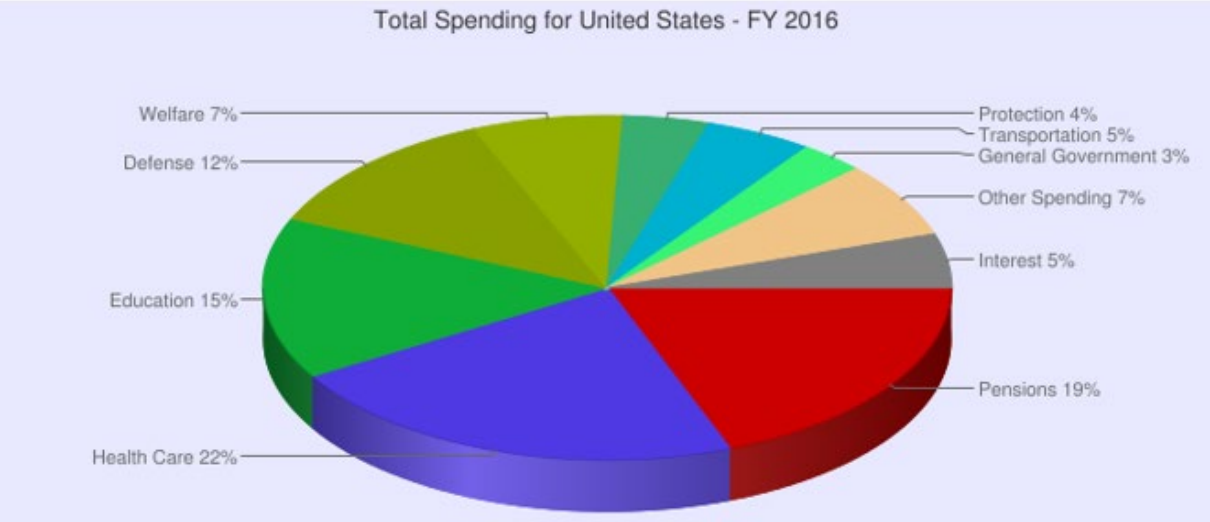


Schaal op y – as is niet oké in grafiek links

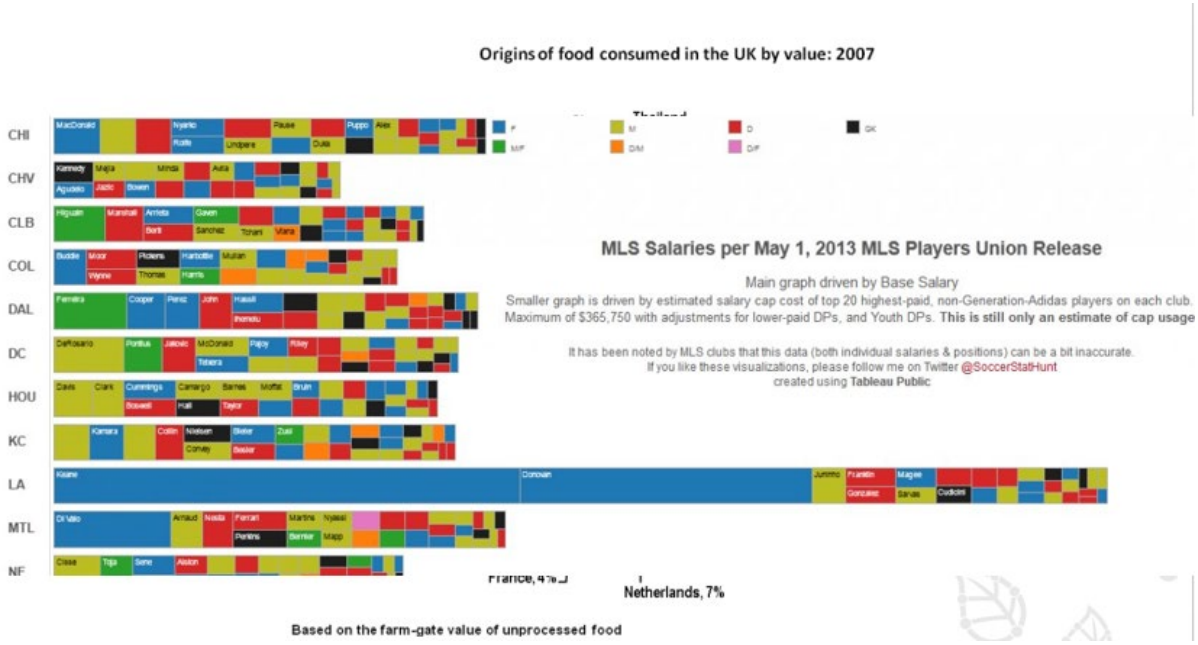


% sommeren niet tot 100…

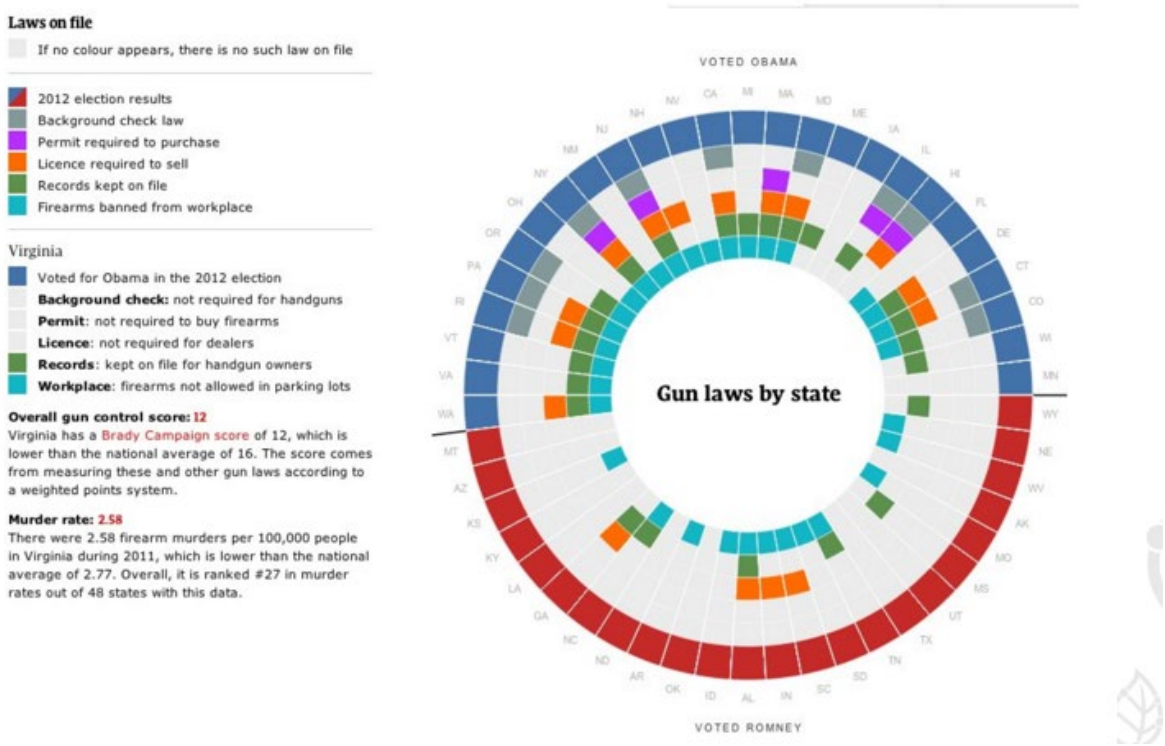
Mogelijkheid tot het steunen van meerdere kandidaten?



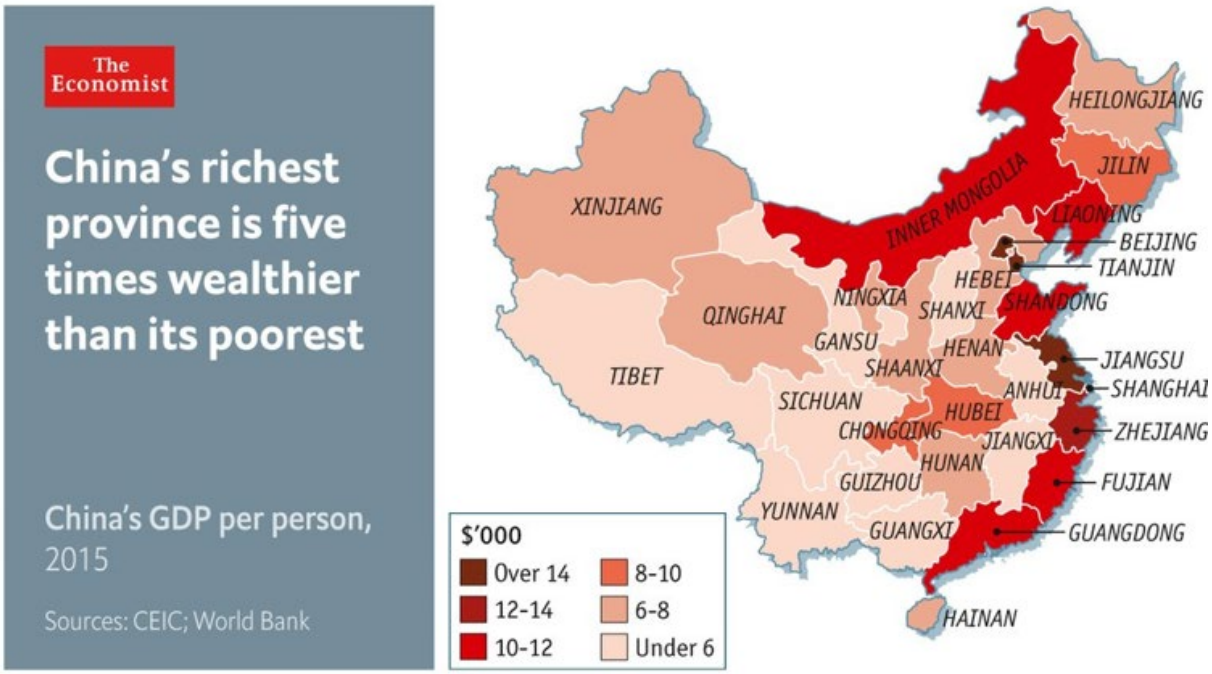
Geen foute grafiek maar zeer moeilijk in te schatten hoe groot de individuele taartstukken zijn. Met een staafdiagram zouden deze gegevens overzichtelijker voorgesteld kunnen worden en daardoor dan ook makkelijker interpreteerbaar.



