



Variabelen en hun verdelingen

Zie ook:
Cursusboek hoofdstuk 20

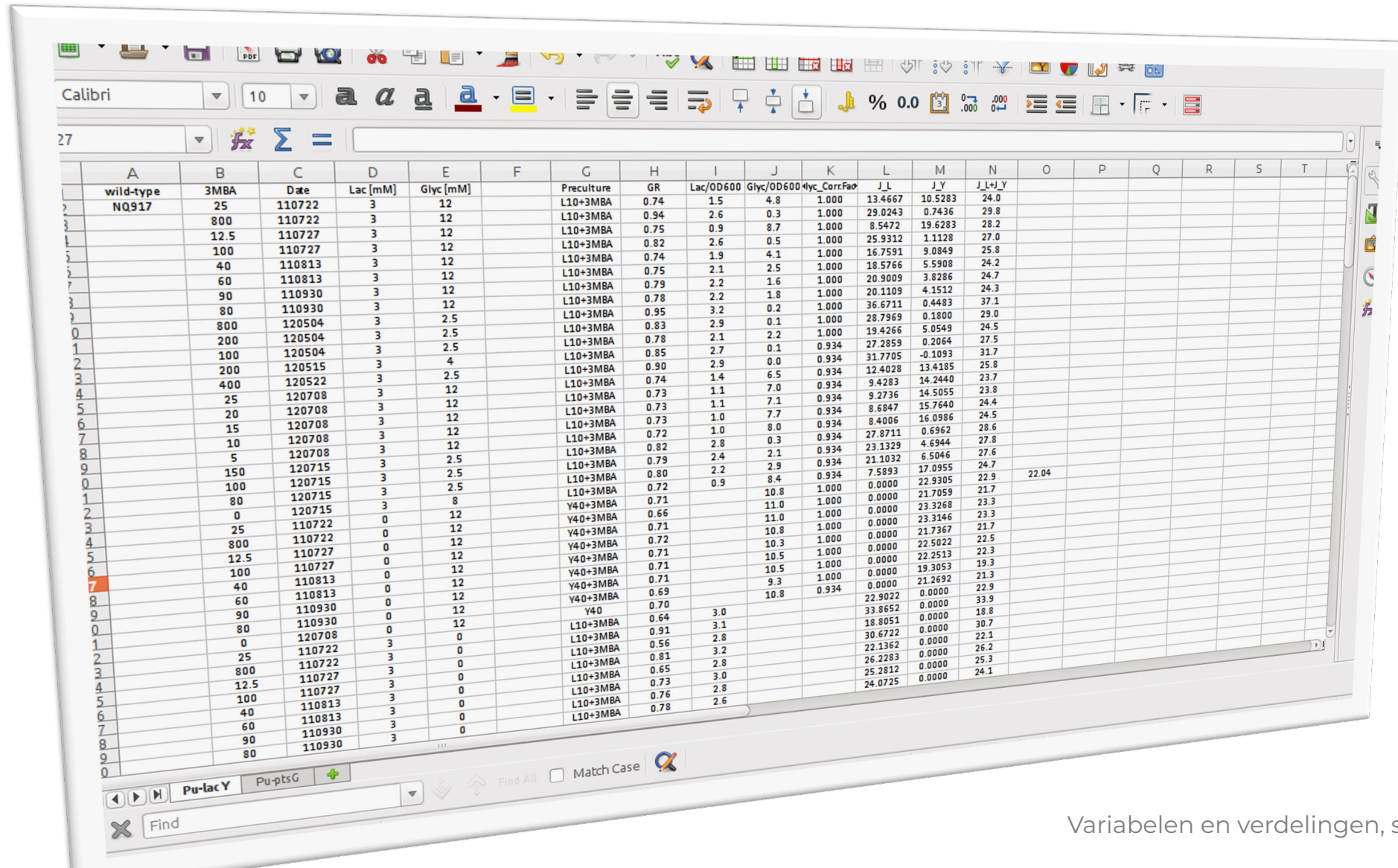
Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. (Frequentie)tabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. (Frequentie)tabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Waarom beschrijvende statistiek?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	wild-type	3MBA	Date	Lac [mM]	Glyc [mM]		Preculture	GR	Lac/OD600	Glyc/OD600	Lac_CorrFao	J_L	J_Y	J_L+J_Y						
2	NQ917	25	110722	3	12		L10+3MBA	0.74	1.5	4.8	1.000	13.4667	10.5283	24.0						
3		800	110722	3	12		L10+3MBA	0.94	2.6	0.3	1.000	29.0243	0.7436	29.8						
4		12.5	110727	3	12		L10+3MBA	0.75	0.9	8.7	1.000	8.5472	19.6283	28.2						
5		100	110727	3	12		L10+3MBA	0.82	2.6	0.5	1.000	25.9312	1.1128	27.0						
6		40	110813	3	12		L10+3MBA	0.74	1.9	4.1	1.000	16.7591	9.0849	25.8						
7		60	110813	3	12		L10+3MBA	0.75	2.1	2.5	1.000	18.5766	5.5908	24.2						
8		90	110930	3	12		L10+3MBA	0.79	2.2	1.6	1.000	20.9009	3.8286	24.7						
9		80	110930	3	12		L10+3MBA	0.78	2.2	1.8	1.000	20.1109	4.1512	24.3						
10		800	120504	3	2.5		L10+3MBA	0.95	3.2	0.2	1.000	36.6711	0.4483	37.1						
11		200	120504	3	2.5		L10+3MBA	0.83	2.9	0.1	1.000	28.7969	0.1800	29.0						
12		100	120504	3	2.5		L10+3MBA	0.78	2.1	2.2	1.000	19.4266	5.0549	24.5						
13		200	120515	3	4		L10+3MBA	0.85	2.7	0.1	0.934	27.2859	0.2064	27.5						
14		400	120522	3	2.5		L10+3MBA	0.90	2.9	0.0	0.934	31.7705	-0.1093	31.7						
15		25	120708	3	12		L10+3MBA	0.74	1.4	6.5	0.934	12.4028	13.4185	25.8						
16		20	120708	3	12		L10+3MBA	0.73	1.1	7.0	0.934	9.4283	14.2440	23.7						
17		15	120708	3	12		L10+3MBA	0.73	1.1	7.1	0.934	9.2736	14.5055	23.8						
18		10	120708	3	12		L10+3MBA	0.73	1.1	7.7	0.934	8.6847	15.7640	24.4						
19		5	120708	3	12		L10+3MBA	0.72	1.0	8.0	0.934	8.4006	16.0986	24.5						
20		150	120715	3	2.5		L10+3MBA	0.82	2.8	0.3	0.934	27.8711	0.6962	28.6						
21		100	120715	3	2.5		L10+3MBA	0.79	2.4	2.1	0.934	23.1329	4.6944	27.8						
22		80	120715	3	2.5		L10+3MBA	0.80	2.2	2.9	0.934	21.1032	6.5046	27.6						
23		0	120715	3	8		L10+3MBA	0.72	0.9	8.4	0.934	7.5893	17.0955	24.7	22.04					
24		25	110722	0	12		Y40+3MBA	0.71		10.8	1.000	0.0000	22.9305	22.9						
25		800	110722	0	12		Y40+3MBA	0.66		11.0	1.000	0.0000	21.7059	21.7						
26		12.5	110727	0	12		Y40+3MBA	0.71		11.0	1.000	0.0000	23.3268	23.3						
27		100	110727	0	12		Y40+3MBA	0.72		10.8	1.000	0.0000	23.3146	23.3						
28		40	110813	0	12		Y40+3MBA	0.71		10.3	1.000	0.0000	21.7367	21.7						
29		60	110813	0	12		Y40+3MBA	0.71		10.5	1.000	0.0000	22.5022	22.5						
30		90	110930	0	12		Y40+3MBA	0.71		10.5	1.000	0.0000	22.2513	22.3						
31		80	110930	0	12		Y40+3MBA	0.69		9.3	1.000	0.0000	19.3053	19.3						
32		0	120708	0	0		Y40+3MBA	0.70		10.8	0.934	0.0000	21.2692	21.3						
33		25	110722	3	0		Y40	0.64	3.0			22.9022	0.0000	22.9						
34		800	110727	3	0		L10+3MBA	0.91	3.1			33.8652	0.0000	33.9						
35		12.5	110727	3	0		L10+3MBA	0.56	2.8			18.8051	0.0000	18.8						
36		40	110813	3	0		L10+3MBA	0.81	3.2			30.6722	0.0000	30.7						
37		60	110930	3	0		L10+3MBA	0.65	2.8			22.1362	0.0000	22.1						
38		80	110930	3	0		L10+3MBA	0.73	3.0			26.2283	0.0000	26.2						
39		0	120708	0	0		L10+3MBA	0.76	2.6			25.2812	0.0000	25.3						
40		25	110722	3	0		L10+3MBA	0.78				24.0725	0.0000	24.1						

