Data Advanced

Hoofdstuk 3

Data Representatie

DE HOGESCHOOL
MET HET NETWERK

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

Gegevens verzamelen

Wat zijn gegevens?

- Kwalitatieve gegevens
  - Nominaal
  - Ordinaal

- Kwantitatieve gegevens
  - Discrete gegevens
  - Continue gegevens

Gegevens verzamelen

Waar halen we deze gegevens?

- Reeds verzameld

- Zelf verzamelen
  - Waarnemend onderzoek
  - Experimenteel onderzoek

Gegevens verzamelen

Hoe betrouwbaar zijn deze gegevens?

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

Lesvrije week of krokusvakantie?

Observatie 78 in dataset: slaagcijfers

Frequentietabel voor kwalitatieve 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**

Frequentietabel voor kwantitatieve discrete gegevens

Examenresultaten op 20 (slaagcijfers 2016-2017_verkort):

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

Frequentietabel voor kwantitatieve 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

Frequentietabel voor kwantitatieve 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

Frequentietabel voor kwantitatieve 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

Frequentietabel voor kwantitatieve 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

Frequentietabel voor kwantitatieve 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**		

Frequentietabel voor kwantitatieve continue gegevens

Werkwijze:

- 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).

Frequentietabel voor kwantitatieve 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**		

Frequentietabel voor kwantitatieve continue gegevens

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

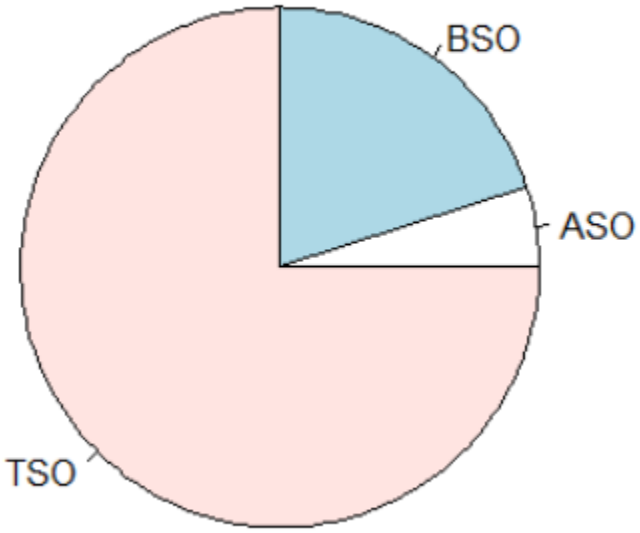
- **Klassenbreedte**: is het verschil tussen de grootste en kleinste klassengrens van een klasse.

- **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.

