Data Advanced

DATA REPRESENTATIE

DE HOGESCHOOL
MET HET NETWERK

Hogeschool PXL – Elfde-Liniestraat 24 – B-3500 Hasselt
www.pxl.be - www.pxl.be/facebook

1. – 2. Doel + Inhoud

Gegevens voorstellen + conclusies trekken

- Gegevens verzamelen
- Frequentietabellen
- Grafische representaties
- Kengetallen
- Verbanden tussen variabelen
- Populatie – steekproef
- Dichtheidsfunctie

3. Gegevens verzamelen

Wat zijn gegevens?

- Categorische gegevens
  - Nominaal
  - Ordinaal

- Numerieke gegevens
  - Discrete gegevens
  - Continue gegevens

3. Gegevens verzamelen

Waar halen we deze gegevens?

- Reeds verzameld

- Zelf verzamelen
  - Waarnemend onderzoek
  - Experimenteel onderzoek

3. Gegevens verzamelen

Hoe betrouwbaar zijn deze gegevens?

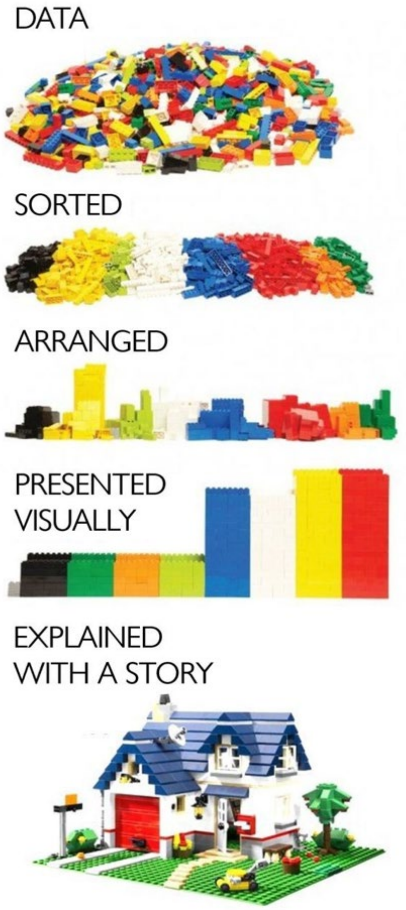
- Slechte steekproef

- Foutieve gegevens

- Foutieve conclusies

4. Gegevens voorstellen

- Frequentietabel
  - Categorische gegevens
  - Numerieke discrete gegevens
  - Numerieke continue gegevens
- Grafische presentatie



4.1. Frequentietabel categorische gegevens

Vooropleiding Algemeen	f_i
ASO	1
TSO	15
BSO	4
	20

f_i = absolute frequentie
φ_i = relatieve frequentie

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

4.2. Frequentietabel numerieke discrete gegevens

Examenresultaten op 20:

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

4.2. Frequentietabel numerieke discrete gegevens

x_i	f_i	φ_i
2		
5		
7		
9		
10		
11		
14		
15		
16		
17		

f_i = absolute frequentie
φ_i = relatieve frequentie

4.2. Frequentietabel numerieke discrete gegevens

x_i	f_i	φ_i
2	2	
5	5	
7	4	
9	1	
10	1	
11	1	
14	1	
15	2	
16	2	
17	1	

f_i = absolute frequentie
φ_i = relatieve frequentie

4.2. Frequentietabel numerieke discrete gegevens

x_i	f_i	φ_i
2	2	0,1
5	5	0,25
7	4	0,2
9	1	0,05
10	1	0,05
11	1	0,05
14	1	0,05
15	2	0,1
16	2	0,1
17	1	0,05

f_i = absolute frequentie
φ_i = relatieve frequentie

4.2. Frequentietabel numerieke discrete gegevens

Score OLOD1	f_i	φ_i	cf_i	$c\varphi_i$
2	2	0.1	2	0.1
5	5	0.25	7	0.35
7	4	0.2	11	0.55
9	1	0.05	12	0.6
10	1	0.05	13	0.65
11	1	0.05	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		

f_i = absolute frequentie
φ_i = relatieve frequentie
cf_i = cumulatieve absolute frequentie
$c\varphi_i$ = cumulatieve relatieve frequentie

4.3. Frequentietabel numerieke continue gegevens

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**		

4.3. Frequentietabel numerieke continue gegevens

<u>Werkwijze:</u>

- Zoek het grootste en kleinste waarnemingsgetal (min=1.89, max=16.86).

- Bereken het verschil tussen de extreme waarden (16.86 – 1.89 = 14.97).

- Deel dit verschil door 5 en door 15 en kies een klassenbreedte b tussen deze uitkomsten ($0.998 \leq$ klassenbreedte ≤ 2.994 ; kies b = 2).

4.3. Frequentietabel numerieke continue gegevens

Score OLOD1 (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**		

4.3. Frequentietabel numerieke continue gegevens

<u>Definities:</u>

- **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

- Cirkeldiagram
- Staafdiagram
- Histogram
- Boxplot (zie ook pg 85)
- Spreidingsdiagram of scatterplot

4.4. Grafische voorstelling: cirkeldiagram

```
#Cirkeldiagram via "pandas"
#Selecteer kolom "vooropleiding"
vooropleiding = slaagcijfers_kort['vooropleiding']
```

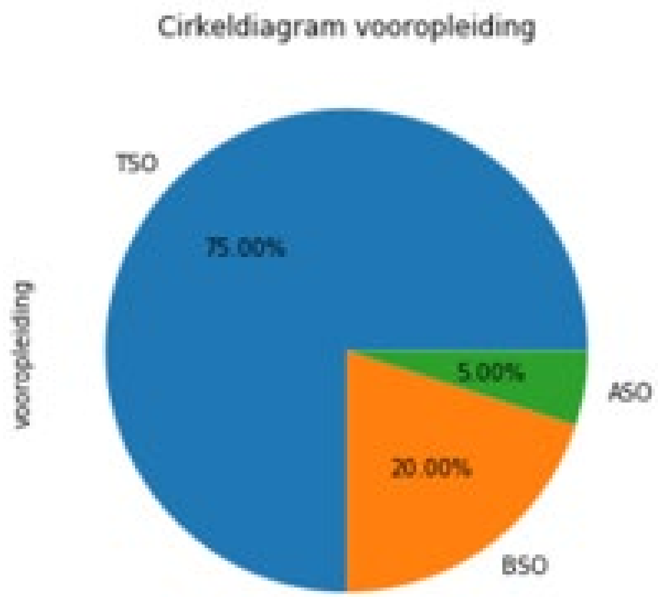
```
#via methode value_counts() tellen we de absolute aantallen
data_pie_plot = vooropleiding.value_counts()
print(data_pie_plot)
```

```
TSO    15
BSO     4
ASO     1
```

4.4. Grafische voorstelling: cirkeldiagram

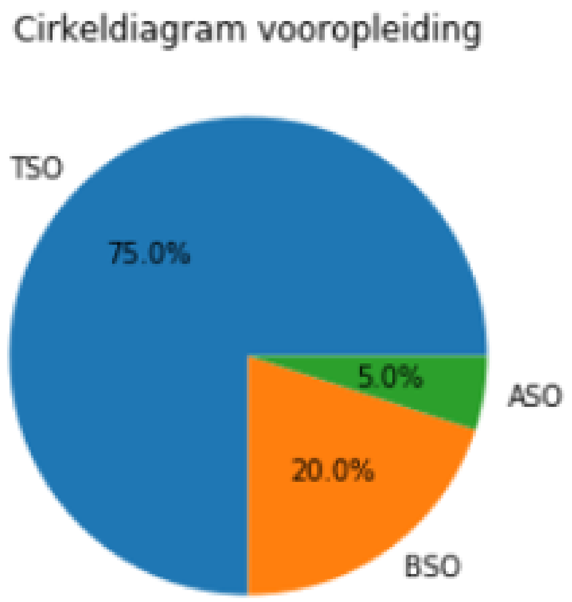
```
#via plot.pie maken we een cirkeldiagram van het categorisch gegeven "vooropleiding"
pie_plot = data_pie_plot.plot.pie(figsize=(5,5), autopct='%3.2f%%', title = 'Cirkeldiagram vooropleiding')
print(pie_plot)
```