Drie oplossingen

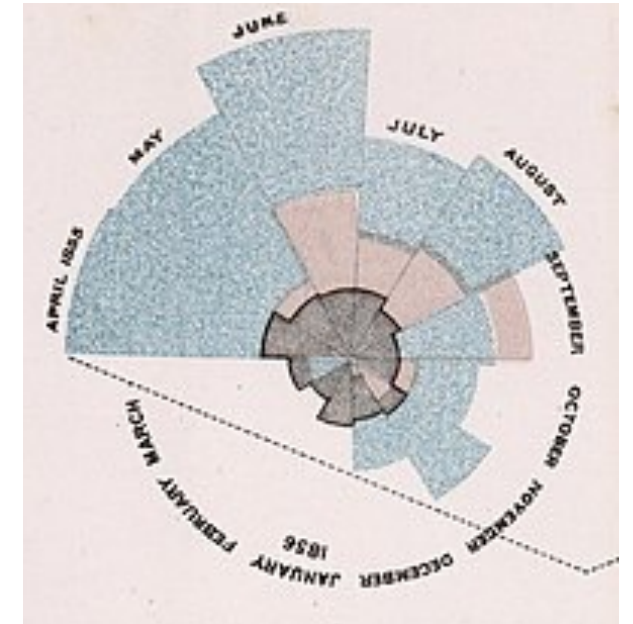
1. Kengetallen / statistieken

- typische waarde
- variatie
- proporties / percentage
- ...

2. Frequentietabellen

Gezondheidsstatus	Absolute frequentie
Slecht	187
Matig	1010
Goed	2956
Zeer Goed	2508
Uitstekend	878

3. Visualiseren

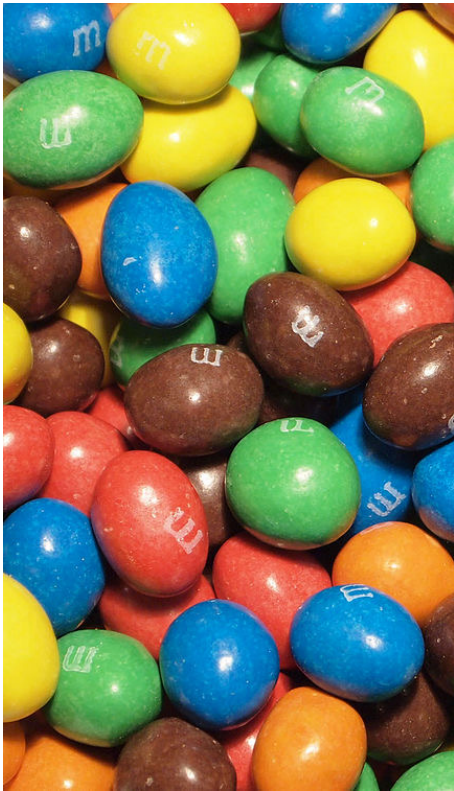


Inhoud

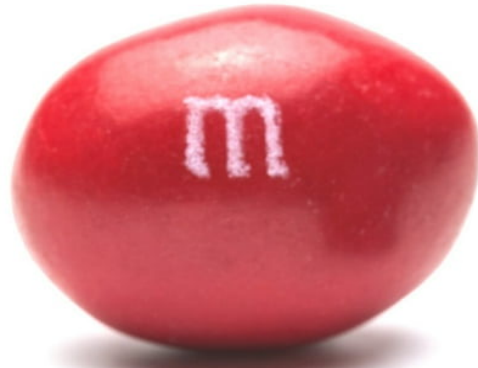
1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. Frequentietabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Samenvatting en conclusies