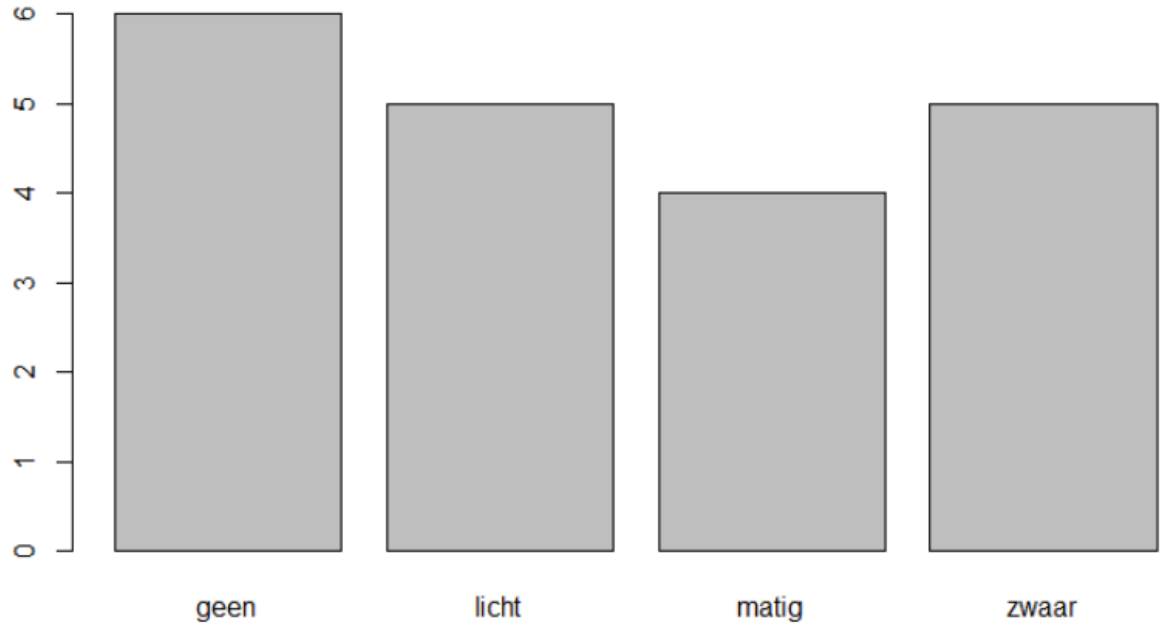
Grafische voorstelling

Cirkeldiagram



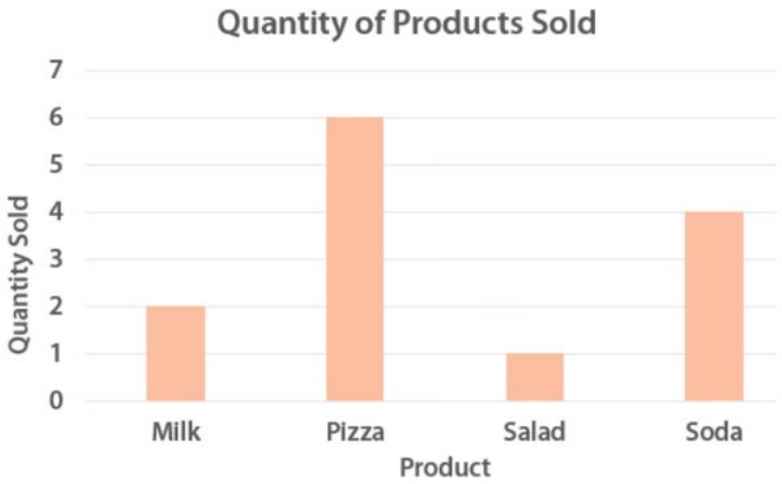
Grafische voorstelling

Staafdiagram