Teveel informatie in 1 grafiek is nadelig voor de leesbaarheid ervan. Je hebt bijna een “handleiding” nodig om deze grafiek te begrijpen. Niet optimaal.



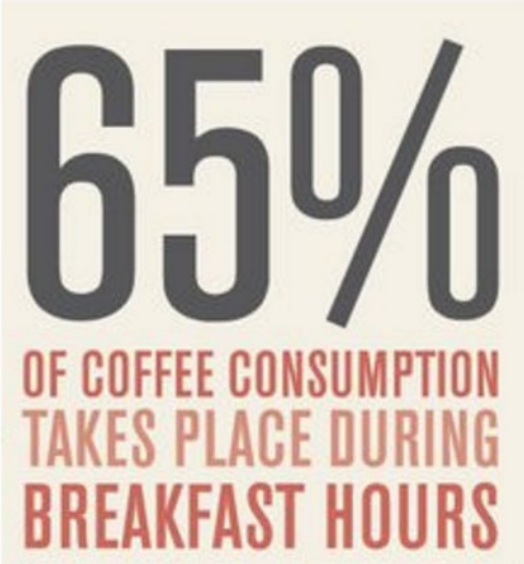
Welke info kunnen we hier uit halen? Onduidelijk wat leesbaarheid betreft.



Duidelijk kleurengebruik; geen overvloed aan informatie in deze grafiek.

Large view (Source: The Economist)

Achtergrond maakt de grafiek minder leesbaar (druk). Grafiek start op 1 gemeenschappelijk punt wat de indruk geeft dat er in New York een lage crime rate is.



1 cijfer kan ook voldoende zijn om een boodschap over te brengen

5. Gegevens samenvatten

Om **numerieke gegevens** verder te analyseren, associeert men er zogenaamde centrum- en spreidingsmaten aan, die de locatie en spreiding van de gegevens weergeven.

5.1. Kengetallen voor locatie

Definitie: rekenkundig gemiddelde

De plaats of locatie van gegevens vatten we samen door één enkel getal: het **rekenkundig gemiddelde**. Dit gemiddelde geeft een soort "midden" aan van de dataset.

Het gemiddelde van n observaties $x_1, \cdots, x_n$ is $\bar{\boldsymbol{x}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i}$

Opmerking:

Indien de gegevens gegeven zijn in een frequentietabel (n gegevens waarvan k verschillend)

Dan wordt het gemiddelde $\bar{x}$ als volgt berekend: $\bar{\boldsymbol{x}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{x_i}\, \boldsymbol{f_i}$

Indien de gegevens gegeven zijn in een frequentietabel met klassenindeling, dan gebruiken we voor de berekening van $\bar{x}$ niet x_i maar het klassenmidden m_i met de bijbehorende klassenfrequenties.

Dit geeft een benadering voor $\bar{\boldsymbol{x}}$: $\bar{\boldsymbol{x}} \approx \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{m_i}\, \boldsymbol{f_i}$

Een groot **nadeel** aan het gemiddelde als spreidingsmaat is dat het zeer gevoelig is voor extremen. Dit wordt duidelijk in volgend voorbeeld:

A: 5 10 5 3 7 $\quad \bar{x}_A = \frac{5+10+5+3+7}{5} = 6$

B: 5 10 1000 3 7 $\quad \bar{x}_B = \frac{5+10+1000+3+7}{5} = 205$

Een mogelijke oplossing is de meest extreme uitkomsten te verwijderen voordat men het gemiddelde bepaalt. Men spreekt dan van een **trimmed mean**.

B: 5 10 3 7 $\quad \bar{x}_B = \frac{5+10+3+7}{4} = 6.25$

Als men over een voldoend aantal waarnemingen beschikt, kan men de hoogste en laagste 5% van de waarnemingen weglaten.

Er zijn situaties waarin de formule van het rekenkundige gemiddelde niet zomaar mag toegepast worden omdat de uitkomsten $x_1, \dots, x_n$ een verschillend gewicht hebben. Bijvoorbeeld bij het berekenen van je eindpercentage op je rapport, krijgt niet elk OLOD hetzelfde gewicht. Dit gewicht is afhankelijk van de studiepunten.

Definitie: gewogen rekenkundig gemiddelde

Het gewogen (rekenkundig) gemiddelde van een reeks numerieke gegevens $x_1, \dots, x_n$ met gewichten $w_1, \dots, w_n$ is de som van alle waarnemingen vermenigvuldigd met het juiste gewicht, gedeeld door de som van de gewichten:

$$\overline{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots w_n x_n}{w_1 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Voorbeeld: bereken het percentage van student 373 (slaagcijfers.xlsx)

Spreidingsmaten voor locatie die niet gevoelig zijn aan uitschieters / extreme waarden, zijn oa de mediaan, de kwartielen en de modus.

Definitie

De mediaan van een rij van n gegevens (**gerangschikt van klein naar groot**) is

De middelste waarde als n oneven is

Het rekenkundig gemiddelde van de middelste twee gegevens als n even is

Voorbeeld

5 gegevens van klein naar groot: 32 42 46 46 54 ➔ mediaan = 46

8 gegevens van klein naar groot: 2 3 4 7 8 10 10 15 ➔ mediaan = (7+8)/2 = 7.5

De zogenaamde kwartielen vormen een uitbreiding van de mediaan. De mediaan verdeelt een geordende rij gegevens in 2 gelijke delen. Wanneer we onze geordende rij gegevens in 4 willen delen, kan dit met behulp van kwartielen:

Definitie: kwartielen

Een kwartiel is één van de drie waarden die een geordende set data in vier gelijke delen opdeelt. Men spreekt van eerste, tweede en derde kwartiel en noteert deze als Q_1 , Q_2 en Q_3

Definitie: modus

De modus is de observatie die het vaakst voorkomt.

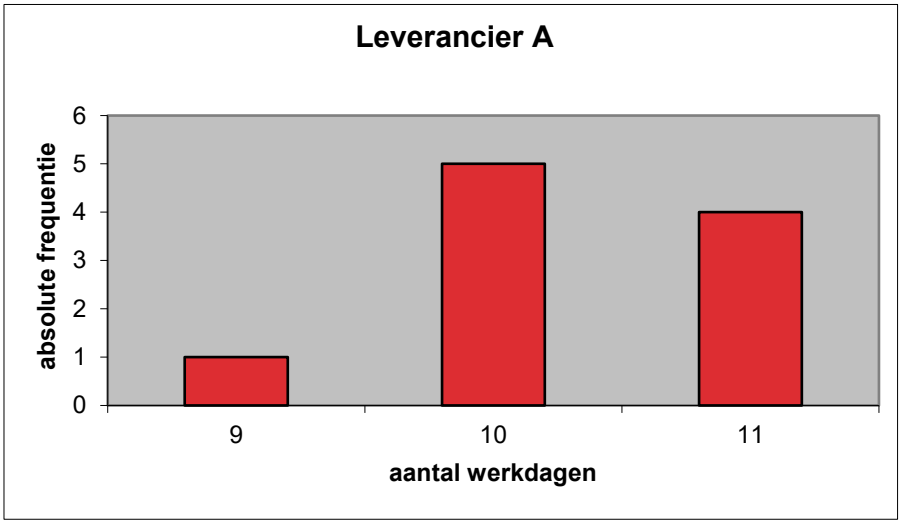
5.2. Kengetallen voor spreiding

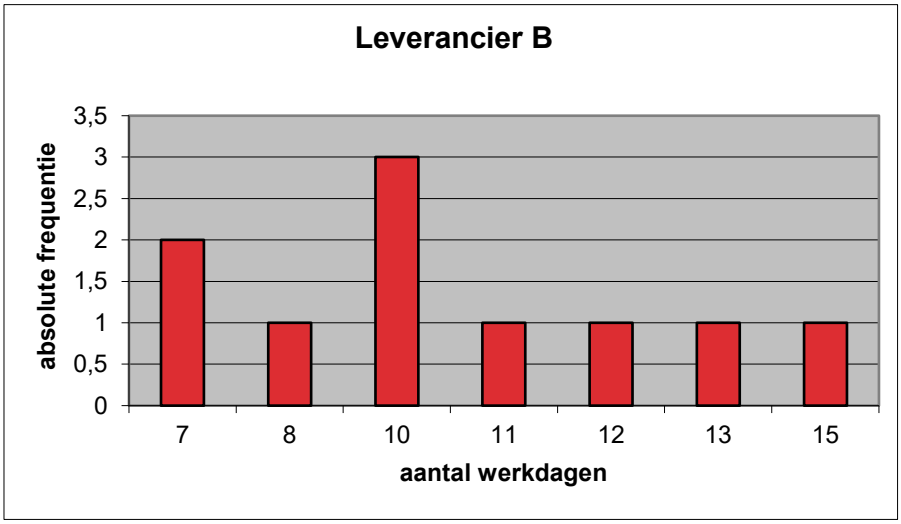
Voorbeeld

De afdeling inkoop van een grote fabrikant plaatst regelmatig bestellingen bij twee verschillende leveranciers. Volgende tabel geeft het aantal dagen weer dat nodig was om 10 bestellingen te leveren.