Populaties, eenheden, variabele, verdeling

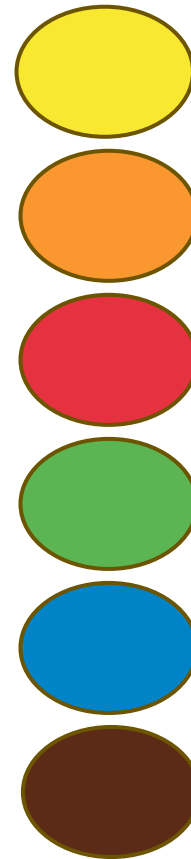
Populatie



Eenheid van de populatie



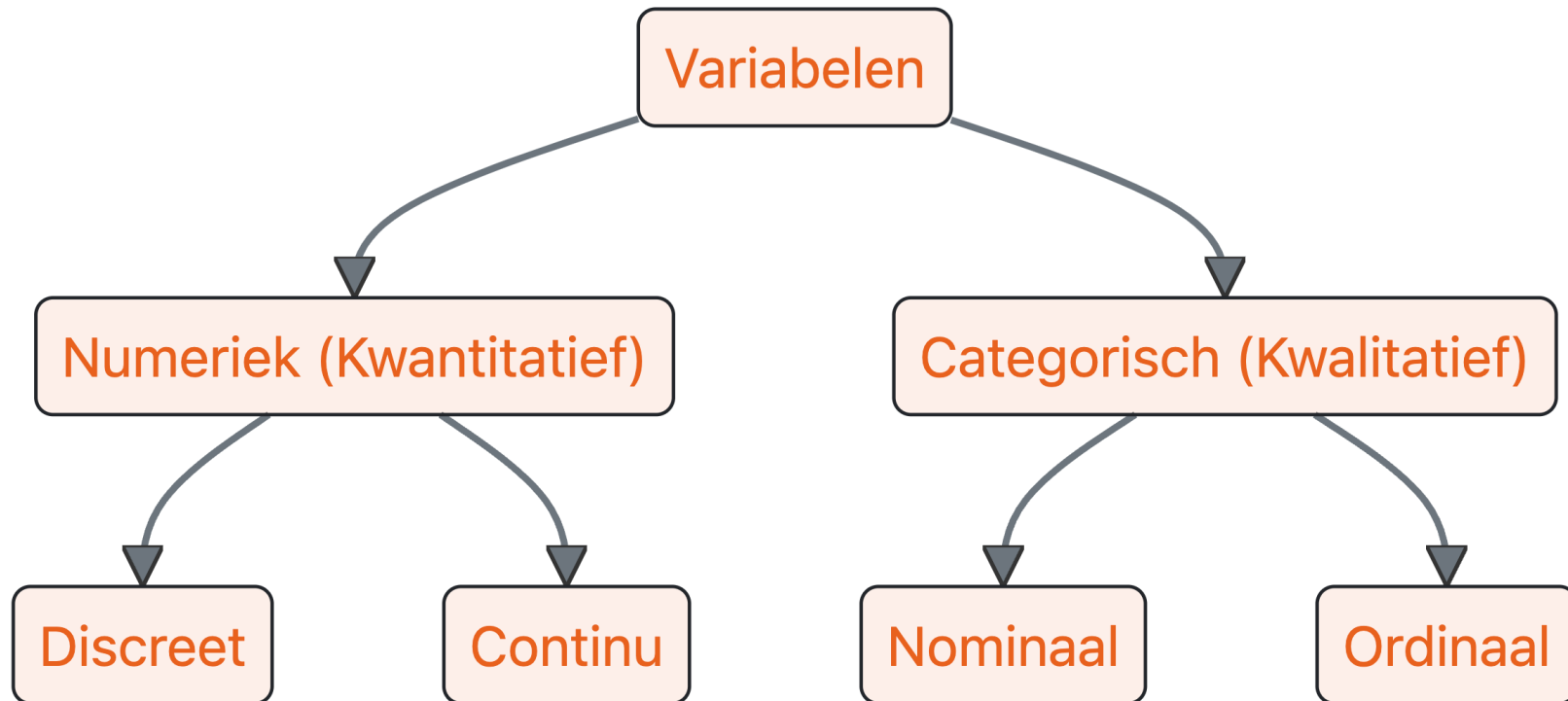
Variabele



Verdeling



Indeling van variabelen



Indeling variabelen



Van welk type is de variabele “kleur” bij M&M’s?

- ☐ Discreet
- ☐ Continu
- ☐ Nominaal
- ☐ Ordinaal



Indeling variabelen



Van welk type variabele is het stadium van de levenscyclus van een organisme?
(bijvoorbeeld: ei, rups, pop, vlinder)

- ☐ Discreet
- ☐ Continu
- ☐ Nominaal
- ☐ Ordinaal



Indeling variabelen

Van welk type variabele is het aantal puppy's in een worp honden?

- ☐ Discreet
- ☐ Continu
- ☐ Nominaal
- ☐ Ordinaal



De dataset NHANES

- **National Health and Nutrition Examination Survey:** grootschalig gezondheidsonderzoek in de VS



- **NHANES dataset in R:** steekproef van 10000 personen library in R



- **Groot aantal variabelen** waaronder gezondheidsstatus, lichaamsgewicht, lengte, bloeddruk, burgerlijke staat, ...