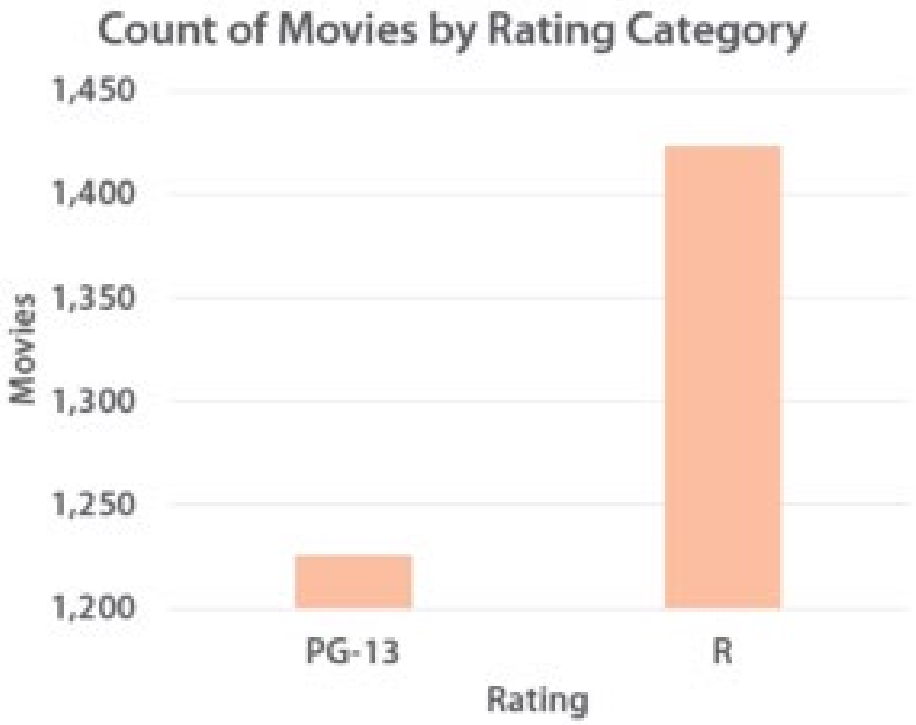


Grafische voorstelling

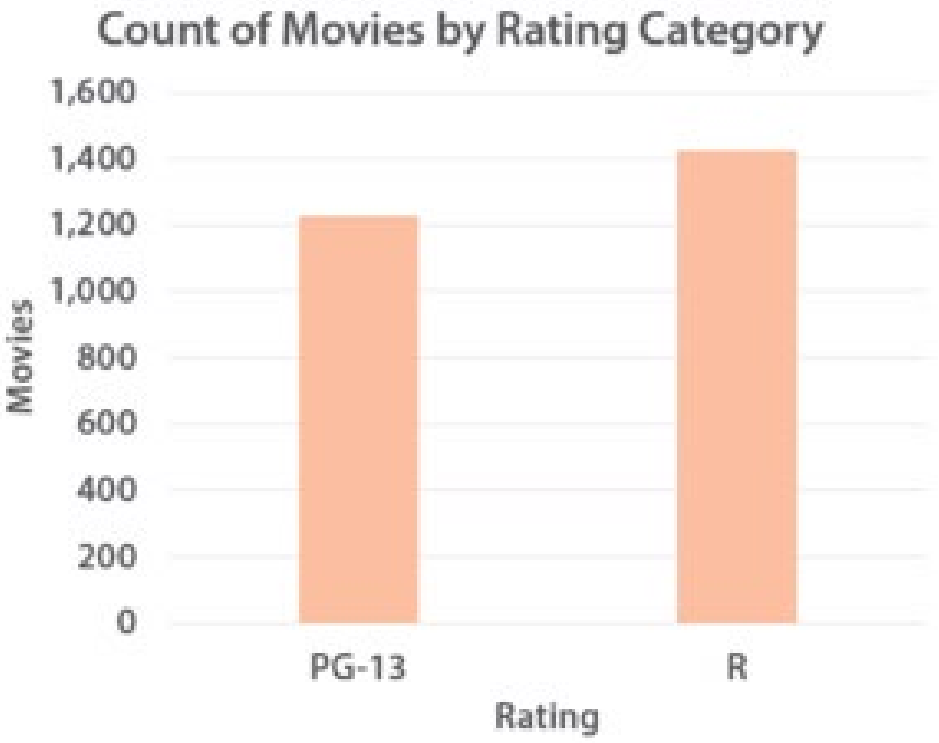
ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1