A:	11	10	9	10	11	11	10	11	10	10
B:	8	10	13	7	10	11	10	7	15	12

Wanneer we het gemiddelde aantal werkdagen berekenen zowel voor leverancier A als B, bekomen we 10.3. Toch kunnen we niet stellen dat beide leveranciers even betrouwbaar zijn wat betreft het op tijd leveren. Aan volgende staafdiagrammen zien we dat de spreiding rond het gemiddelde van deze twee leveranciers verschillend is (de spreiding bij leverancier B is groter dan bij A).





Als we willen weten hoe sterk de individuele gegevens $x_1, \cdots, x_n$ afwijken van hun gemiddelde maken we gebruik van de variantie.

Definitie

- De **variantie** s^2 van de gegevens $x_1, \cdots, x_n$ met gemiddelde $\bar{x}$ is gelijk aan

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

- De standaardafwijking s van de gegevens $x_1, \cdots, x_n$ is de positieve vierkantswortel uit de variantie

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

Definitie: spreidingsbreedte R (Range)

De **spreidingsbreedte** van n gegevens $\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}$: R is grootste waarde – kleinste waarde.

Net zoals het gemiddelde is de spreidingsbreedte uiterst gevoelig voor uitschieters.

Definitie

De interkwartielafstand van n gegevens $x_1, \cdots, x_n$ is gedefinieerd als

$$\text{IKA} = Q_3 - Q_1$$

De interkwartielafstand geeft de spreidingsbreedte weer van de middelste 50% van de gegevens en is dus niet gevoelig aan uitschieters.

5.3. Visuele voorstelling van locatie en spreiding

De boxplot is een eenvoudige grafische samenvatting van enkele belangrijke kengetallen van een gegevensset. Met een boxplot kan in één oogopslag informatie over locatie en spreiding van de verschillende gegevensverzamelingen vergeleken worden.

De boxplot wordt getekend mbv. de 5 volgende getallen (**vijf - getallenrésumé**)

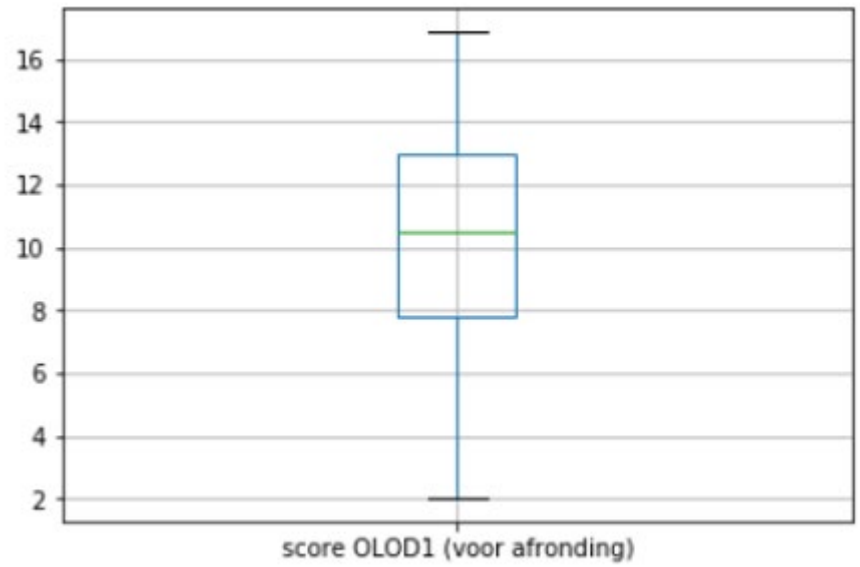
- het kleinste gegeven
- het eerste kwartiel Q_1
- de mediaan of tweede kwartiel Q_2
- het derde kwartiel Q_3
- het grootste gegeven

Tussen elk van de opeenvolgende getallen ligt telkens 25% van de gegevens.

Boxplots zijn interessant wanneer je meerdere verzamelingen van gegevens met elkaar wil vergelijken.

Je kan deze onder mekaar zetten of “kantelen”.

Voorbeeld



Een boxplot geeft ons informatie over:

- centrum: mediaan (eventueel gemiddelde)
- spreiding: de lengte van de box is de IKA
- Scheefheid: de positie van de mediaan ten opzichte van het eerste en het derde kwartiel geeft reeds aan of er asymmetrie is of niet. Bij symmetrische verdelingen (zie later) ligt de mediaan in het midden van de box, bij rechtsscheve verdelingen in de onderste helft en bij linksscheve verdelingen in de bovenste helft van de box. Bovendien kunnen we naar de lengtes van de 'takken' kijken. Rechtse scheefheid geeft aanleiding tot boxplots waarbij de bovenste tak langer is dan de onderste. Bij linkse scheefheid is dit natuurlijk andersom
- uitschieters

6. Verbanden tussen variabelen

Bij onderzoek naar het verband tussen twee variabelen onderscheiden we een aantal vragen:

1. Theoretisch: ontdek je een verband in de analyses?
2. Kracht: hoe sterk is het verband tussen die variabelen?
3. Richting: is het verband positief of negatief?
4. Causaliteit: is er sprake van een oorzakelijk verband?

6.1 Twee categorische variabelen

Kruistabellen zijn handige hulpmiddelen om het verband of associatie tussen twee **categorische** variabelen te analyseren. Zo'n tabel telt hoeveel keer elke combinatie van twee variabelen voorkomt. Dit zijn **absolute frequenties**. Bovendien kunnen er rij- en kolomtotalen aan de tabel toegevoegd worden

In plaats van te werken met absolute frequenties kunnen we de kruistabel ook voorstellen aan de hand van **relatieve frequenties**, waarbij we dan elk element in de tabel delen door het totaal aantal observaties.

Voorbeeld:

We willen de samenhang nagaan tussen Hogeschool waar men gestudeerd heeft en werkprestatie. Van drie hogescholen (A, B en C) worden 30 studenten per hogeschool geselecteerd. Bedrijven stellen aan de hand van een beoordelingslijst vast hoe de student functioneert in dat bedrijf. We gaan ervan uit dat er geen verschil is in wijze van beoordelen tussen de bedrijven.

De kruistabel van bovenstaande gegevens ziet er als volgt uit:

hogeschool prestatie	A	B	C	All
goed	25	25	25	75
middelmatig	4	4	4	12
slecht	1	1	1	3
All	30	30	30	90

De verdeling van de werkprestaties bij elke Hogeschool is hetzelfde. De twee variabelen ‘type Hogeschool’ en ‘werkprestatie’ hangen samen indien het kennen van de variabele ‘type Hogeschool’ informatie geeft over de variabele ‘werkprestatie’. Daar is in dit geval **geen sprake van**! Als je een willekeurige student van bvb Hogeschool B tegenkomst, kan je geen uitspraak doen over de werkprestatie waarmee deze school zich onderscheidt van de andere. De werkprestatie zal naar alle waarschijnlijkheid goed zijn. Maar dat is ook het geval bij de andere Hogescholen. Het omgekeerde geldt ook niet: van iemand die goed functioneert in het bedrijf kan niet eenduidig gezegd worden van welk type Hogeschool hij / zij komt. Er bestaat dus geen verband tussen ‘type Hogeschool’ en ‘werkprestatie’; anders gezegd **het verband of correlatie is 0**. De variabele ‘werkprestatie’ is onafhankelijk van de variabele ‘type Hogeschool’. Er is in dit geval geen correlatie tussen deze twee variabelen.

Merk op dat in dit geval de verhouding van de frequenties binnen de kolom voor alle kolommen gelijk is aan 25 : 4 : 1 : 30. Voor de rijen is die verhouding steeds gelijk aan 1 : 1 : 1 : 3. Een **typisch kenmerk voor onafhankelijkheid** van variabelen.

Relatieve frequenties

In dit voorbeeld is het aantal studenten dat per Hogeschool geselecteerd is gelijk per Hogeschool nl 30. In het geval dat **deze aantallen niet gelijk** zijn, werk je beter met relatieve frequenties.

Plaatsen we in de cellen van de kruistabel in plaats van absolute frequenties f_i de relatieve frequenties φ_i, dan ziet de kruistabel er als volgt uit:

hogeschool prestatie	A	B	C	All
goed	0.277778	0.277778	0.277778	0.833333
middelmatig	0.044444	0.044444	0.044444	0.133333
slecht	0.011111	0.011111	0.011111	0.033333
All	0.333333	0.333333	0.333333	1.000000

Een **regel voor onafhankelijkheid** op basis van relatieve frequenties: de **relatieve frequenties in de cellen blijken gelijk te zijn aan de producten van de bijbehorende rij- en kolomtotalen**:

0.2777 = 0.3333 * 0.8333 en 0.0444 = 0.3333 * 0.1333

Dit alles samengevat in 3 regels: onafhankelijkheid tussen twee **categorische** variabelen (aan de hand van een kruistabel):

- De frequenties (of relatieve frequenties) hebben in elke kolom dezelfde verhouding ofwel
- De frequenties (of relatieve frequenties) hebben in elke rij dezelfde verhouding ofwel
- De relatieve frequentie in elke cel is gelijk aan het product van de bijbehorende rij- en kolomtotalen

Als één van deze drie regels geldt, gelden ook automatisch de andere twee.