StatistiekRutger - R

NHANES

Filter Cols: << 1 - 50 >>

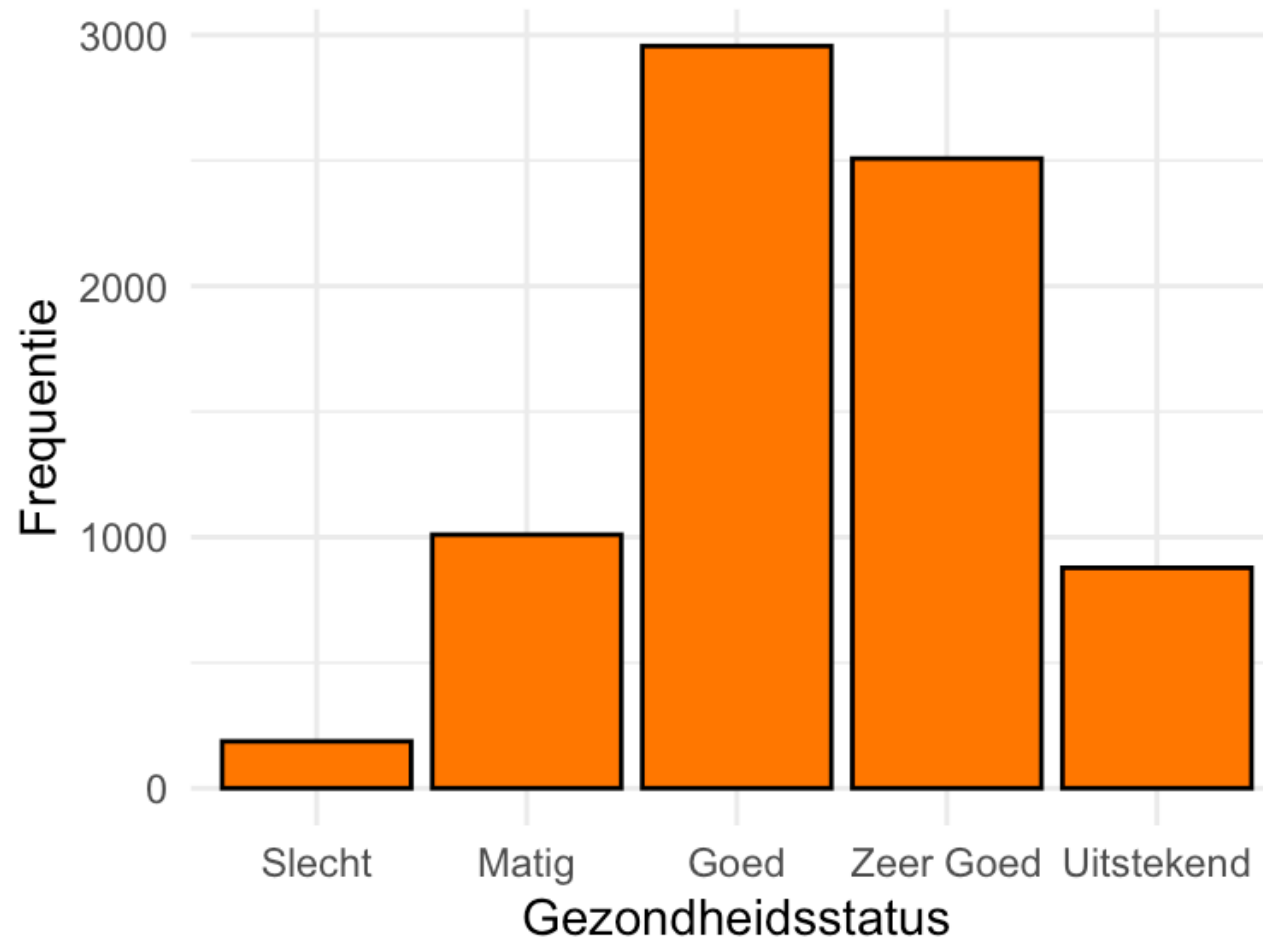
	ID	SurveyYr	Gender	Age	AgeDecade	AgeMonths	Race1	Ra
1	51624	2009_10	male	34	30-39	409	White	Na
2	51624	2009_10	male	34	30-39	409	White	Na
3	51624	2009_10	male	34	30-39	409	White	Na
4	51625	2009_10	male	4	0-9	49	Other	Na
5	51630	2009_10	female	49	40-49	596	White	Na
6	51638	2009_10	male	9	0-9	115	White	Na
7	51646	2009_10	male	8	0-9	101	White	Na
8	51647	2009_10	female	45	40-49	541	White	Na
9	51647	2009_10	female	45	40-49	541	White	Na
10	51647	2009_10	female	45	40-49	541	White	Na
11	51654	2009_10	male	66	60-69	795	White	Na
12	51656	2009_10	male	58	50-59	707	White	Na