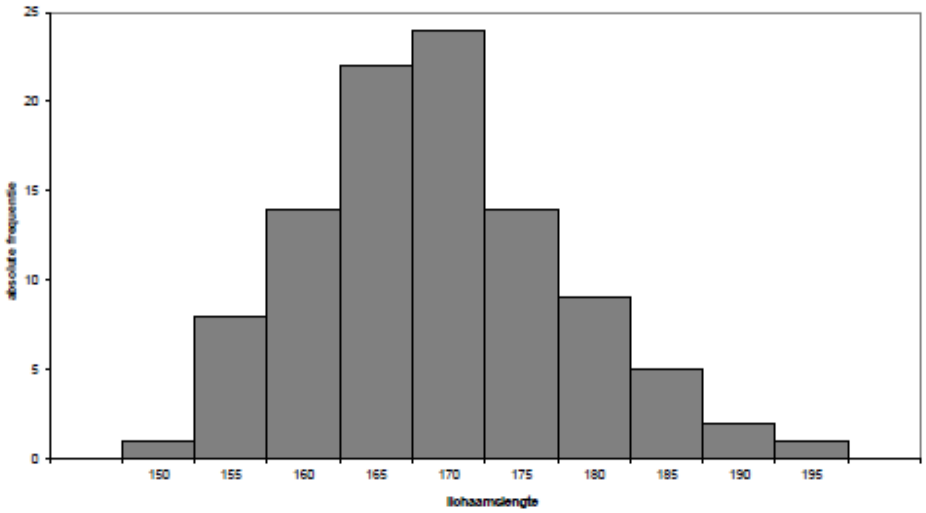
Grafische voorstelling

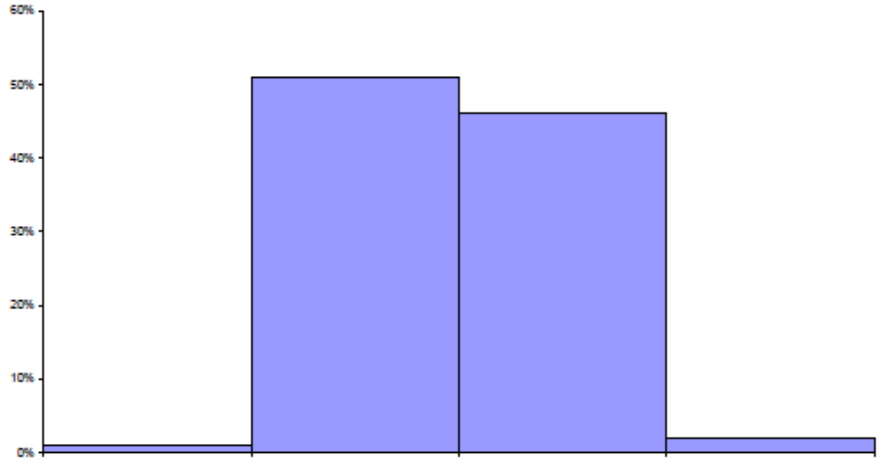


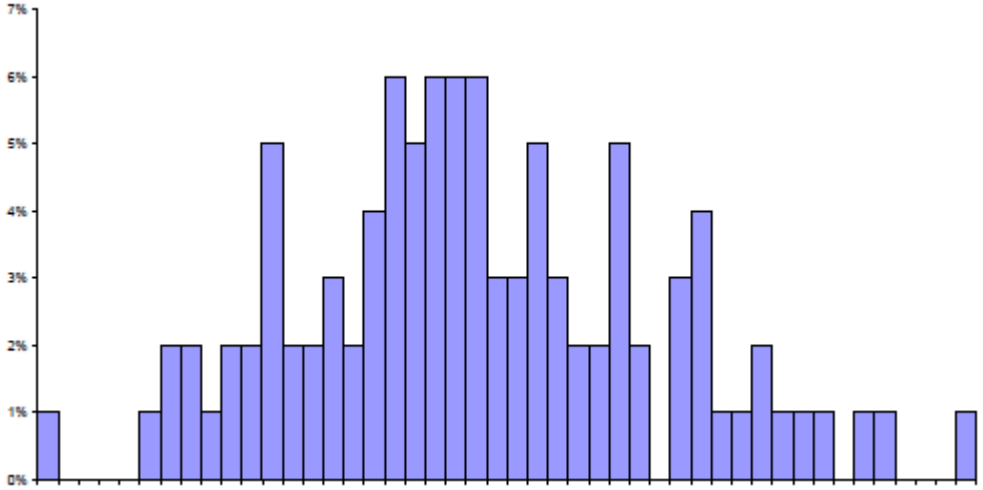
Grafische voorstelling



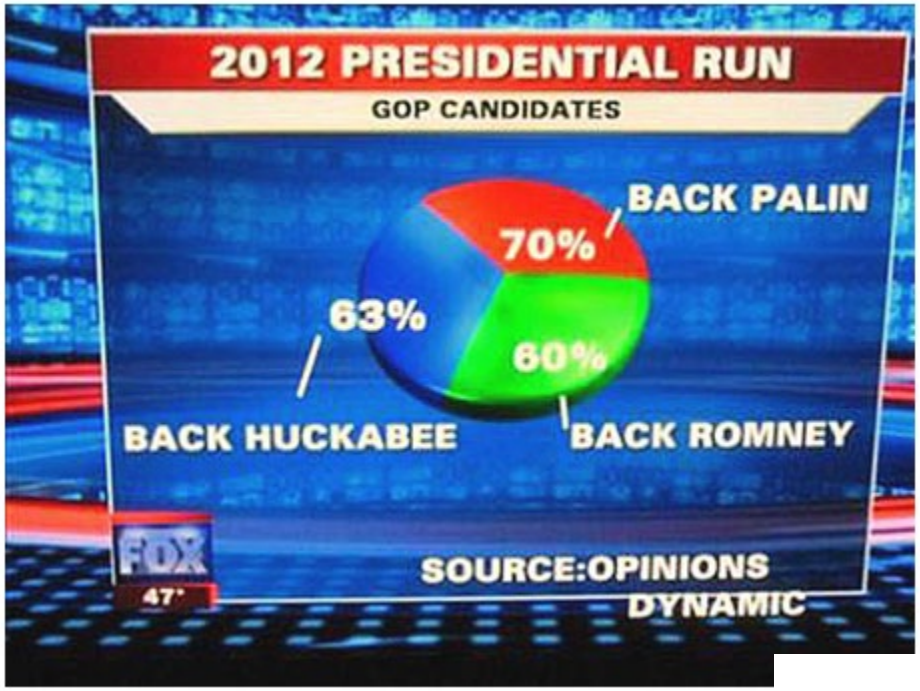
Grafische voorstelling

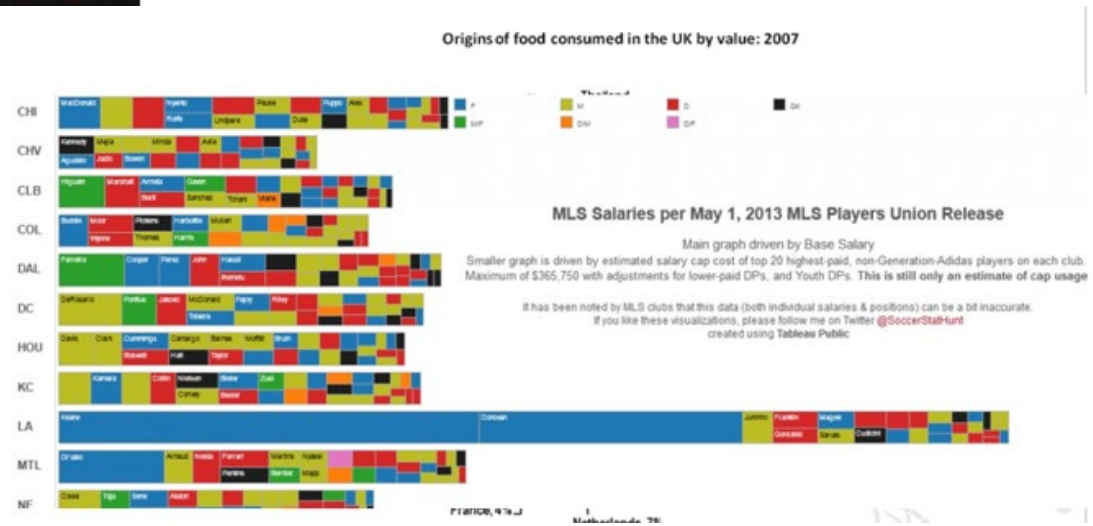






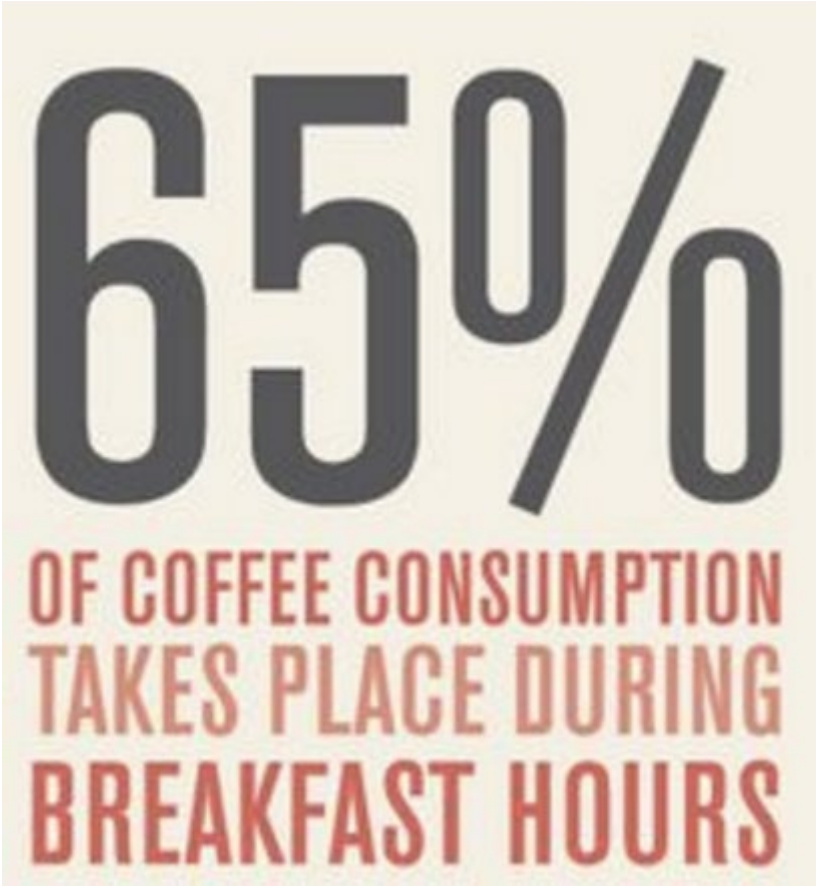
Waar het fout kan gaan





Waar het fout kan gaan

Soms is 1 cijfer voldoende…



Gegevens samenvatten

- Kengetallen voor
  - Locatie
  - Spreiding

Kengetallen voor locatie

5	2	7	6	10

105	102	107	106	110

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

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

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

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

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

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

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

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

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:

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

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

Kengetallen voor locatie

32	42	46	46	54

Mediaan = 46

2	3	4	7	8	10	10	15

Mediaan = (7 + 8)/2 = 7.5

Kengetallen voor locatie

Q_1 = het eerste kwartiel

Het getal met rangnummer $\frac{n+1}{4}$ in de geordende rij gegevens.

Q_2 = de mediaan of het tweede kwartiel

Het getal met rangnummer $\frac{n+1}{2}$ in de geordende rij gegevens.

Q_3 = het derde kwartiel

Het getal met rangnummer $3\frac{n+1}{4}$ in de geordende rij gegevens.