Stel nu dat de verdeling eruit ziet als in volgende kruistabel:

hogeschool / prestatie	A	B	C	All
goed	25	1	4	30
middelmatig	4	4	25	33
slecht	1	25	1	27
All	30	30	30	90

Als je weet van welke Hogeschool een student komt, kan je iets zeggen over de werkprestatie van die student. Er bestaat nu **wel een verband** of correlatie tussen 'type Hogeschool' en 'werkprestatie'. Het verband is weliswaar niet perfect (in real – life is een verband nooit perfect), maar duidelijk aanwezig: informatie over schooltype levert ook enige informatie over werkprestatie. Kom je nl een willekeurige student tegen van Hogeschool B, dan is het erg waarschijnlijk dat zijn/haar werkprestatie slecht is. Kom je een student tegen van resp. Hogeschool A/C, dan is het erg waarschijnlijk dat hun werkprestatie goed respectievelijk middelmatig is. Het omgekeerde geldt ook: kom je een goede student tegen, dan is deze hoogstwaarschijnlijk afkomstig van Hogeschool A.

Van een perfecte samenhang is sprake in onderstaande kruistabel. Werkprestatie hangt volledig af en type Hogeschool en omgekeerd.

hogeschool / prestatie	A	B	C	All
goed	30	0	0	30
middelmatig	0	0	30	30
slecht	0	30	0	30
All	30	30	30	90

Naast een kruistabel zijn er ook **grootheden** die je kan berekenen om het verband tussen categorische variabelen na te gaan (bvb Cramers V; hier gaan wij niet verder op in.

6.2 Twee numerieke variabelen

Wanneer we te maken hebben met numerieke variabelen (continue gegevens of discrete gegevens met een groot aantal verschillende waarden), kunnen we niet op voorgaande manier werken. De kruistabel zou immers al snel heel onoverzichtelijk worden.

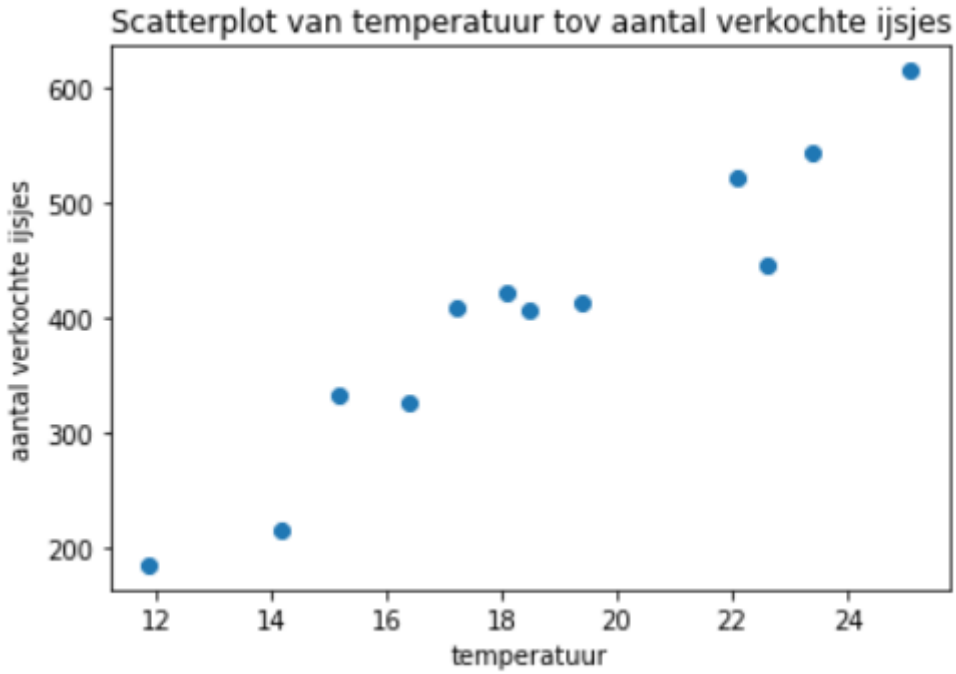
Om een mogelijk verband te detecteren tussen twee numerieke variabelen kunnen we gebruik maken van een **spreidingsdiagram (scatterplot).** Een spreidingsdiagram is een tweedimensionale grafiek, met de waarden van de ene variabele langs de verticale as, en de waarden van de andere langs de horizontale as.

Als een toename in één variabele in het algemeen samenhangt met een toename in de tweede variabele, zeggen we dat de twee variabelen **'positief samenhangen' of 'positief correleren'**. Een andere mogelijkheid is dat één variabele de neiging heeft af te nemen, naarmate de andere toeneemt. We zeggen dan dat de variabelen **'negatief correleren'**.

Uiteraard zijn er ook situaties waarbij er geen relatie bestaat tussen de twee variabelen.

Voorbeeld

De lokale ijsjeszaak heeft de afgelopen weken bijgehouden hoeveel ijsjes er verkocht werden alsook de middagtemperatuur.



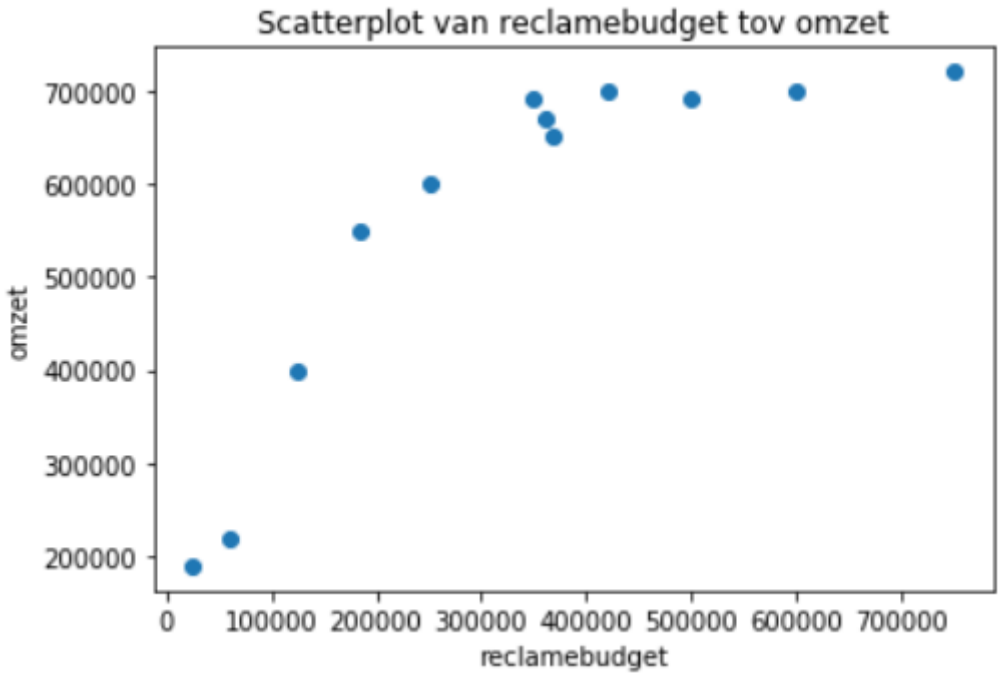
Het verband in voorgaande grafiek kan blijkbaar redelijk goed omschreven worden door een **rechte**: we spreken dan van een **lineair verband**.

Naast grafische manieren zijn er ook kengetallen die correlaties tussen variabelen weergeven bvb. de correlatiecoëfficiënt van Pearson. Deze geeft aan hoe sterk het lineair verband tussen twee variabelen is. Deze coëfficiënt ligt steeds tussen -1 en 1. Wanneer deze coëfficiënt 0 is, is er geen lineair verband tussen de 2 variabelen.

Een “heatmap” in python geeft je een overzichtelijk beeld van de correlaties tussen verschillende variabelen.

Verder hoeft het verband tussen twee variabelen uiteraard niet altijd lineair te zijn.

Voorbeeld



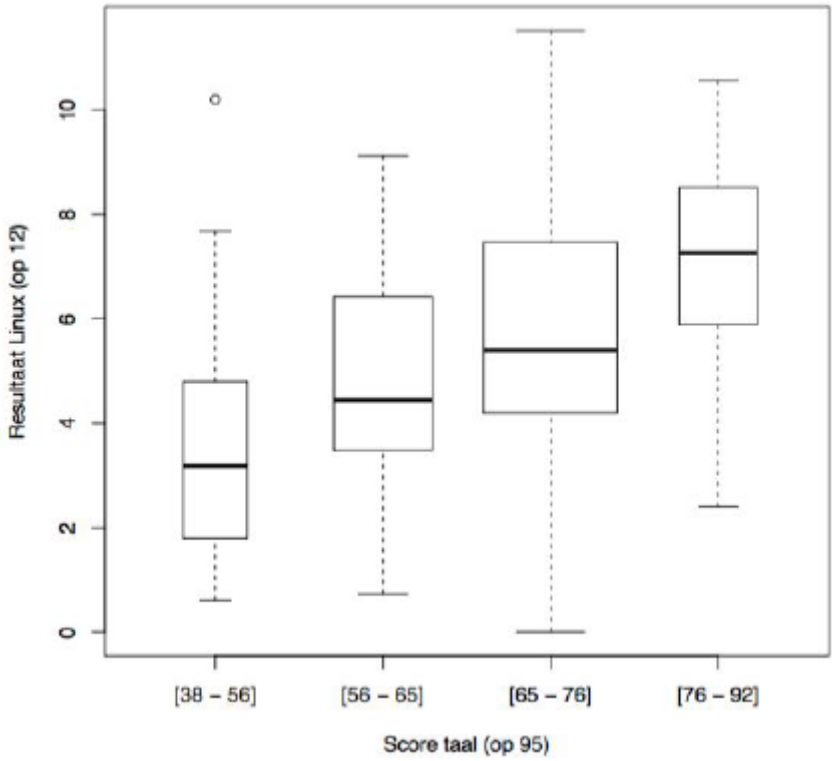
Let op voor

- Aanwezigheid van **uitschieters (zie pg 96)** in de dataset (zeker als het aantal gegevens beperkt is). Uitschieters kunnen zorgen voor een “verkeerdelijk” verband tussen variabelen.