Showing 1 to 13 of 10,000 entries, 76 total columns

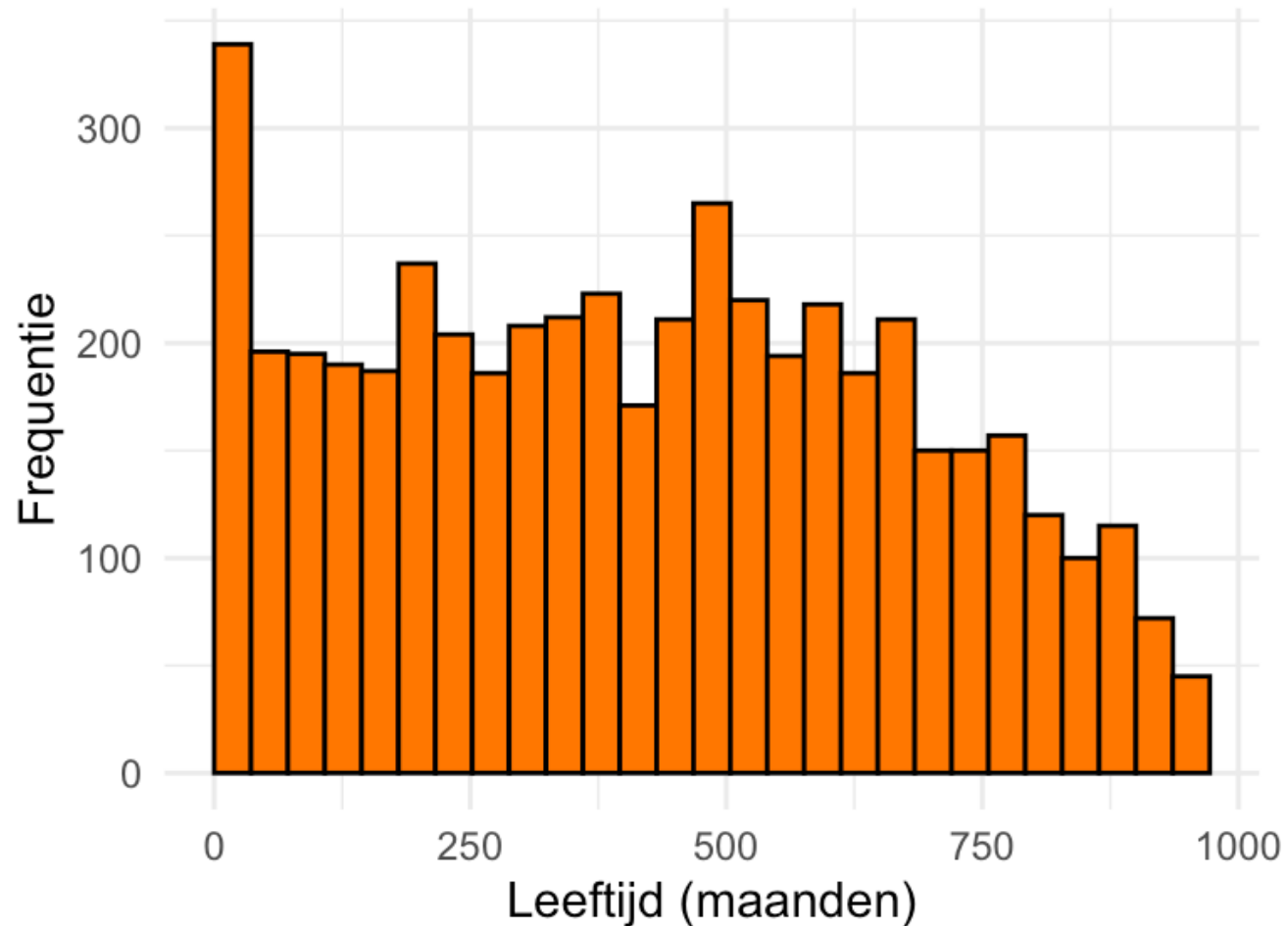
Console Terminal Background Jobs

R 4.4.3 ~ /Library/CloudStorage/OneDrive-UniversiteitUtrecht/Documents/Teaching/Kwantitatieve

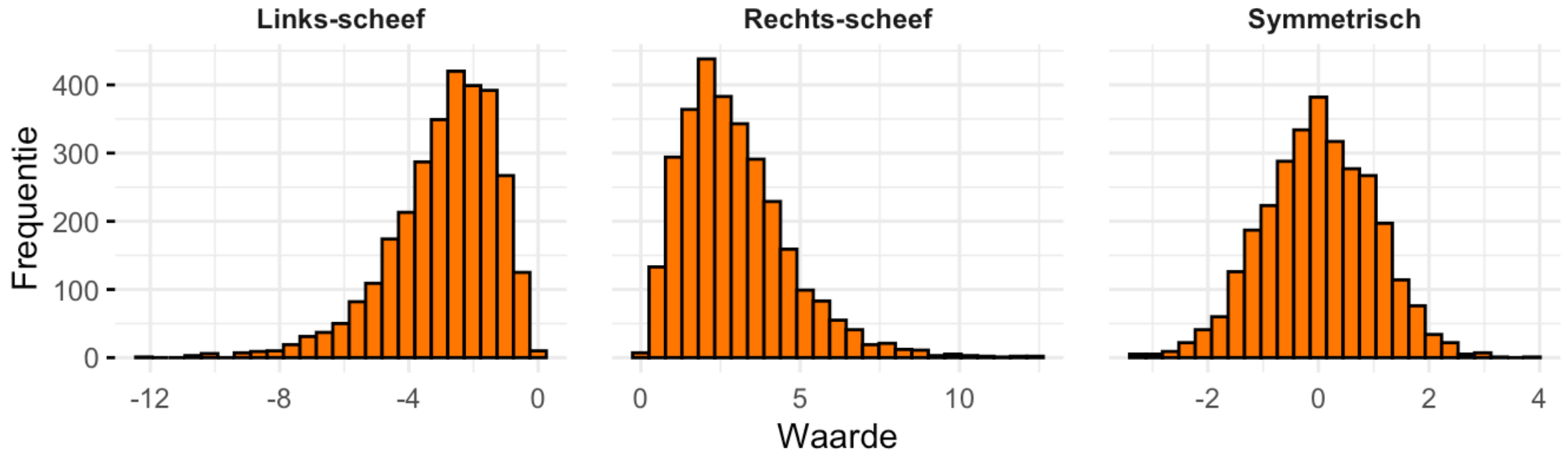
Verdeling van een categorische variabele