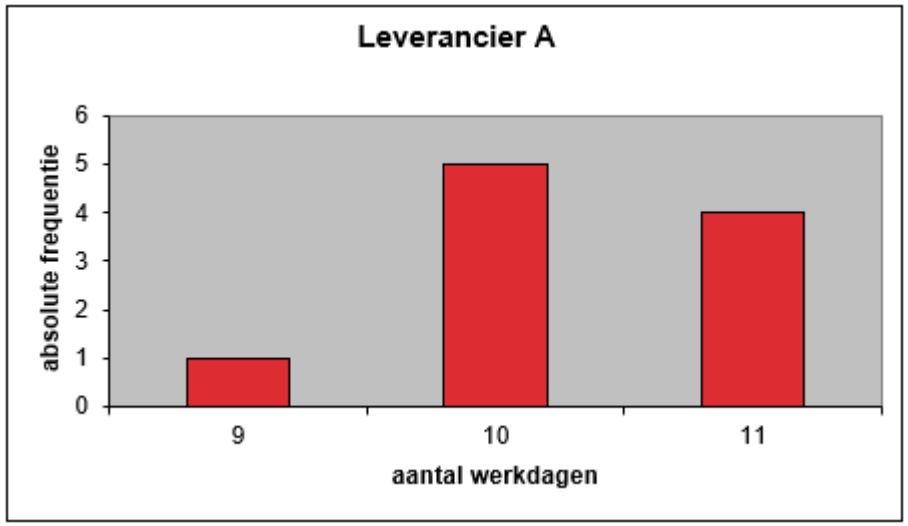
Kengetallen voor locatie

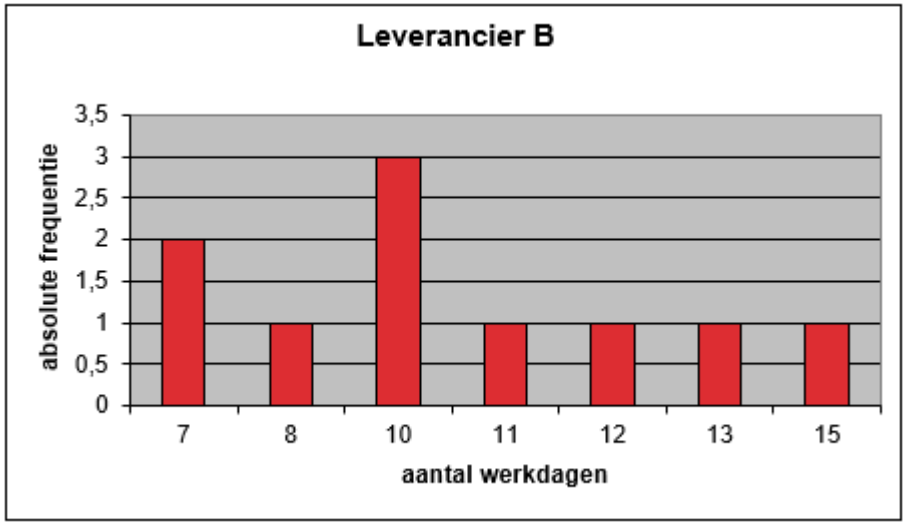
Modus = observatie met de hoogste frequentie

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





Kengetallen voor spreiding

11	10	9	10	11	11	10	11	10	10

8	10	13	7	10	11	10	7	15	12

Variantie:

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

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

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

Kengetallen voor spreiding

Variantie voor frequentietabellen:

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

Variantie voor frequentietabellen met klassenindeling:

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

Standaardafwijking:

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

Kengetal voor spreiding

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

32	42	46	46	54

Spreidingsbreedte = 54 – 32 = 22

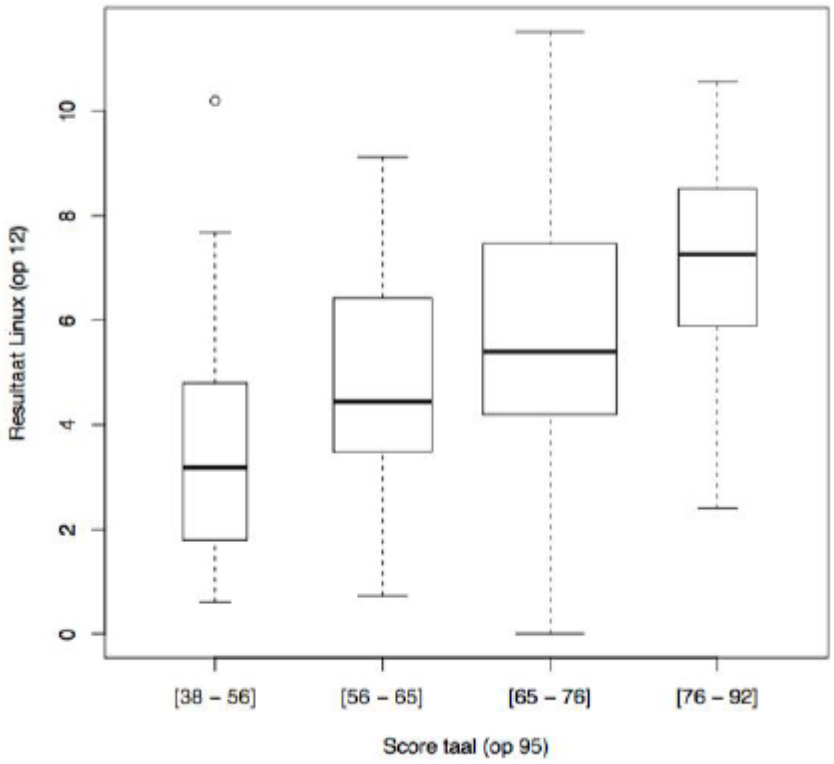
Kengetal voor spreiding

- Interkwartielafstand (IKA) = $Q_3 - Q_1$

32	42	46	46	54

Interkwartielafstand = 50 – 37 = 13

Boxplot


Resultaat Linux (op 12)
0
2
4
6
8
10
[38 – 56]
[56 – 65]
[65 – 76]
[76 – 92]
Score taal (op 95)