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



4.4. Grafische voorstelling: cirkeldiagram

```
#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()
```



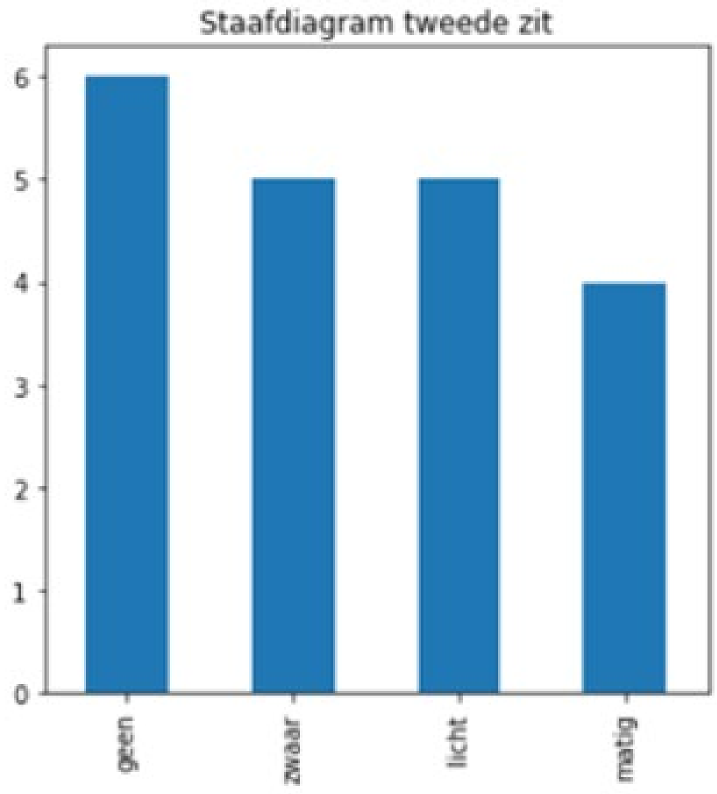
4.4. Grafische voorstelling: staafdiagram

```python
# 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
```