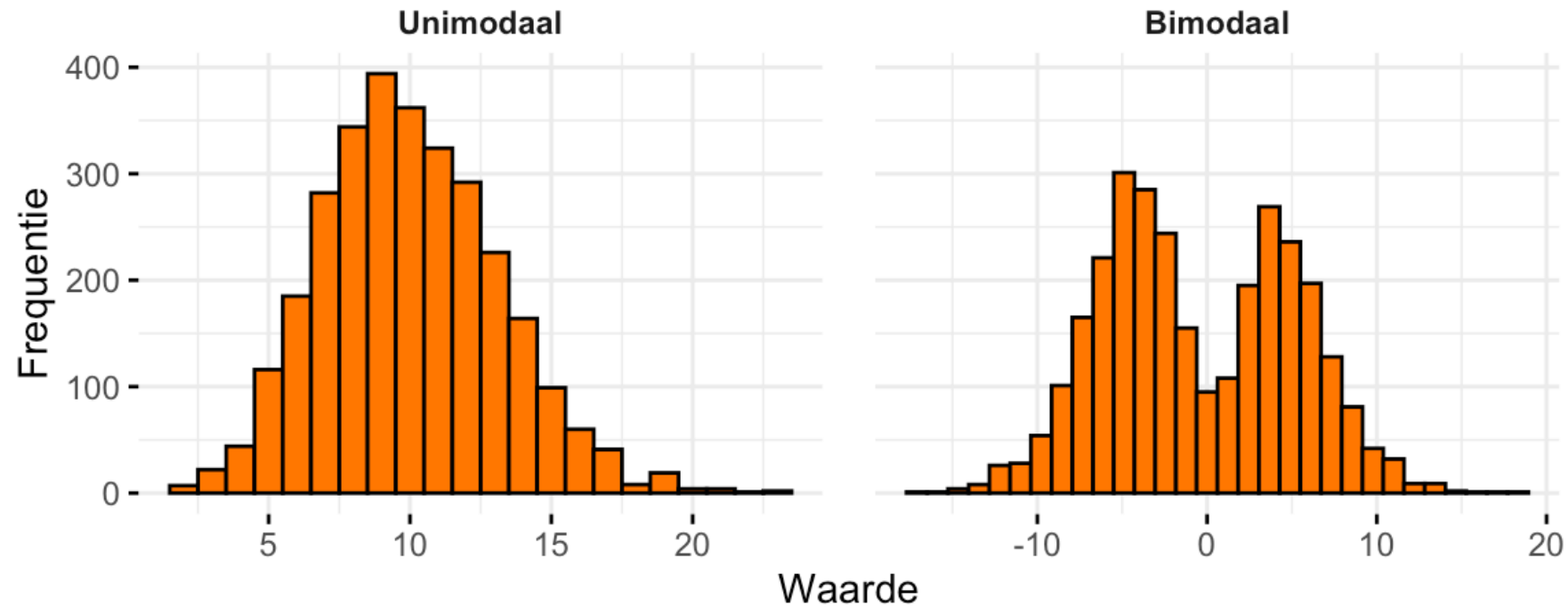
Verdeling van een continue variabele



De vorm van verdelingen: scheef vs symmetrisch



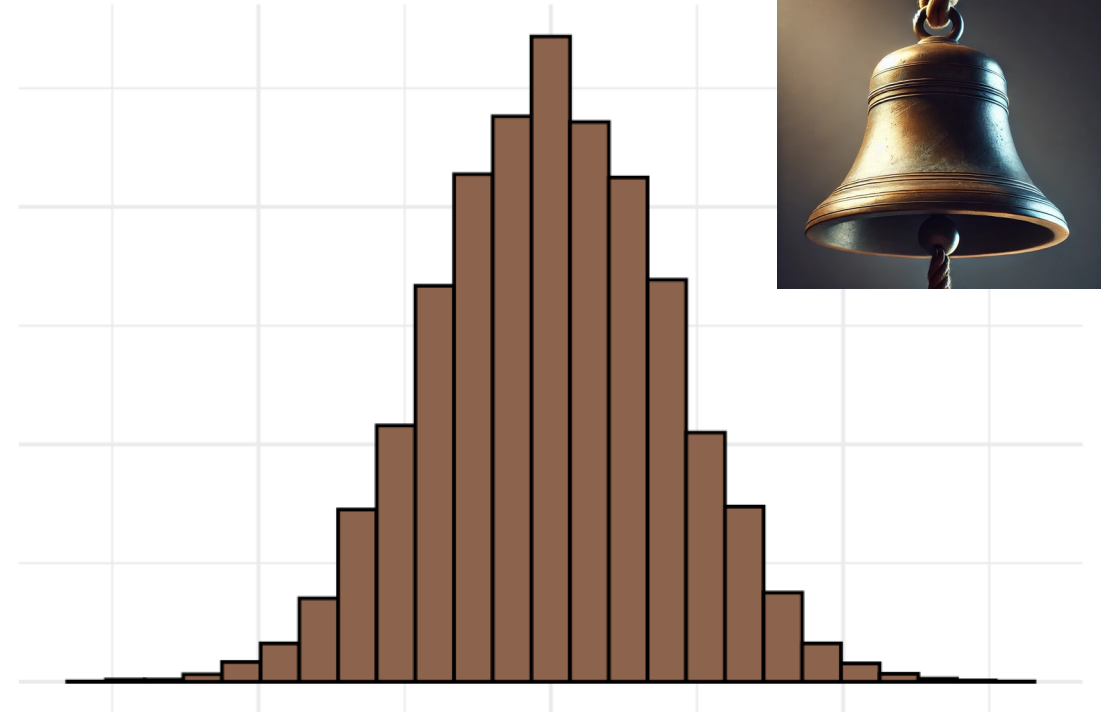
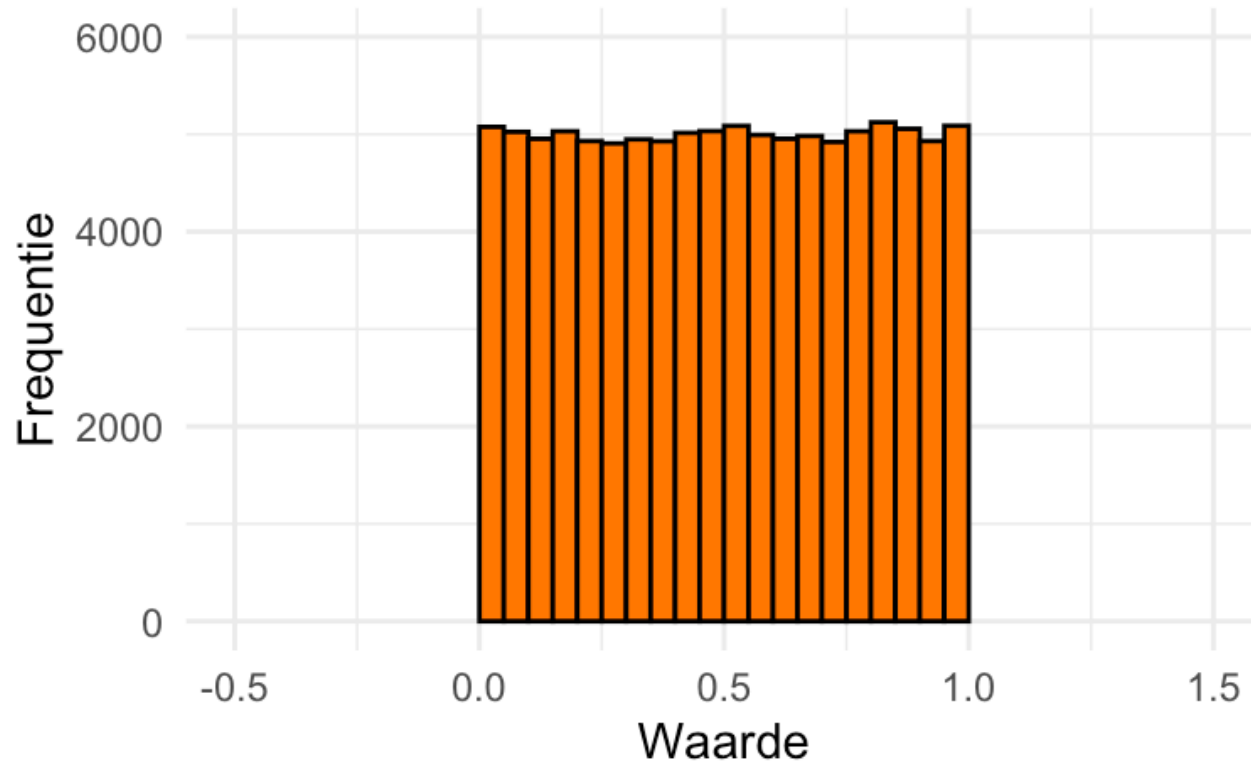
De vorm van verdelingen: unimodaal, bimodaal