- **Verstrengelende factoren**: een ijsverkoper aan zee beweert dat er een causaal verband is tussen het aantal mensen dat in zee zwemt en het aantal ijsjes dat hij verkoopt. Dit is een foutieve conclusie. Het aantal mensen dat in zee zwemt en het aantal ijsjes dat aan zee verkocht wordt, wordt nl beïnvloed door een andere variabele: temperatuur. Omdat beiden afhangen van temperatuur, lijkt er een verband te zijn tussen beide. Temperatuur noemen we een verstrengelende factor. **Correlatie impliceert geen causaliteit. Een veelgemaakte fout is dan ook “samenhangen” te verwarren met “veroorzaken”.**

6.3 Een categorische en een numerieke variabele

Om het verband tussen één categorische variabele en één numerieke variabele na te gaan, maken we gebruik van een boxplot (zie eerder).



7. Missing Values

Missing values treden op wanneer sommige respondenten niet antwoorden op bepaalde vragen of wanneer de antwoorden niet opgenomen worden in de dataset. Het is belangrijk om het concept van missing values (ontbrekende waarden) te begrijpen bij gegevensverwerking. Missing values “negeren”, kan leiden tot onnauwkeurige en zelfs compleet foutieve conclusies over de gegevens.

Verschillende soorten missing values

Wij bespreken twee soorten missing values: MCAR (missing completely at random) en MAR (missing at random).

MCAR: de missing values zijn willekeurig verspreid zijn over alle observaties

MAR: de missing values zijn niet willekeurig verspreid over de observaties maar komen voor in bepaalde delen van de dataset (bvb enkel bij mannen indien geslacht aanwezig is in de dataset). Deze vorm treedt vaker op dan MCAR.

8. Uitschieters (Outliers)

Definitie

Een waarneming (of meetwaarde) die ongewoon klein of groot is ten opzichte van de overige waarden in een gegevensverzameling wordt een outlier (uitschieter) genoemd.

Outliers zijn in het algemeen het gevolg van één van de volgende oorzaken:

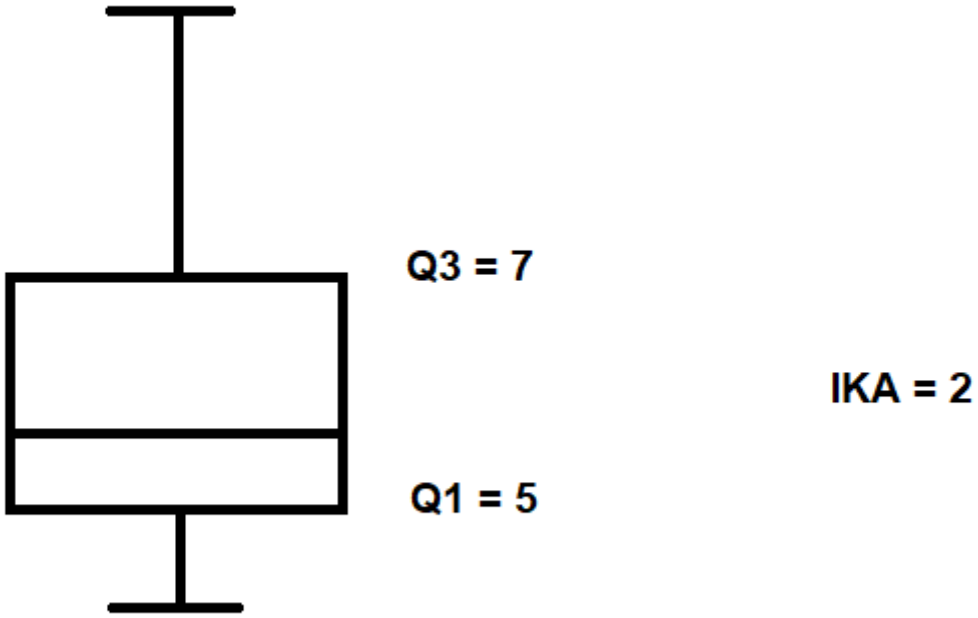
- De meetwaarde is onjuist waargenomen, geregistreerd of in de computer ingevoerd

- De meetwaarde is juist maar vertegenwoordigt een (toevallige) zeldzame gebeurtenis

Twee bruikbare methodes voor het opsporen van outliers zijn box-plots en z – scores. Wij beperken ons tot boxplots.

Herinner: de boxplot is gebaseerd op de kwartielen van een gegevensverzameling. Kwartielen zijn waarden die de gegevensverzameling verdelen in vier groepen, die elk 25% van de meetwaarden bevatten. De boven- en onderzijde van de box zijn getekend ter hoogte van het eerste resp. derde kwartiel. Een uitschieter wordt gedefinieerd als een waarnemingsgetal dat minstens 1,5 keer de interkwartielafstand beneden $\boldsymbol{Q_1}$of boven $\boldsymbol{Q_3}$ligt. De boxplot bevat 2 categorieën van uitschieters:

- Extreme waarden liggen meer dan 3 keer de box-lengte boven of onder de box
- Waarden die tussen 1,5 en 3 keer de box-lengte verwijderd zijn van de box, noemt men outliers



Omgaan met missing values / outliers

Er zijn verschillende manieren van missing value /outlier correctie, ieder met voor- en nadelen.

- **Negeer missing values / outliers:** één mogelijkheid is om de waarde te negeren en de data precies te laten als het is.

- De **variabele (feature) verwijderen (kolom)**: een variabele met veel missing values / outliers kan worden verwijderd. Dit kan natuurlijk alleen als de variabele niet essentieel is voor het beantwoorden van de onderzoeksvraag. Zeker als er veel variabelen zijn gemeten kan deze methode worden overwogen.

- **De observatie (met missing values / outliers) verwijderen (rij):** Veronderstel dat het aantal missing values / outliers extreem klein is (in verhouding met de totale dataset), dan kan de onderzoeker beslissen om de observaties waar missing values / outliers in voorkomen gewoon te "deleten" of te negeren in bepaalde analyses. Vaak wordt hiervoor de 5% regel gebruikt: indien de missing values / outliers minder dan 5% van de gehele dataset betreffen, kunnen we deze observaties weglaten. Een nadeel van deze aanpak is het risico op 'selection bias', het fenomeen waarbij alleen de datapunten gunstig voor de plannen van de onderzoeker worden geselecteerd ("Een onderzoeker kan bewijzen wat hij wil bewijzen indien hij maar de juiste data gebruikt!"). Zo verteken je de data dus in je eigen voordeel. Deze aanpak, hoewel makkelijk in gebruik, is de meest drastische methode van omgaan met missing values / outliers.

- Waardes **imputeren** (‘imputing values’): het imputeren van data houdt in dat de waarde van een bepaald datapunt (de missing value / outlier) wordt vervangen door een nieuwe, voorspelde waarde. Deze techniek wordt vaak gebruikt voor missing data. Er zijn verschillende manieren om de te imputeren waarde(s) te voorspellen; zo kan het gemiddelde van de dataset zonder de missing values / outliers worden genomen als nieuwe waarde (dit is een slordige aanpak), of er kan op de rest van de data een toepasselijk model worden gefit waarmee de nieuwe waarde van missing values / outliers kan worden voorspeld. De imputatie vereist wel dat er genoeg vergelijkbare data is om een voorspelde waarde op te baseren.

- **Transformeren**: het transformeren van data houdt in dat de waarde van ieder datapunt van een variabele/dataset volgens dezelfde wiskundige functie wordt omgezet naar een andere waarde. Zo blijft de onderliggende datastructuur hetzelfde terwijl andere eigenschappen van de data, zoals de verdeling of de variantie, kunnen veranderen. Er zijn veel verschillende vormen van datatransformatie, ieder met eigen voorwaarden, voordelen en nadelen. Wij behandelen enkel:
    - **Binning**: Numerieke data wordt omgezet naar categorische data. Outliers hebben minder effect op categorische data.

Voorbeeld

	Score / 20	Imputation (gem)	Transformation (binning)
Student 1	**0**	**0**	**[0 , 12 [**
Student 2	**18.01**	**18.01**	**[18 , 21 [**
Student 3	**15.18**	**15.18**	**[15 , 18 [**
Student 4	**15.77**	**15.77**	**[15 , 18 [**
Student 5	**16.32**	**16.32**	**[15 , 18 [**
Student 6	**??**	**13.06**	**[15 , 18 [**

We zien dat student 1 een outlier is. Deze heeft een groot effect het gemiddelde:

Zonder Student 1: 16.32 - Met Student 1: 13.06

We willen de missing value van Student 6 opvullen.