4.4. Grafische voorstelling: staafdiagram

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

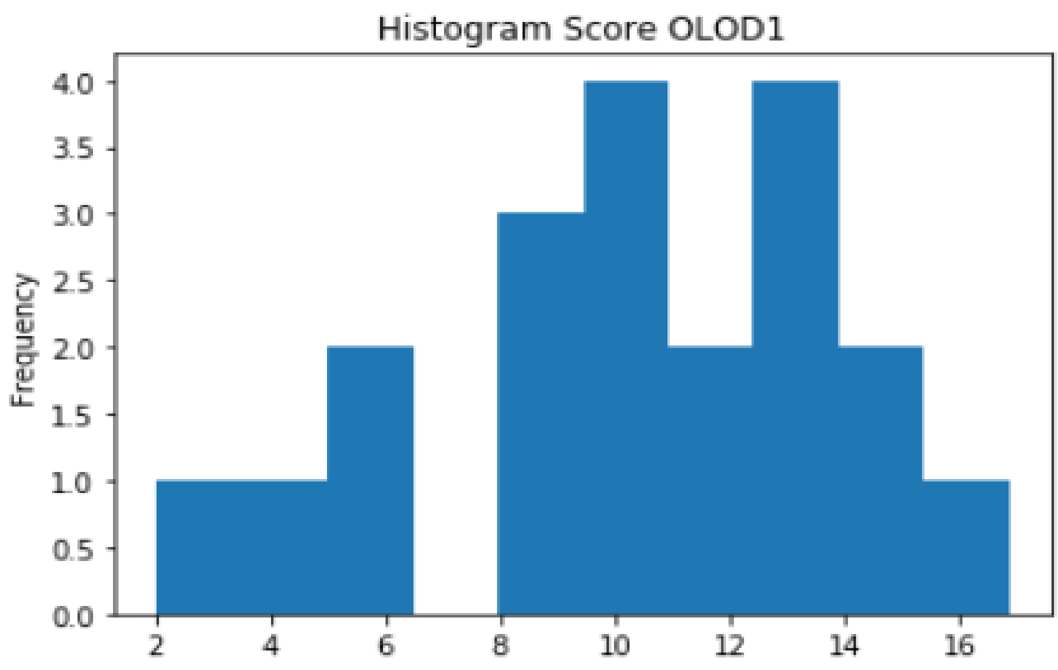


4.4. Grafische voorstelling: histogram

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

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

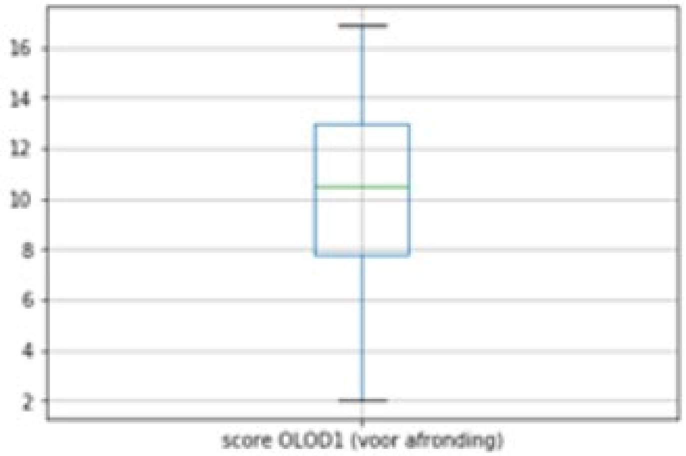
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f3a8a50128>
```



4.4. Grafische voorstelling: boxplot

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

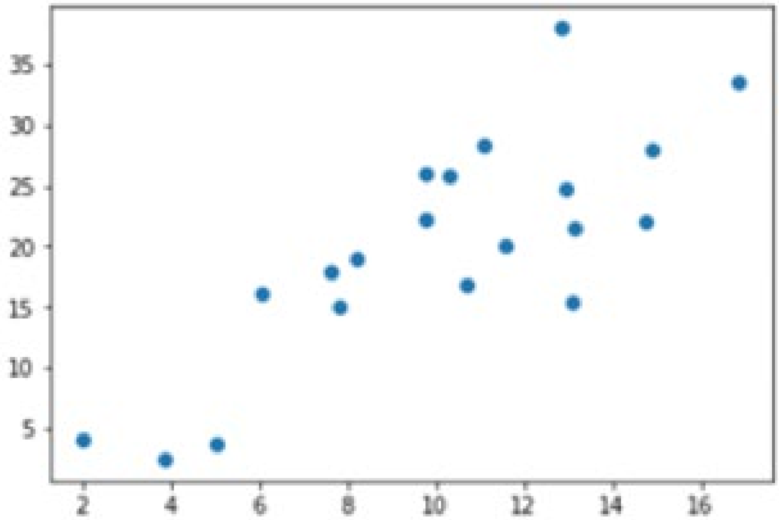
```
<matplotlib.axes._subplots.AxesSubplot at 0x1deb2050208>
```



4.4. Grafische voorstelling: scatterplot

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

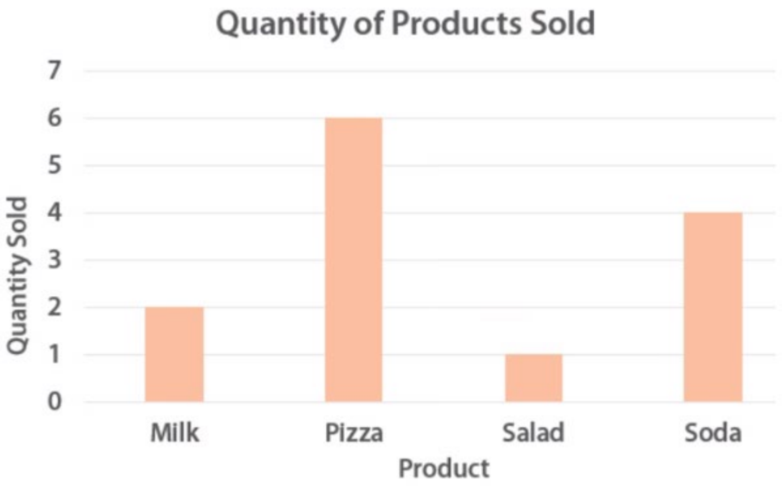
```
<matplotlib.collections.PathCollection at 0x1deb2192a58>
```



4.5. Verwarrende / foutieve representaties

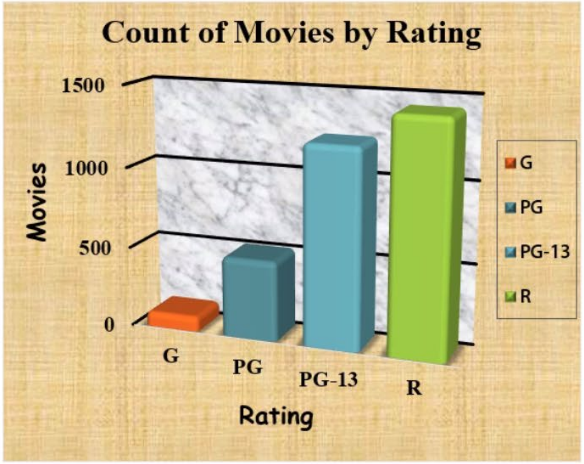
Zelfde data, andere representatie…

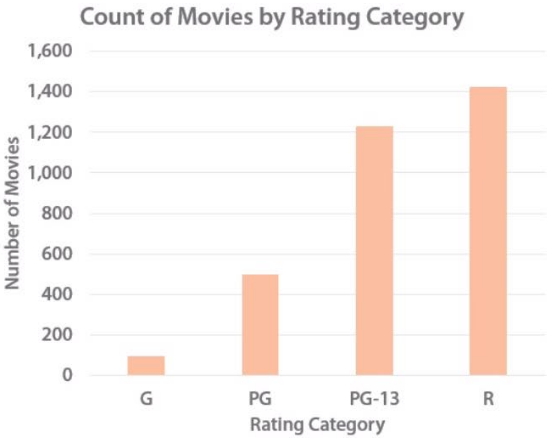
ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1



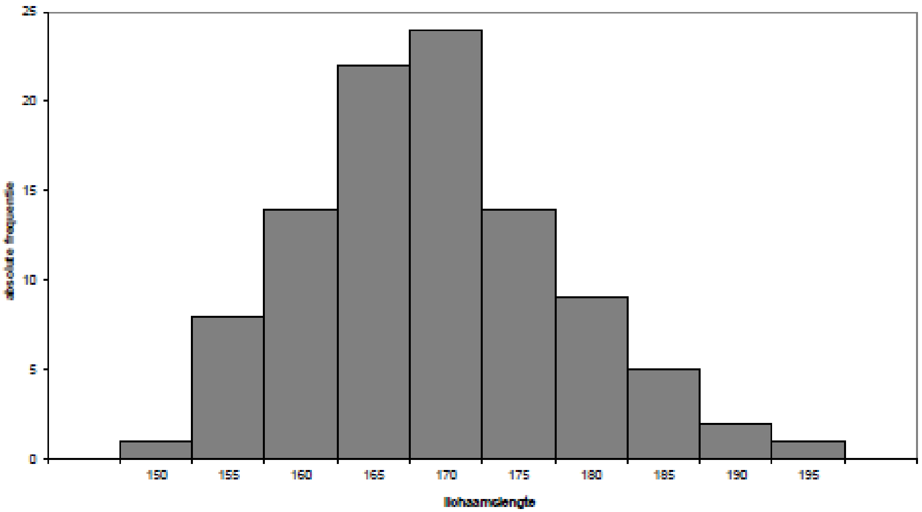
4.5. Verwarrende / foutieve representaties

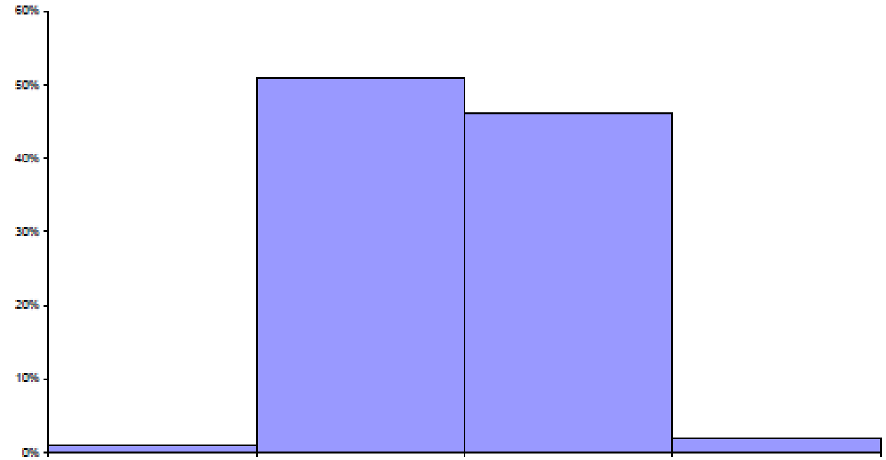
Zelfde data, andere representatie…

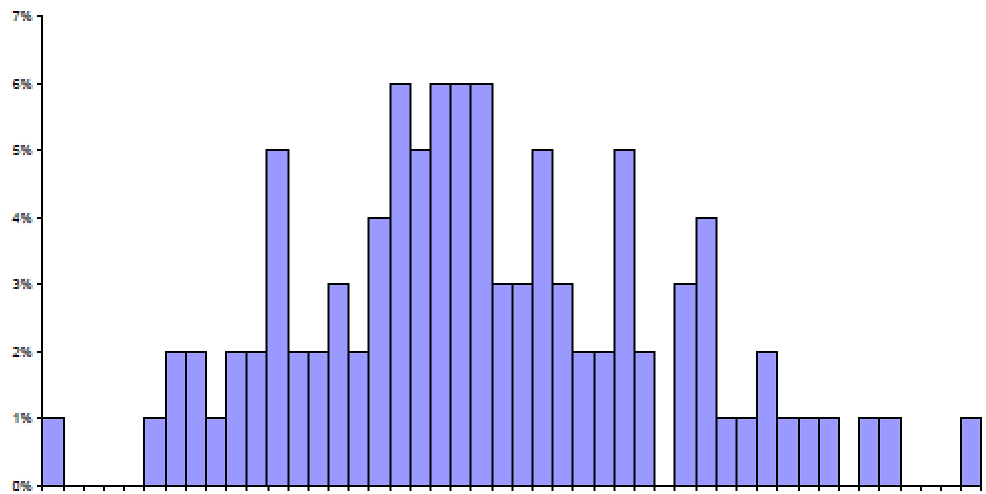




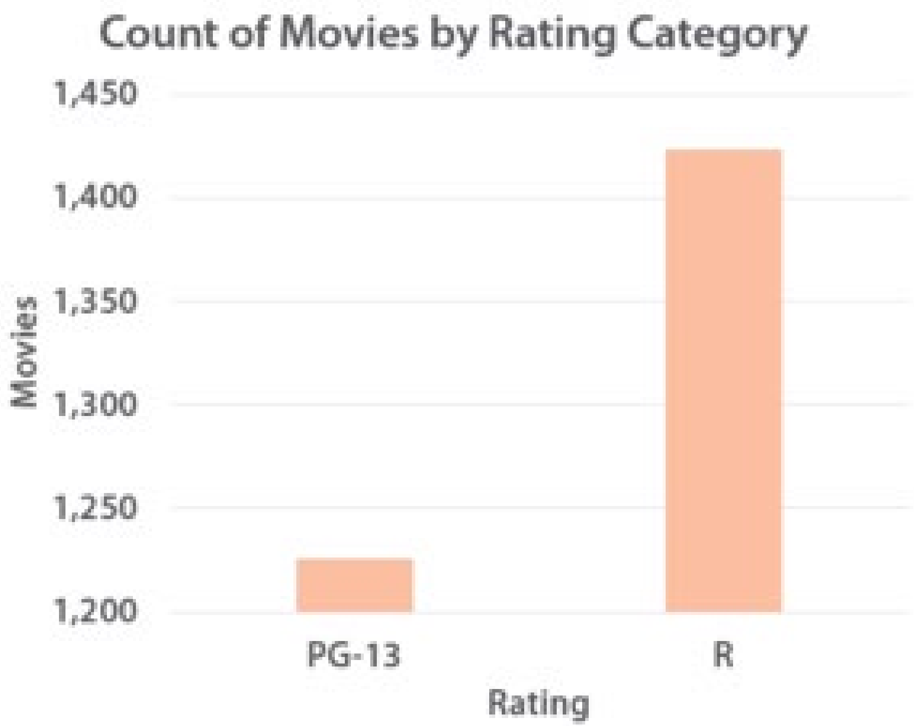
4.5. Verwarrende / foutieve representaties



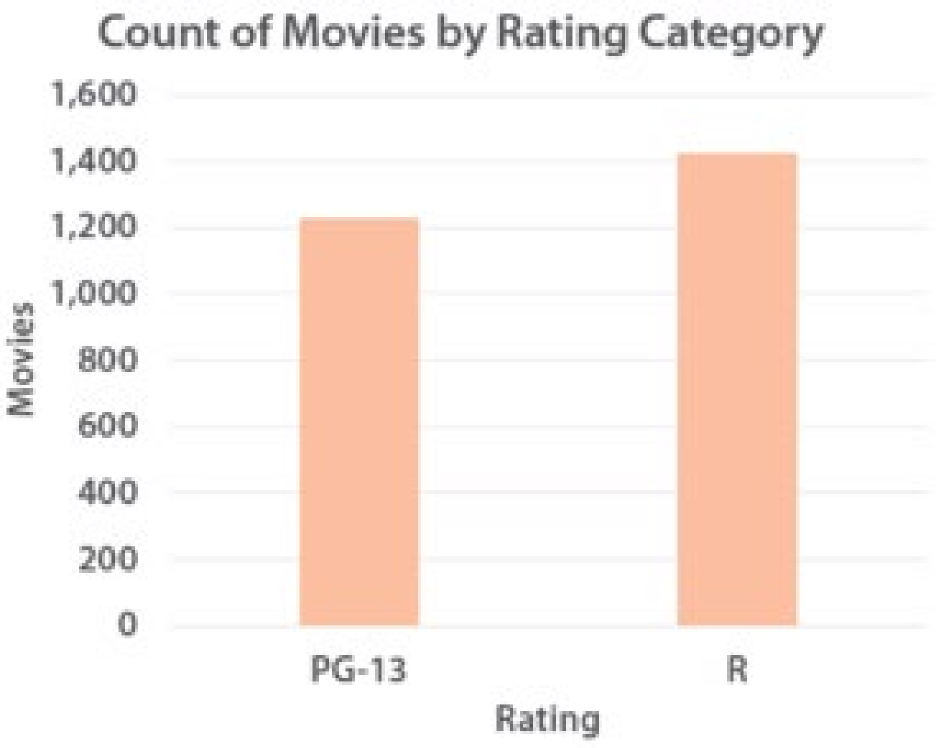




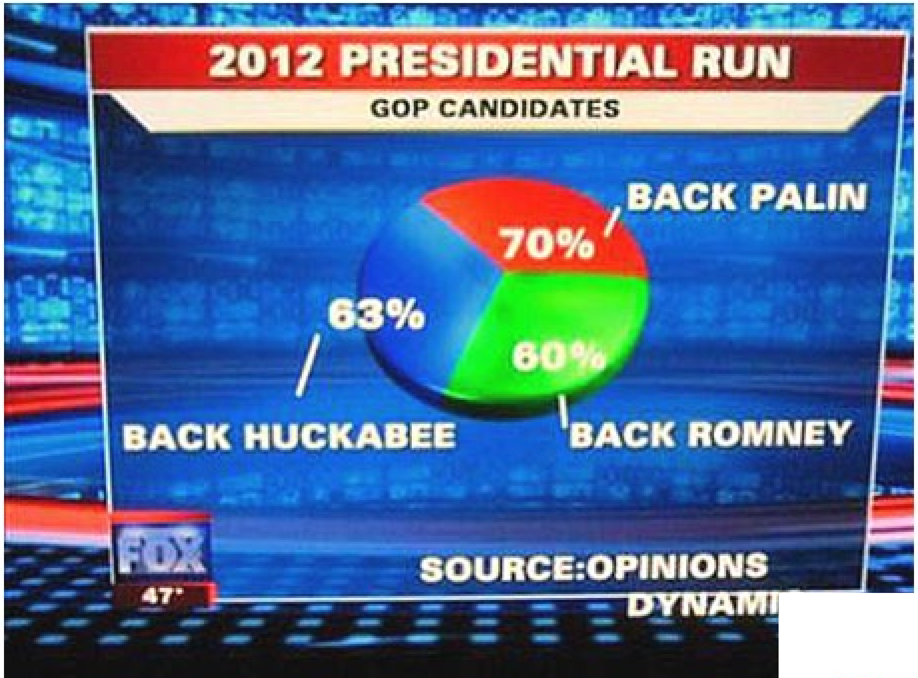
4.5. Verwarrende / foutieve representaties

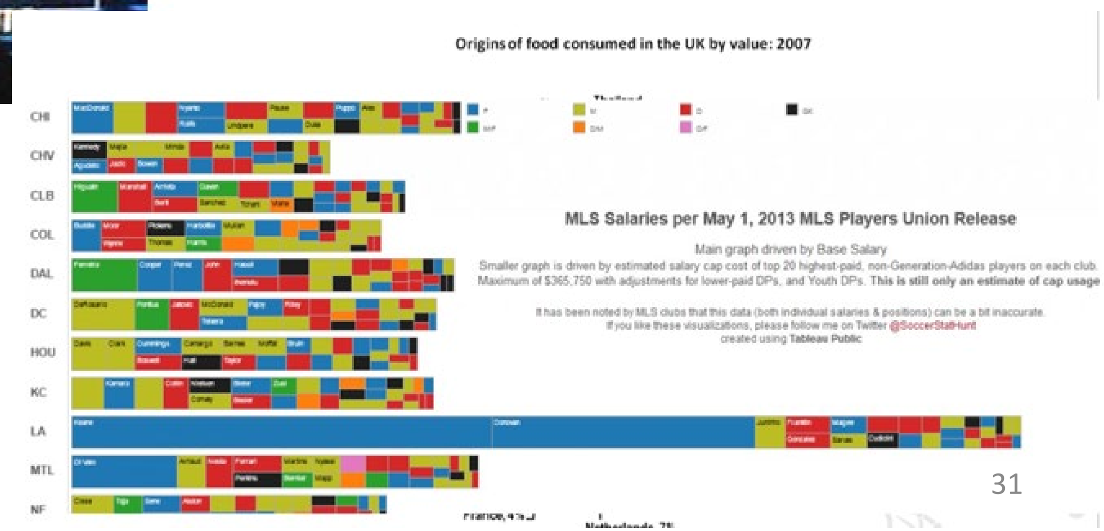


4.5. Verwarrende / foutieve representaties



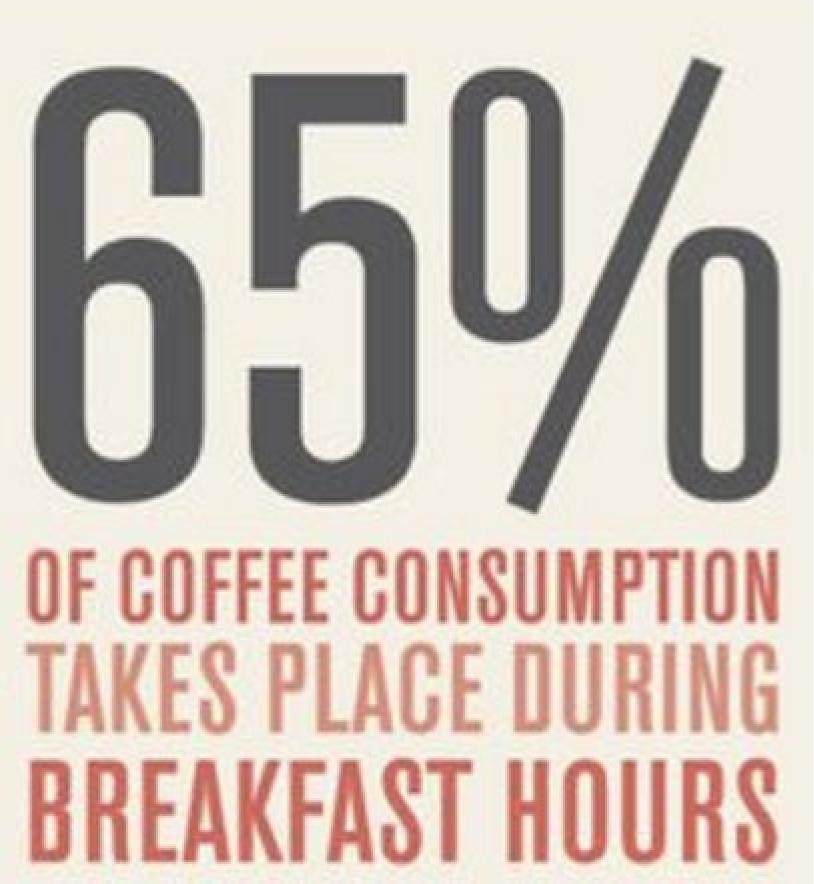
4.5. Verwarrende / foutieve representaties





4.5. Verwarrende / foutieve representaties

Soms is 1 cijfer voldoende…



5. Gegevens samenvatten

- Kengetallen voor locatie
  - Rekenkundig gemiddelde
  - Gewogen rekenkundig gemiddelde
  - Mediaan
  - Kwartielen
  - Modus
- Kengetallen voor spreiding