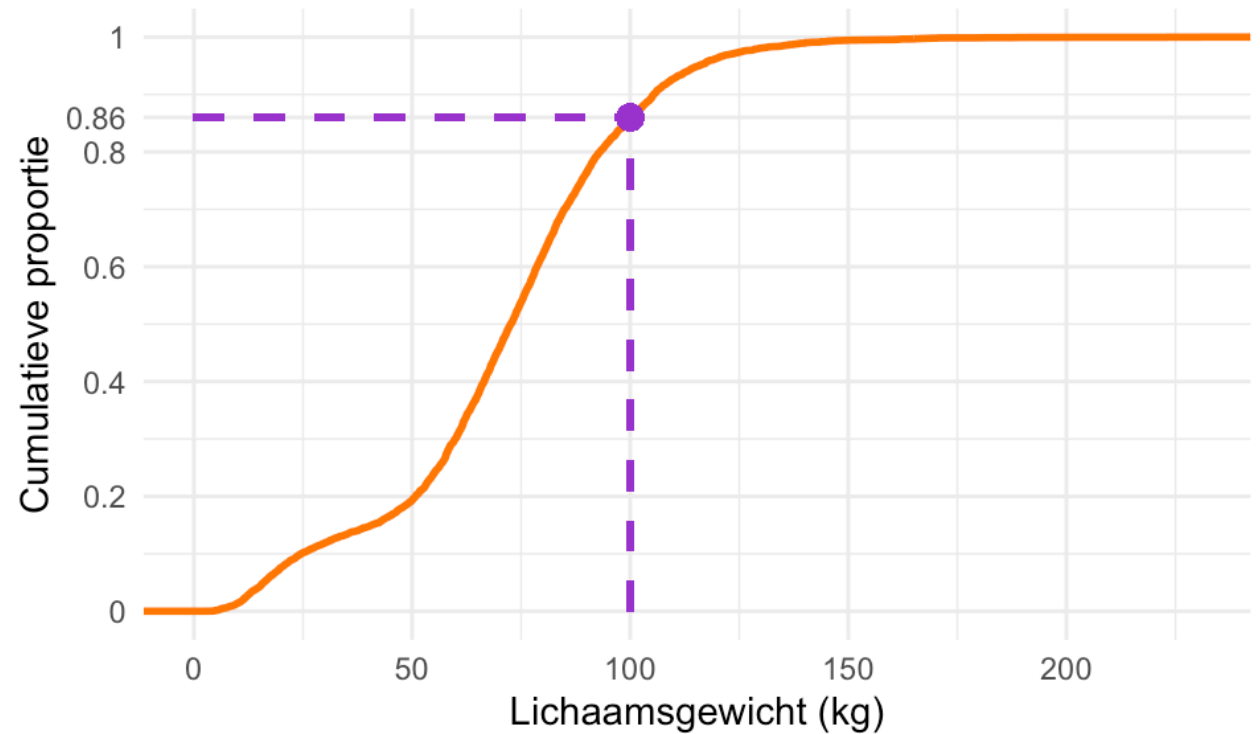
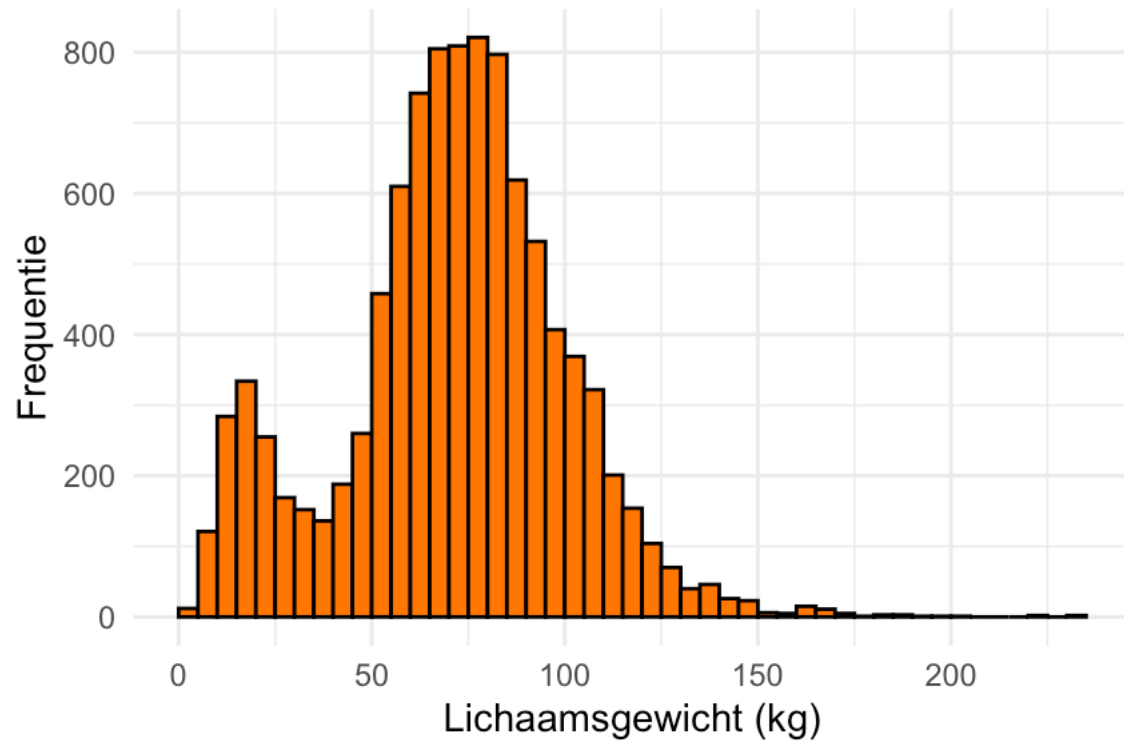
Uniform...

en...