Wanneer we **Imputation** toepassen zonder aandacht te besteden aan de outlier, kunnen we Student 6 het gemiddelde van de andere 5 studenten geven: 13.06

Gemiddelde (alle studenten) na imputation : 13.06

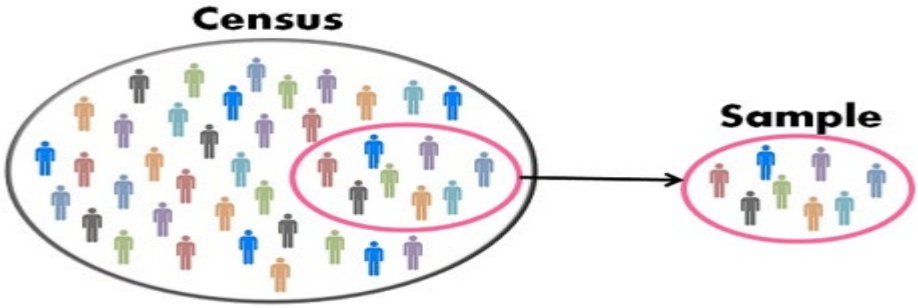
Om het effect van de outlier te minimaliseren, passen we eerst **Transformation (binning)** toe**:** we maken van de continue variabele een categorische variabele (met intervallen met ongelijke lengte). Het gemiddelde, berekend op de categorieën, van de andere 5 studenten is 15: we kennen categorie [15 , 18 [toe aan Student 6.

Gemiddelde (alle studenten) na transformation: 15.25 (de outlier heeft minder effect).

9. Populatie – steekproef

Veel onderzoek is gebaseerd op steekproeven: op basis van gegevens van een kleine groep observaties worden conclusies getrokken die gelden voor de hele bevolking (= populatie). De populatie is het grotere geheel waarvan we bepaalde karakteristieken te weten willen komen, maar die we om allerlei redenen (van praktische, financiële, … aard) niet volledig kunnen observeren.

Die kleine groep moet dan wel goed gekozen worden, zodat de groep representatief is voor de populatie waarover uitspraken worden gedaan.



Van deze steekproeven bekijken we enkele belangrijke grootheden: het steekproefgemiddelde (schatter voor het populatiegemiddelde) en de steekproefvariantie (schatter voor de populatievariantie).

Als de samenstelling van de te onderzoeken populatie verhoudingsgewijs overeenkomst met de samenstelling van de steekproef, dan spreken we van een **selecte of** gerichte steekproef.

Bij een **aselecte steekproef** nemen we de te onderzoeken elementen willekeurig uit de populatie. Elk element heeft dezelfde kans om in de steekproef opgenomen te worden. We laten als het ware het lot beslissen om een representatieve steekproef te bekomen.

Voorbeeld:

Een sigarettenproducent wil de gemiddelde hoeveelheid teer meten in een nieuw merk sigaretten. Hij neemt hiervoor een steekproef van 100 sigaretten, laat de hoeveelheid teer meten (mg) en berekent het gemiddelde. Dit leidt tot een steekproefgemiddelde van 14.8mg; wat een schatting geeft van het gemiddelde van de hele populatie (= alle sigaretten van dit merk). Het is duidelijk dat deze schatting toevallig is, een andere steekproef zal waarschijnlijk een ander gemiddelde opleveren en dus een andere schatting.

	Steekproef	Populatie
Grootte	n	N
Gemiddelde	$\overline{X}\ (\bar{x})$	μ
Variantie	$S^2\ (s^2)$	σ^2
Standaard afwijking	$S\ (s)$	σ

10. De dichtheidsfunctie bij numerieke gegevens

In voorgaande hebben we datasets (steekproef) beschreven aan de hand van kengetallen voor locatie - kengetallen voor spreiding - grafische methodes…

Bij grotere datasets (de populatie in het achterhoofd houdend) krijgen we niet altijd een goed beeld mbv bovenstaande methoden:

- Frequentietabellen worden (te) lang waardoor details onderdrukt worden of afhankelijk worden van de klasseindeling.
- Staafdiagrammen / histogrammen worden te uitgebreid…

De vraag die we ons kunnen stellen is of er een manier bestaat die toelaat aan de hand van één uitdrukking de volledige dataset (ook wel verdeling genoemd) te beschrijven. Deze "ideale" beschrijving (uitschieters en kleine onregelmatigheden buiten beschouwing gelaten) noemen we de **dichtheidsfunctie**.
De dichtheidsfunctie volgt het algemeen patroon van de gegevens maar verwaarloost onregelmatigheden en uitschieters.

Voorbeeld

Volgend histogram geeft het aantal mensen (in 10-tallen) op een observatie van 500 mensen dat bereid is gedurende een bepaalde tijd te wachten wanneer je naar een call-center telefoneert.

We zien dat deze gegevens een regelmatig patroon vertonen.

Het histogram is min of meer symmetrisch: de top ligt ongeveer in het midden en beide staarten gaan geleidelijk naar 0.

Op basis van dit histogram kunnen we een antwoord geven op volgende vragen:

- Hoeveel % van de mensen wil ten hoogste 6 minuten wachten?

- Hoeveel % van de mensen wacht tussen 4 en 6 minuten?

De vloeiende curve (**dichtheidsfunctie**) die doorheen dit histogram getekend is, is een goede beschrijving van het algemeen patroon van deze gegevens.

Nu blijkt dat wanneer de oppervlakte onder dichtheidsfunctie berekend wordt, ongeveer dezelfde resultaten bekomen worden.

Deze oppervlaktes kunnen heel eenvoudig via Python berekend worden.

```
import scipy.stats as stats
import math
```

- Hoeveel % van de mensen wil ten hoogste 6 minuten wachten?

```
stats.norm.cdf(6,5,math.sqrt(4))
```

```
0.6914624612740131
```

- Hoeveel % van de mensen wacht tussen 4 en 6 minuten?

```
stats.norm.cdf(6,5,math.sqrt(4)) - stats.norm.cdf(4,5,math.sqrt(4))
```

```
0.38292492254802624
```

We komen in het verloop van dit deel nog terug op bovenstaande commando's.

11. Soorten dichtheidsfuncties en verdelingen

Verdelingen (dataset) en dus ook dichtheidsfuncties kunnen in verschillende gedaanten voorkomen.

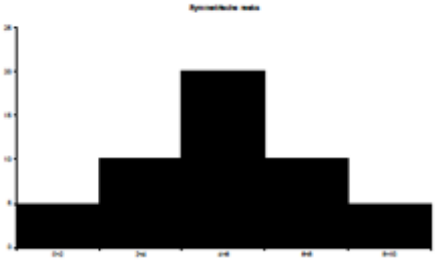
Voorbeeld

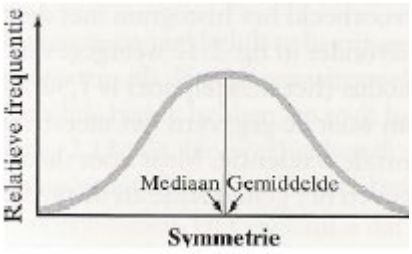
We houden de score bij op 3 verschillende OLOD's voor een steekproef van 50 studenten. Dit geeft volgende frequentietabellen met klassenindeling met bijhorende histogrammen.

Symmetrische verdelingen en dichtheidsfuncties

- Deze gedragen zich op dezelfde wijze aan de linkse en rechtste zijde van de figuur
- Mediaan en gemiddelde zijn (ongeveer) gelijk
- De linkertak en rechtertak vormen elkaars spiegelbeeld

Klassen	m_i	f_i
0<2	1	5
2<4	3	10
4<6	5	20
6<8	7	10
8<10	9	5
		50

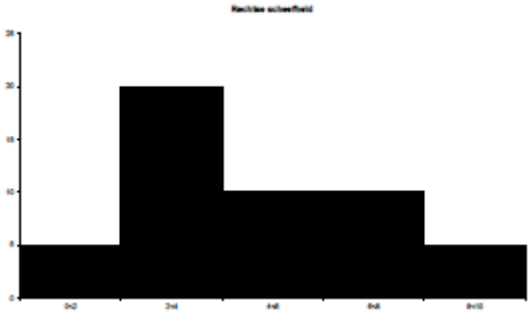


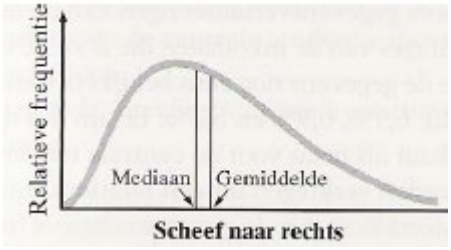


Rechts-scheve verdelingen en dichtheidsfuncties

- Het gemiddelde is groter dan de mediaan
- De dichtheidsfunctie en het histogram hebben een ‘rechterstaart’

Klassen	m_i	f_i
0<2	1	5
2<4	3	20
4<6	5	10
6<8	7	10
8<10	9	5
		50

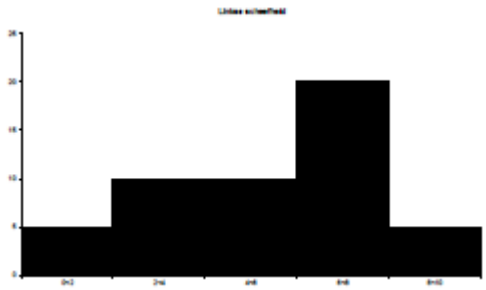


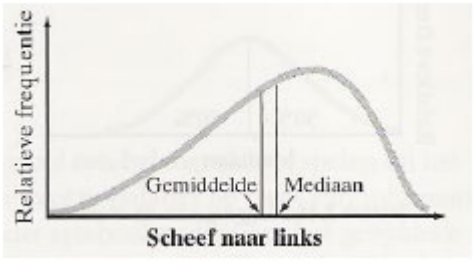


Links-scheve verdelingen en dichtheidsfuncties

- Het gemiddelde is kleiner dan de mediaan
- De dichtheidsfunctie en het histogram hebben een ‘linkerstaart’