  - Variantie
  - Standaardafwijking
  - Spreidingsbreedte
  - Interkwartielafstand

5.1. Kengetallen voor locatie

5	2	7	6	10

105	102	107	106	110

- Gemiddelde: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\bar{x}_A = \frac{5+2+7+6+10}{5} = 6$$

$$\bar{x}_B = \frac{105+102+107+106+110}{5} = 106$$

5.1. Kengetallen voor locatie

- Gemiddelde voor frequentietabellen:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$$

x_i	f_i
1	1
3	2
5	7

$$\bar{x} = \frac{1*1+2*3+7*5}{10} = \frac{1+6+35}{10} = 4.2$$

5.1. Kengetallen voor locatie

- Gemiddelde voor frequentietabellen met klassenindeling:

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} m_i f_i$$

klasse	m_i	f_i
[1;3[	2	2
[3;5[	4	5
[5;7]	6	3

$$\bar{x} \approx \frac{2*2+4*5+6*3}{10} = \frac{4+20+18}{10} = 4.2$$

5.1. Kengetallen voor locatie

5	10	5	3	7

5	10	1000	3	7

Zeer gevoelig voor uitschieters:

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

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

Mogelijke oplossing: trimmed mean

5.1. Kengetallen voor locatie

Definitie: gewogen rekenkundig gemiddelde

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

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \cdots w_nx_n}{w_1 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

5.1. Kengetallen voor locatie

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

5.1. Kengetallen voor locatie

32	42	46	46	54

Mediaan = 46

2	3	4	7	8	10	10	15

Mediaan = (7 + 8)/2 = 7.5

5.2. Kengetallen voor locatie

Een **kwartiel** is één van de drie waarden die een geordende dataset in vier gelijke delen opdeelt. Men spreekt van het eerste, tweede en derde kwartiel en noteert deze als

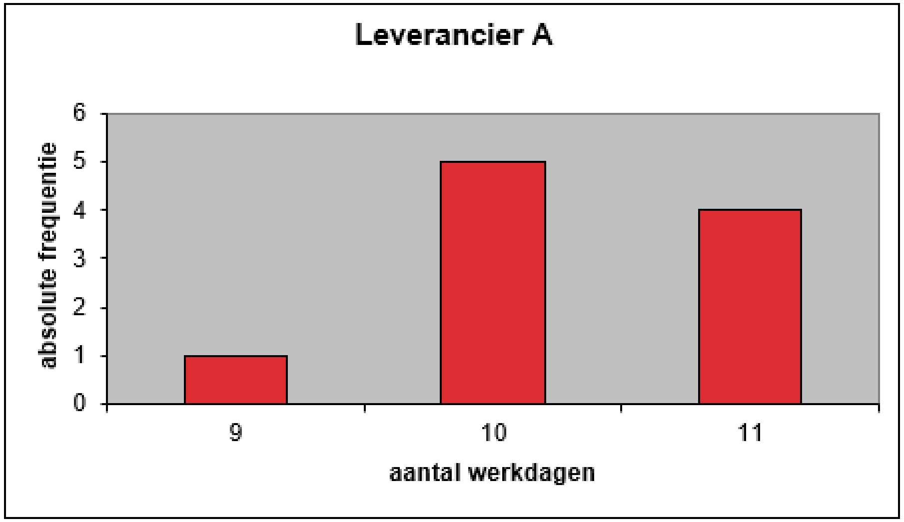
$Q_1 , Q_2 \; en \; Q_3$

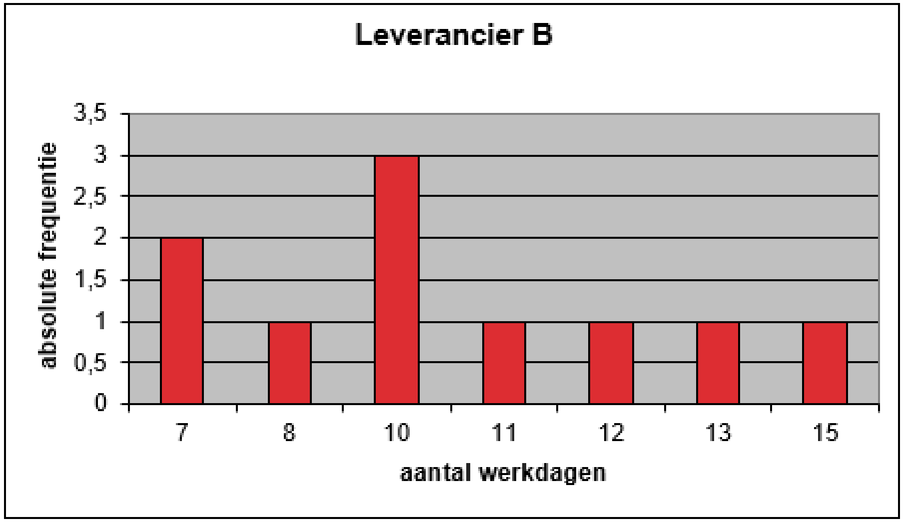
Modus is de observatie met de hoogste frequentie

5.2. Kengetallen voor spreiding