Klassen	m_i	f_i
0<2	1	5
2<4	3	10
4<6	5	10
6<8	7	20
8<10	9	5
		50





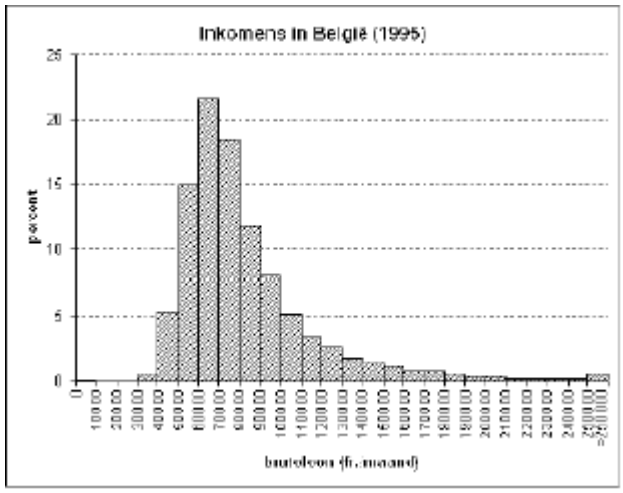
Besluit

- We hebben te maken met een symmetrische verdeling met bijhorende symmetrische dichtheidsfunctie als $\overline{x} \approx$ mediaan $\approx$ modus. Het bijhorende histogram en de dichtheidsfunctie zijn zo goed als symmetrisch.
- We hebben te maken met een links-scheve verdeling met bijhorende symmetrische dichtheidsfunctie als $\overline{x} \leq$ mediaan $\leq$ modus. Het bijhorende histogram en de dichtheidsfunctie vertonen een linkerstaart.
- We hebben te maken met een rechts-scheve verdeling met bijhorende symmetrische dichtheidsfunctie als $\overline{x} \geq$ mediaan $\geq$ modus. Het bijhorende histogram en de dichtheidsfunctie vertonen een rechterstaart.

Om te beoordelen of een observatie symmetrisch, linksscheef of rechtsscheef is, kan je best steeds het bijhorende histogram maken.

Voorbeeld

Histogram van de Belgische gezinsinkomens in het jaar 1995 (n = 474)

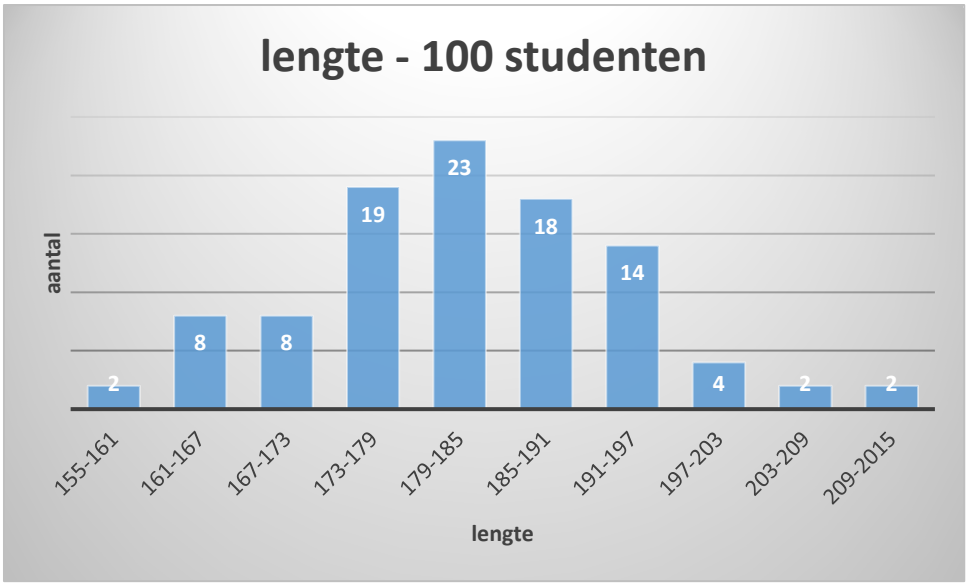


- Is bovenstaande een symmetrische verdeling?
- Welke centrummaat beklemtoont dat een groot aantal gezinnen een laag inkomen hebben?
- Welke maat drukt het best uit hoeveel de doorsnee Belg verdient?
- Welke maat houdt rekening met elk inkomen (zowel van de armen als van de rijken)?

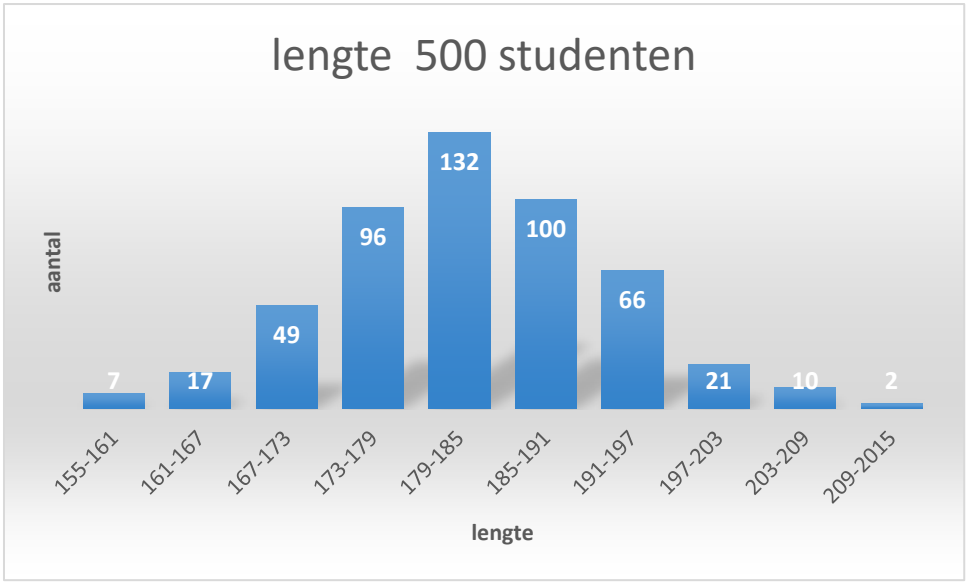
12. De normale verdeling

Een veel voorkomende verdeling / dichtheidsfunctie is deze van de normale verdeling omdat veel gegevens normaal verdeeld zijn. In de praktijk wordt de normale verdeling veel toegepast als het gaat om variabelen als lengte, gewicht, IQ - scores, … Wij beperken ons in deze cursus tot de normale verdeling.