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

gemiddelde A = gemiddelde B = 10,3





5.2. Kengetallen voor spreiding

11	10	9	10	11	11	10	11	10	10

8	10	13	7	10	11	10	7	15	12

Variantie:

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

$s^2_a = 1/9\ [(11-10.3)^2 + (10-10.3)^2 + \ldots\ .\ (10-10.3)^2] = 0.45$

$s^2_b = 1/9\ [(8-10.3)^2 + (10-10.3)^2 + \ldots\ .\ (12-10.3)^2] = 6.67$

5.2. Kengetallen voor spreiding

Standaardafwijking:

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

5.2. Kengetallen voor spreiding

- **Spreidingsbreedte** = grootste waarde – kleinste waarde

32	42	46	46	54

Spreidingsbreedte = 54 – 32 = 22

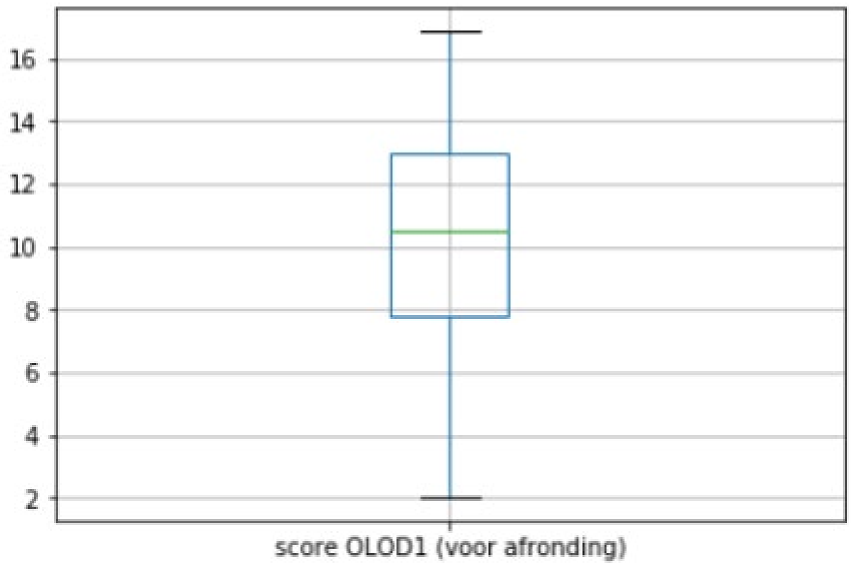
5.2. Kengetallen voor spreiding

- Interkwartielafstand (IKA) = $Q_3 - Q_1$

32	42	46	46	54

Interkwartielafstand = 50 – 37 = 13

5.3. Visuele voorstelling van locatie en spreiding



6. Verbanden tussen variabelen

- Inhoud
- Kracht
- Richting
- Causaliteit

6.1. Twee categorische variabelen

Kruistabel: absolute frequenties

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

➔ geen verband; variabelen = onafhankelijk

6.1. Twee categorische variabelen

Kruistabel: relatieve frequenties

hogeschool	A	B	C	All
prestatie				
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

➔ geen verband; variabelen = onafhankelijk

6.1. Twee categorische variabelen

onafhankelijkheid als:

- 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

6.1. Twee categorische variabelen

Kruistabel: absolute frequenties

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

➔ verband / correlatie tussen variabelen

6.1. Twee categorische variabelen

Kruistabel: absolute frequenties

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

→ Perfecte samenhang tussen variabelen

6.2. Twee numerieke variabelen

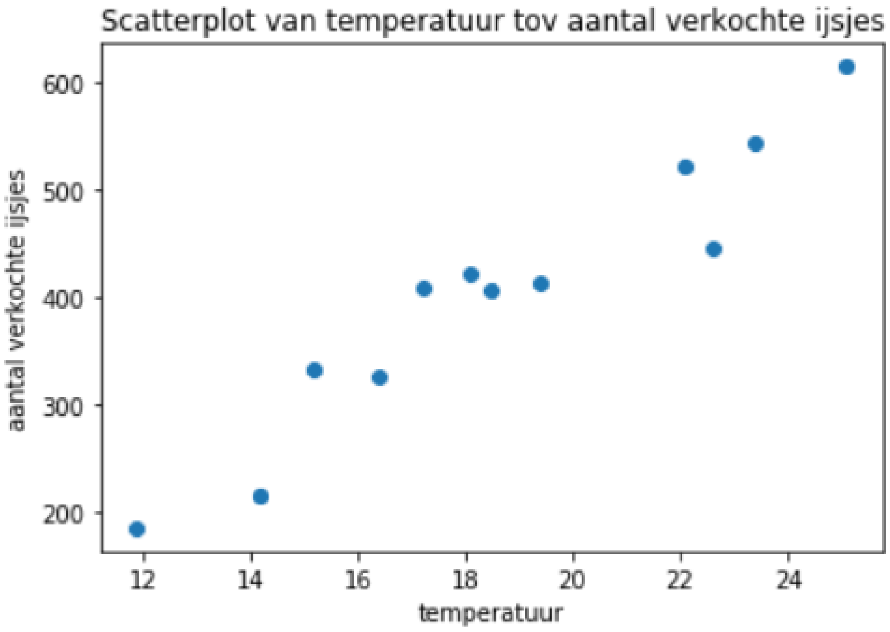