Voorbeeld: staafdiagram lichaamslengte 100 studenten

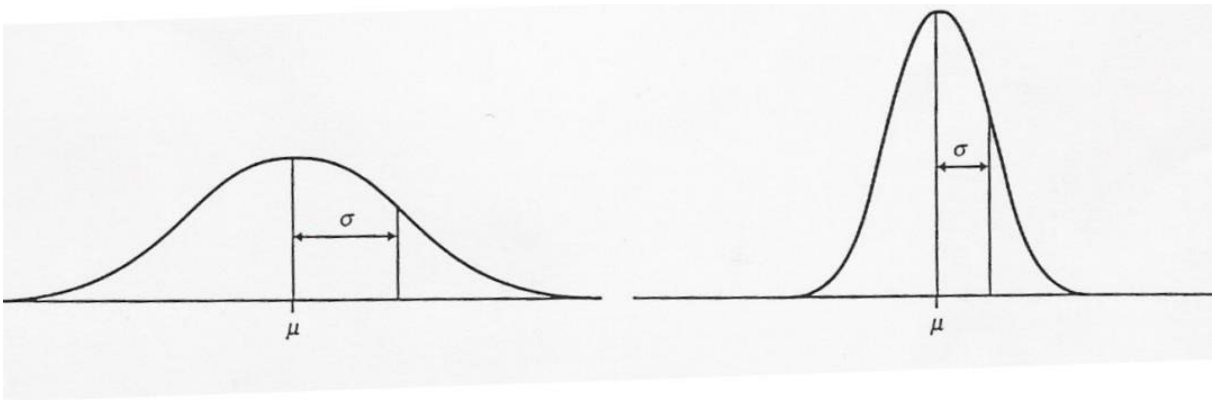


Voorbeeld: staafdiagram lichaamslengte 500 studenten

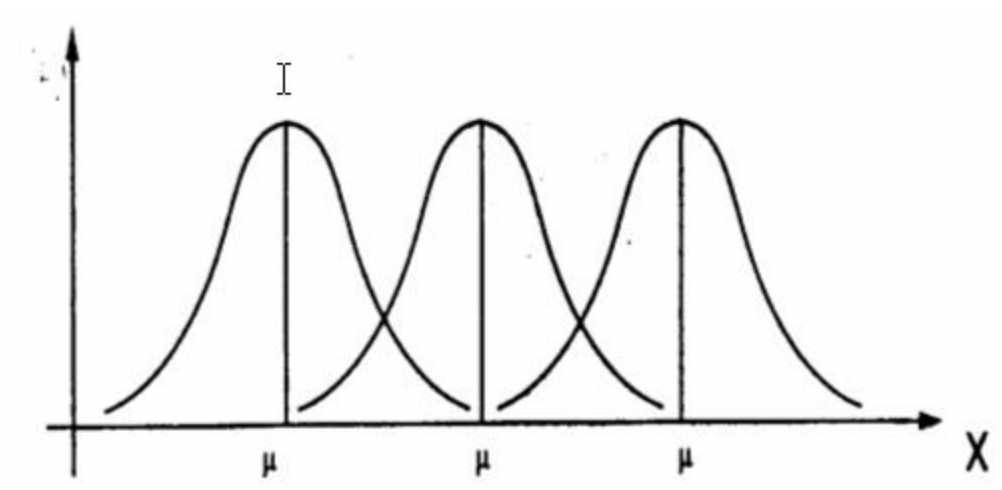


Grafiek

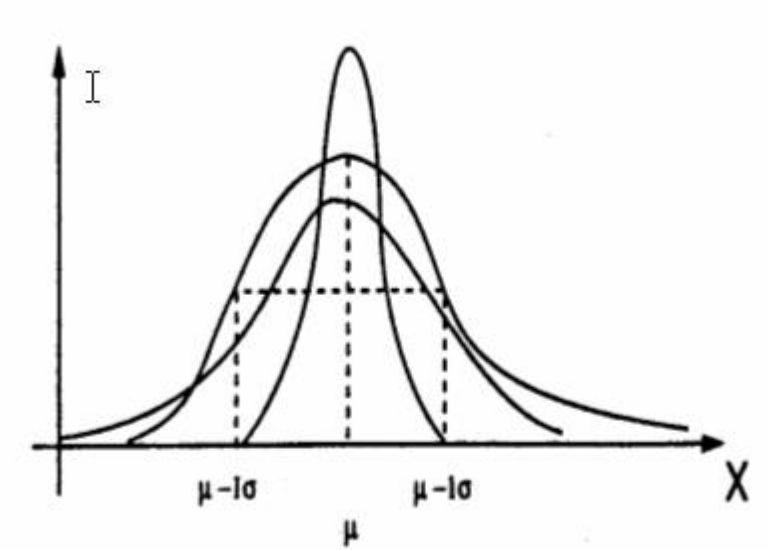
De dichtheidsfunctie van een normaal verdeelde variabele X ~ N (μ, σ^2) heeft een typische klokvorm (en wordt Gausscurve genoemd).
Ze is symmetrisch rond het gemiddelde μ. Bovendien varieert de breedte en de hoogte van de klok afhankelijk van de waarde van de standaard afwijking σ.



Indien we μ zouden wijzigen zonder σ te veranderen, dan verkrijgen we een horizontale verschuiving:



De standaardafwijking σ bepaalt de breedte van de curve. Zoveel te groter σ, zoveel te breder en vlakker de curve, wat een grotere spreiding van de gegevens aangeeft.



- De totale oppervlakte onder de curve is gelijk aan 1 of 100%

- Bij een normaal verdeelde variabele geldt het volgende:
  68 % van de gegevens bevindt zich in het interval $[\mu - \sigma, \mu + \sigma]$

  95 % van de gegevens bevindt zich in het interval $[\mu - 2\sigma, \mu + 2\sigma]$

  99.7% van de gegevens bevindt zich in het interval $[\mu - 3\sigma, \mu + 3\sigma]$

De oppervlaktes onder de dichtheidsfunctie van deze normale verdeling (kansen) kunnen m.b.v. Python commando's eenvoudig berekend worden.

In [ ]:
```
#Normale waarden (oppervlaktes onder dichtheidsfunctie van de normale verdeling) berekenen
```

In [1]:
```
import scipy.stats as stats
import math
# X ~ N (118, 36)
```

In [2]:
```
# P (X < 98)
print("Kans dat X kleiner is dan 98 is", stats.norm.cdf(98, 118, math.sqrt(36)))
```

```
Kans dat X kleiner is dan 98 is 0.0004290603331968372
```

In [4]:
```
# P(X > 120)
print("Kans dat X groter is dan 120 is", 1 - stats.norm.cdf(120, 118, math.sqrt(36)))
```

```
Kans dat X groter is dan 120 is 0.36944134018176367
```

In [5]:
```
# P (116 < X < 122)
print("Kans dat X tussen 116 en 122 ligt is", stats.norm.cdf(122, 118, math.sqrt(36)) - stats.norm.cdf(116, 118, math.sqrt(36)))
```

```
Kans dat X tussen 116 en 122 ligt is 0.3780661222713134
```

In [7]:
```
# Welke oppervlakte onder dichtheidsfunctie van X ligt er links van 118
print("De oppervlaktes links van 118 is", stats.norm.cdf(118, 118, math.sqrt(36)))
```

```
De oppervlaktes links van 118 is 0.5
```

In [8]:
```
# 80% van de oppervlakte ligt links van welke waarde?
print("80% van de oppervlakte ligt links van", stats.norm.ppf(0.8, 118, math.sqrt(36)))
```

```
80% van de oppervlakte ligt links van 123.04972740143748
```

In [9]:
```
# 80% van de oppervlakte ligt rechts van welke waarde?
print("80% van de oppervlakte ligt rechts van", stats.norm.ppf(0.2, 118, math.sqrt(36)) )
```

```
80% van de oppervlakte ligt rechts van 112.95027259856252
```

Voorbeeld 1: Z ~ N (0, 1): Bereken

$$P(Z < -1.23) =$$

$$P(Z > 2.09) =$$

$$P(1.21 < Z < 2.85) =$$

Voorbeeld 2: Z ~ N (0, 1): Bereken a als:

$$P(Z < a) = 0.9936$$

$$P(Z \leq a) = 0.0281$$

$$P(Z > a) = 0.9887$$

Voorbeeld 3:

Zij X de lichaamslengte van een PXL-student. De verdeling kan benaderd worden door een normale verdeling met gemiddelde 170.6cm en standaardafwijking 6.75cm.

a. Hoeveel % van de PXL-studenten hebben een lichaamslengte kleiner dan 180 cm?

b. Hoeveel % van de PXL-studenten hebben een lichaamslengte tussen 160 en 175 cm?

c. Hoeveel % van de PXL-studenten heeft een lichaamslengte gelijk aan 180cm?

d. 60% van de PXL-studenten heeft een lichaamslengte kleiner dan of gelijk aan …….cm.

Voorbeeld 4:

De punten van een leestest zijn $N(430; 100)$ verdeeld. Hoe hoog moet een leerling scoren om bij de 10% beste leerlingen te behoren?

Voorbeeld 5:

Een klas in 'data advanced' heeft normaal verdeelde punten met een gemiddelde van 52 op 100 en een standaardafwijking van 9.

Wat is de kans dat een individuele toevallig genomen student minder dan 50 heeft?

13. Oefeningen

Oefening 1

Veronderstel dat de IQ - scores van studenten normaal verdeeld zijn met gemiddelde 110 en standaardafwijking 25. Van hoeveel van de 200 studenten verwacht je dat ze een IQ - score van meer dan 140 hebben?

Oefening 2

De tijd die nodig is om een examen van een bepaalde cursus af te leggen is normaal verdeeld met een gemiddelde van 80 minuten en een variantie van 100.
a. Hoe groot is de kans dat een student het examen aflegt in meer dan 60 maar minder dan 75 minuten?
b. Neem aan dat in een klas 60 studenten zitten en dat ze maar 90 minuten de tijd krijgen om het examen te maken. Van hoeveel studenten verwacht je dat ze het examen niet afkrijgen?
c. Hoeveel tijd moet voorzien worden als 90% van de studenten voldoende tijd moet krijgen om het examen af te leggen?

Oefening 3

De levensduur van een machineonderdeel is normaal verdeeld met een gemiddelde van 5 jaar en een standaardafwijking van 1 jaar. De fabrikant vervangt het onderdeel gratis zolang het in garantie is. Hoeveel jaar garantie kan hij geven als hij niet meer dan 4% gratis wil vervangen?

Oefening 4

Een anesthesist beschikt over twee soorten medicaties om een patiënt te verdoven. De slaapduur bij de eerste medicatie is normaal verdeeld met een gemiddelde van 6 uur en een standaardafwijking van 1 uur. De slaapduur bij de tweede medicatie is normaal verdeeld met een gemiddelde van 5 uur en een standaardafwijking van 1.5 uur. Gezien de zwakte van de patiënt zou de anesthesist graag de tweede medicatie gebruiken, maar bovendien wil hij nog 99% zekerheid hebben dat de patiënt minstens 2.5 uur slaapt. Kan de anesthesist elk van deze medicaties gebruiken?

Oefening 5

De NV Maesen bestelt bouten bij de NV Publico. Als optimale lengte stelt Maesen een lengte van 60mm voor. Nochtans staat zij ook toe dat er bouten worden geleverd die 7mm afwijken van deze optimale lengte. Publico levert een grote vracht bouten die een gemiddelde lengte blijken te hebben van 58mm en een standaardafwijking van 4mm. We mogen ervan uitgaan dat de lengte van de bouten normaal verdeeld is. Hoeveel procent ‘slechte’ bouten werd er geleverd?

Oefening 6

Een bandenfabrikant heeft een nieuwe band ontwikkeld. De managers denken dat een garantie op duurzaamheid belangrijk kan zijn voor de acceptatie van de band door de consumenten. Daarom willen ze inzicht krijgen in de kansverdeling van het aantal kilometers dat de band meegaat. Op grond van testen schat het technisch bureau van de firma dat de band gemiddeld 36.500 kilometer meegaat met een standaardafwijking van 5000 kilometer. Verder valt uit de gegevens op te maken dat het redelijk is ervan uit te gaan dat het aantal kilometers dat de band meegaat normaal verdeeld is.

a. Hoe groot is de kans dat een band meer dan 40.000 kilometer meegaat?

De bandenfabrikant overweegt een aantal kilometers te garanderen, en een korting te geven op de vervanging van een band als deze niet langer meegaat.

b. Hoeveel kilometers kan de firma garanderen als ze niet wil dat meer dan 10% van alle banden in aanmerking komt voor een korting?