Spreidingsdiagram (scatterplot)

- Toename in één variabele ➔ toename in andere variabele: positief correleren

- Toename in één variabele ➔ afname in andere variabele: negatief correleren

6.2. Twee numerieke variabelen

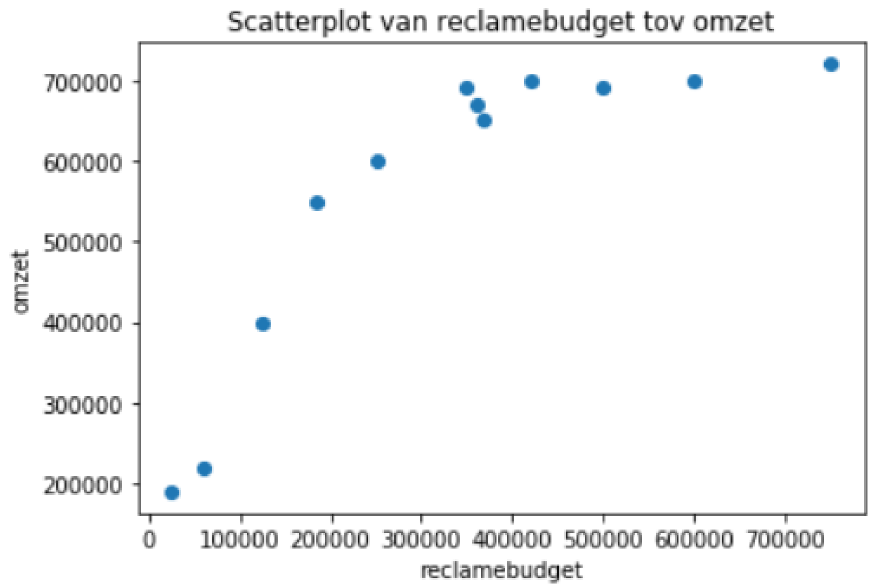
Spreidingsdiagram (scatterplot)



➔ Lineair verband

6.2. Twee numerieke variabelen

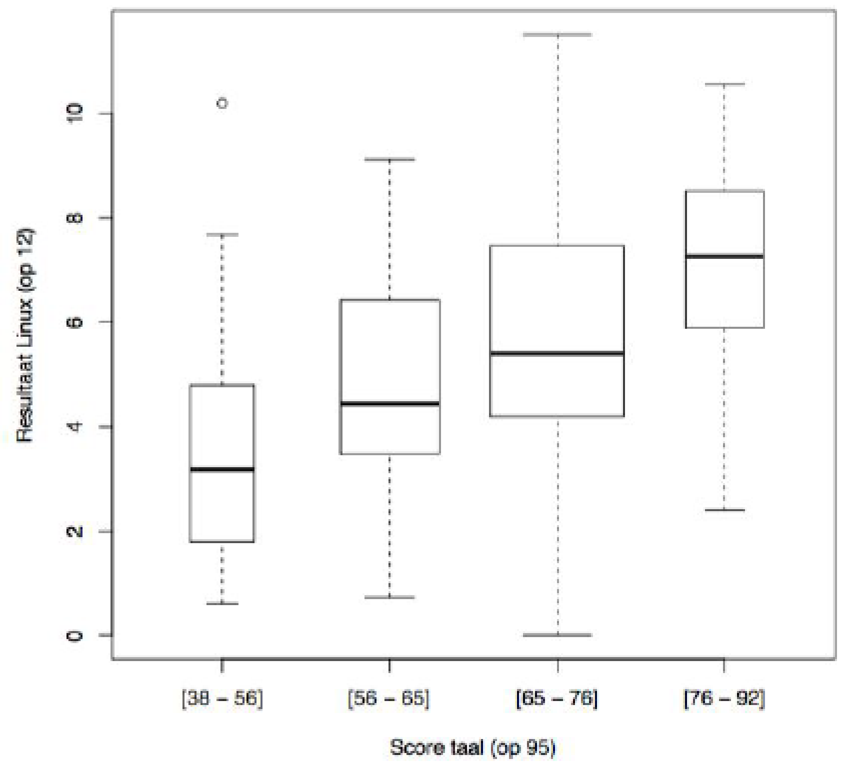
Spreidingsdiagram (scatterplot)



➔ Verband (niet – lineair)

6.3. Eén categorische en één numerieke variabele

Boxplot



7. Missing Values

- Reden
  - Respondenten antwoorden niet
  - Antwoorden niet opgenomen in de dataset

- Soorten
  - MCAR
  - MAR

8. Outliers

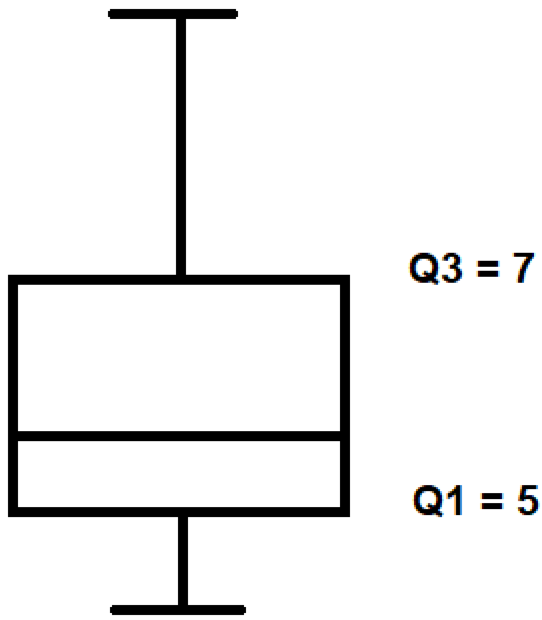
- Definitie:

Waarneming die ongewoon klein / groot is tov andere waarnemingen

- Oorzaak:
  - Onjuiste waarde
  - Juiste waarde: zeldzaam

- Detecteren
  - boxplot
  - z - scores

8. Outliers



IKA = 2

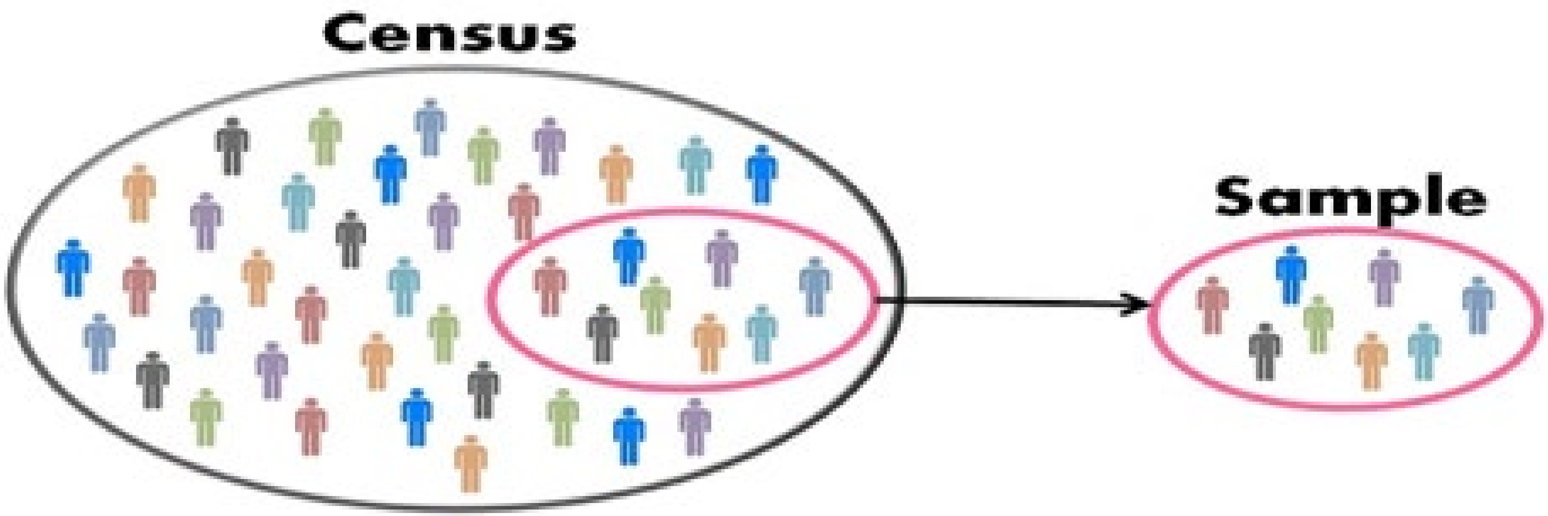
Omgaan met missing values / outliers

- Actie:
  - Negeren
  - Feature (kolom) verwijderen
  - Observatie (rij) verwijderen
  - Waarden imputeren
  - Data transformeren (vb binning)

Imputation <-> Binning

	Score / 20	Imputation (gem)	Transformation (binning)
Student 1	0	0	[0 , 3 [
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	[12 , 15 [

9. Populatie – Steekproef



Selecte steekproef

Aselecte steekproef

9. Populatie – Steekproef

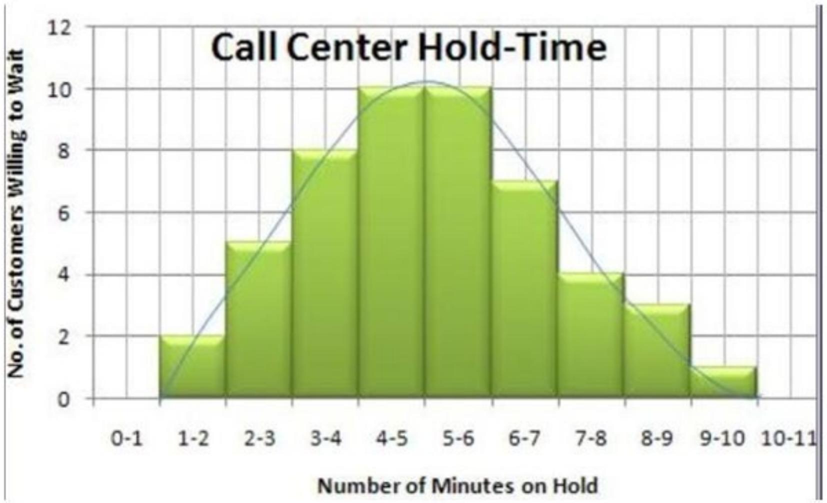
	Steekproef	Populatie
Grootte	n	N
Gemiddelde	$\bar{X}$ $(\bar{x})$	μ
Variantie	S^2 (s^2)	σ^2
Standaard afwijking	S (s)	σ

10. De dichtheidsfunctie bij numerieke gegevens

= “ideale” beschrijving van de volledige dataset

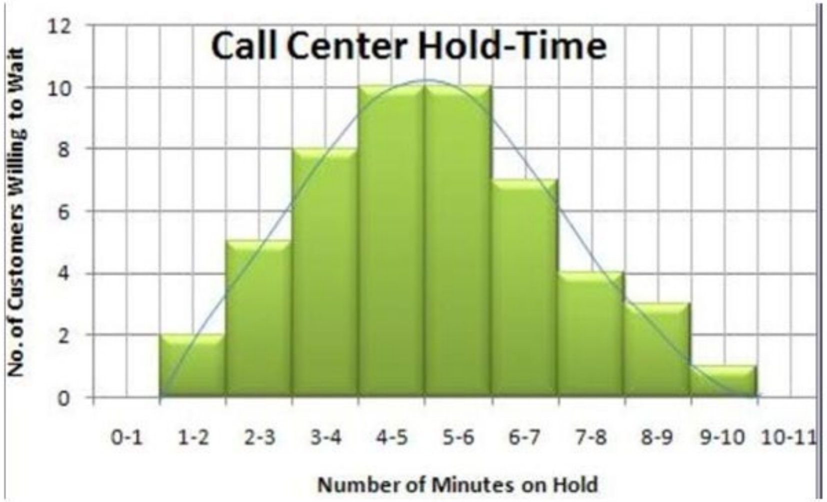
- % mensen dat ten hoogste 6 min wacht?

- % mensen dat tussen 4 en 6 min wacht?



10. De dichtheidsfunctie bij numerieke gegevens

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))
```

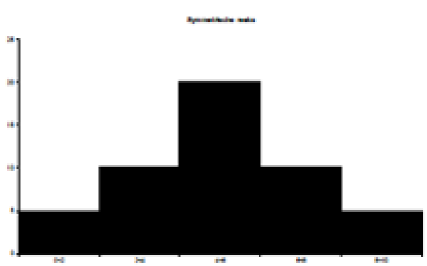
```
0.38292492254802624
```

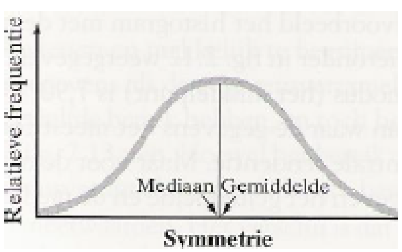
11. Soorten Dichtheidsfuncties en verdelingen

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



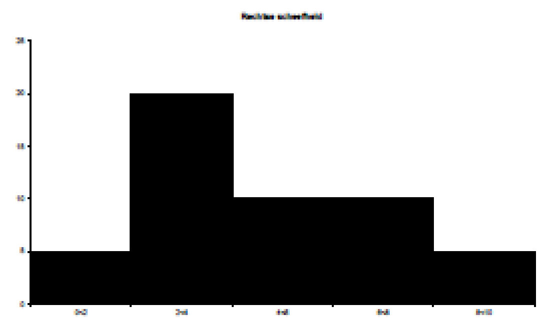


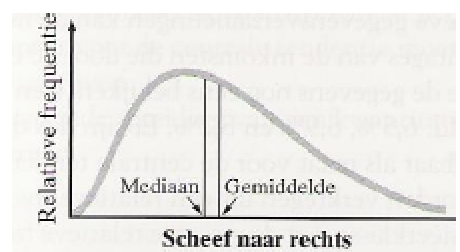
11. Soorten Dichtheidsfuncties en verdelingen

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



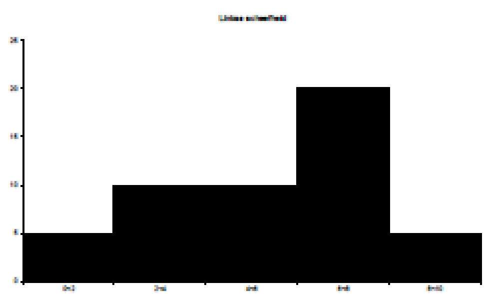


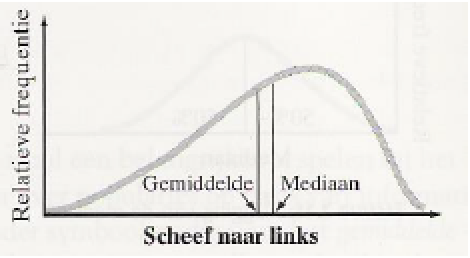
11. Soorten Dichtheidsfuncties en verdelingen

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

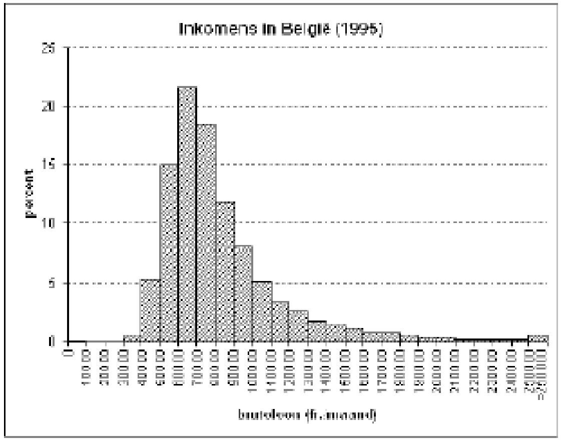




11. Soorten Dichtheidsfuncties en verdelingen

Voorbeeld

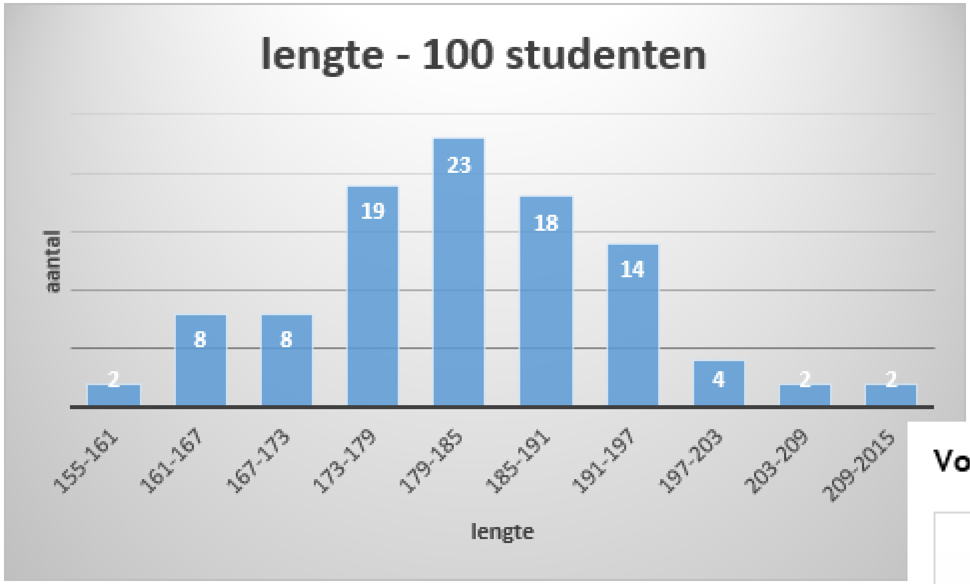
Histogram van de Belgische gezinsinkomens in het jaar 1995 (n = 474)



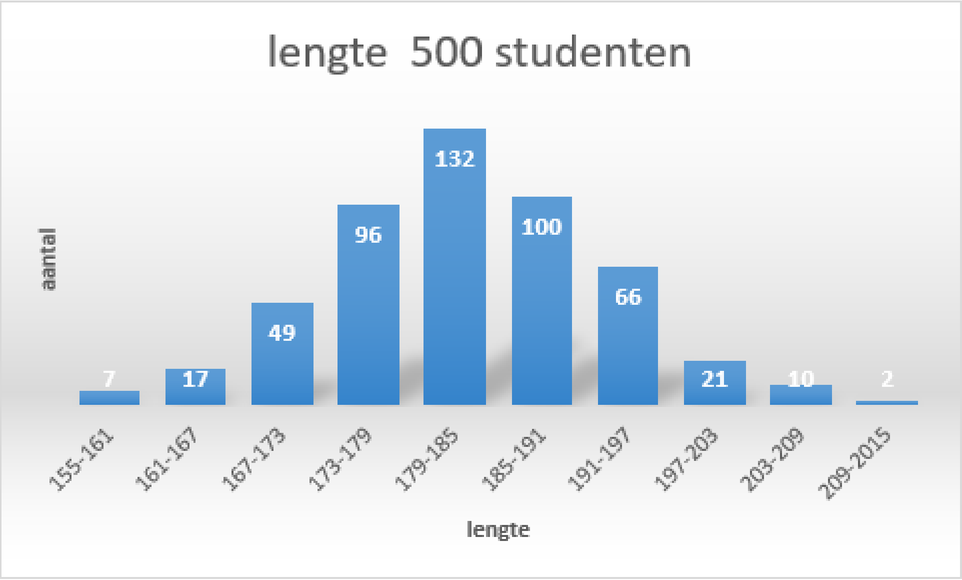
- Is bovenstaande een symmetrische verdeling?
- Welke centrummaat beklemtoont dat een groot aantal gezinnen een laag inkomen hebben?
- Welke maat drukt het best uit hoeveel de doorsnee Belg verdient?
- Welke maat houdt rekening met elke inkomen (zowel van de armen als van de rijken)?

12. De normale verdeling

Voorbeeld: staafdiagram lichaamslengte 100 studenten



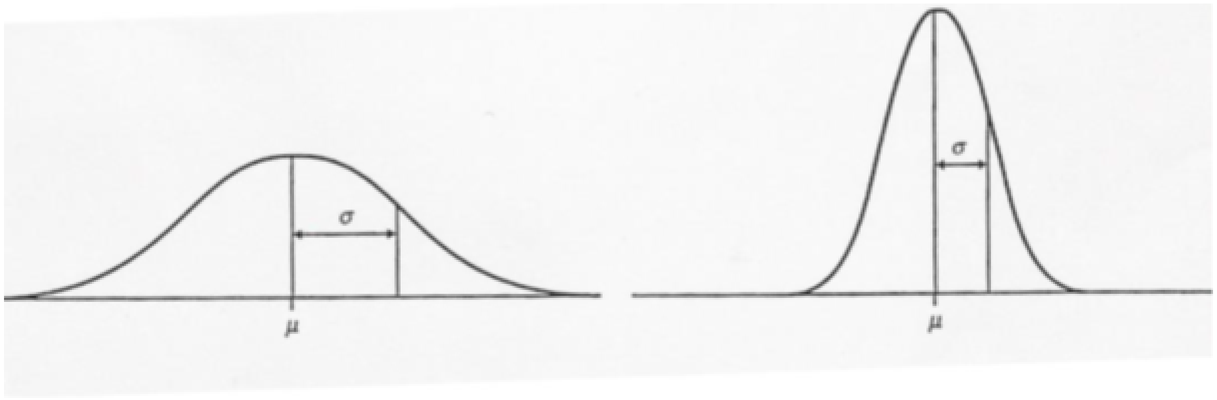
Voorbeeld: staafdiagram lichaamslengte 500 studenten



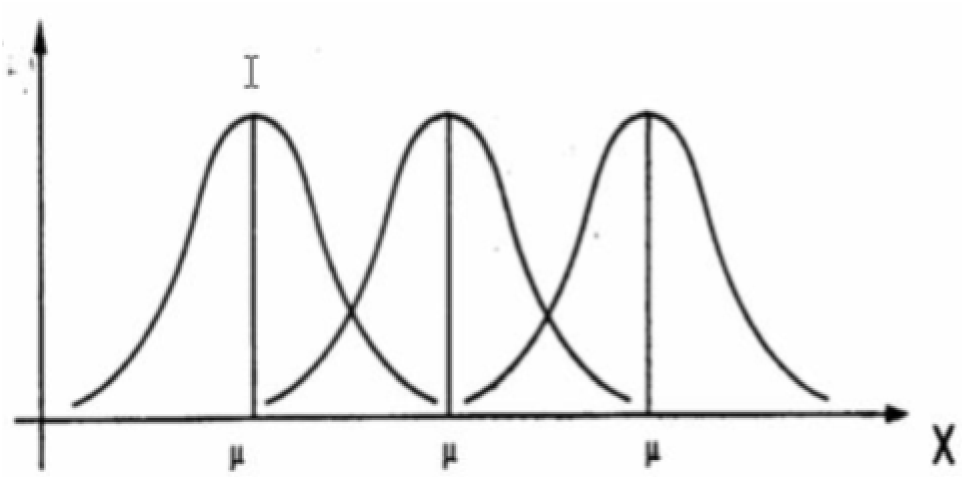
12. De normale verdeling

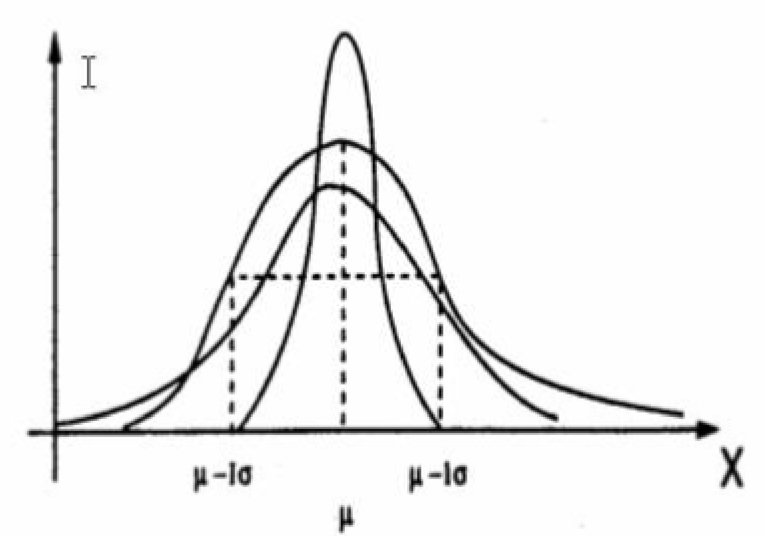
Grafiek

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.



12. De normale verdeling





Normale verdeling: kansen berekenen mbv python

```
#Normale waarden (oppervlaktes onder dichtheidsfunctie van de normale verdeling) berekenen
```

```
import scipy.stats as stats
import math
# X ~ N (118, 36)
```

```
# 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
```

```
# 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
```

```
# 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
```

```
# 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
```

```
# 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
```

```
# 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
```

Voorbeelden (pg 113)

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:

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

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

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

Voorbeelden + oefeningen

Voorbeeld 1 – 2 – 3 – 4 – 5

Oefeningen 1 – 2 – 3 – 6