klokvormig



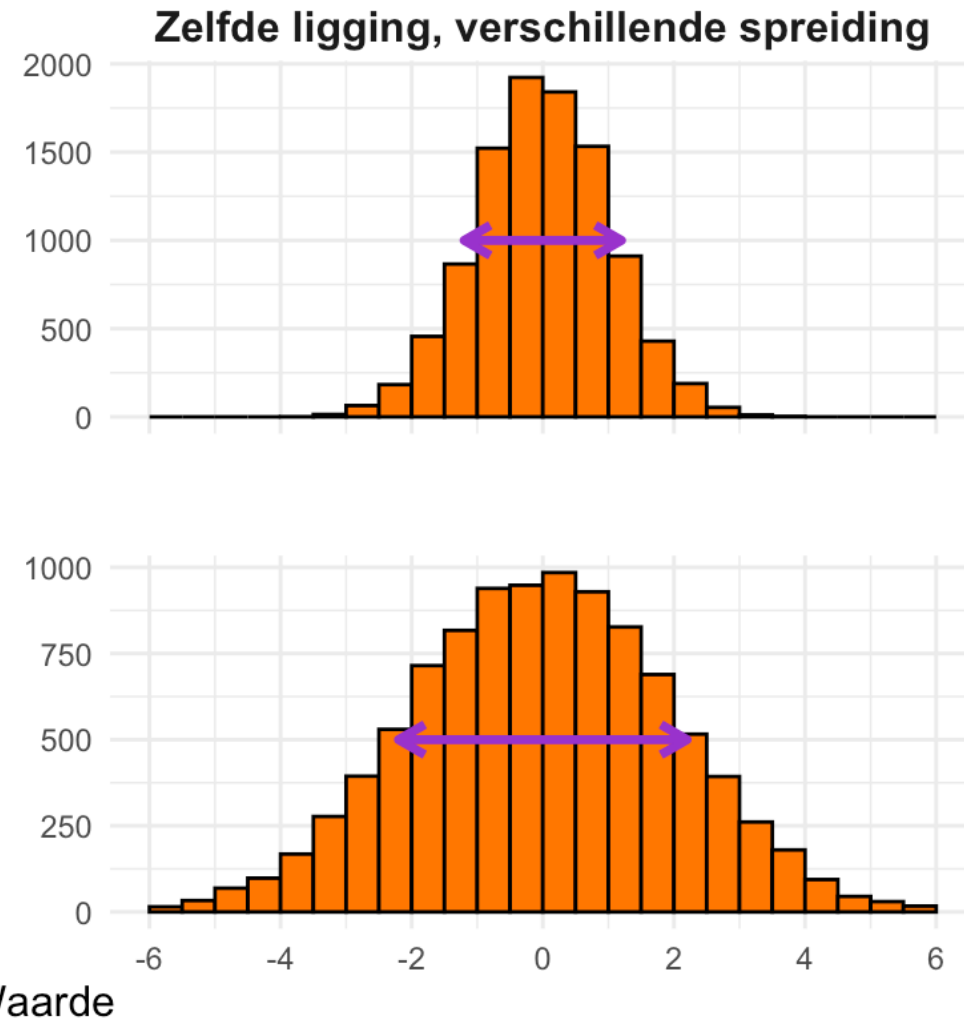
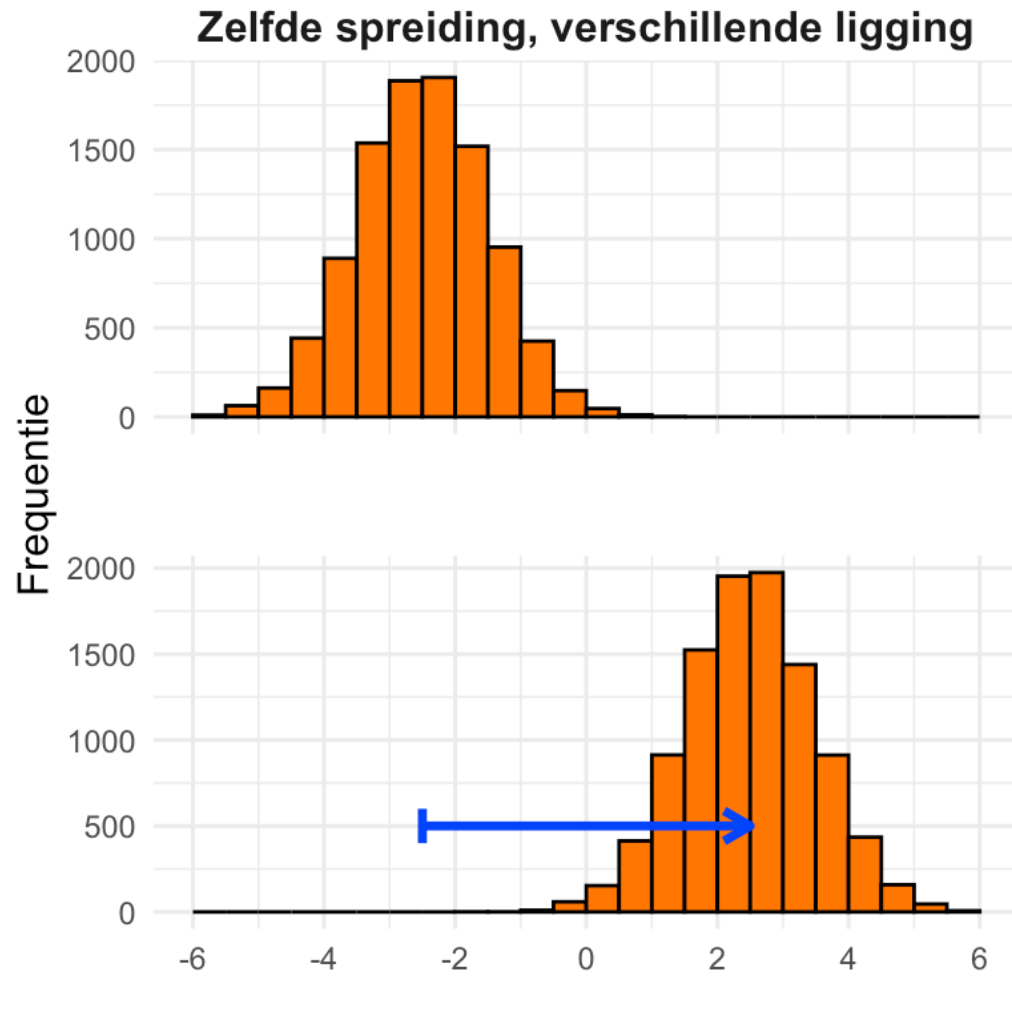
De cumulatieve verdeling



Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. Frequentietabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Ligging en spreiding



Maten van ligging (centrumsmaten) 1: de mediaan

Middelste getal na sorteren op grootte

Tentamencijfers:

7 7,5 1 8,5 9.

Sorteren:

1 7 7,5 8,5 9.

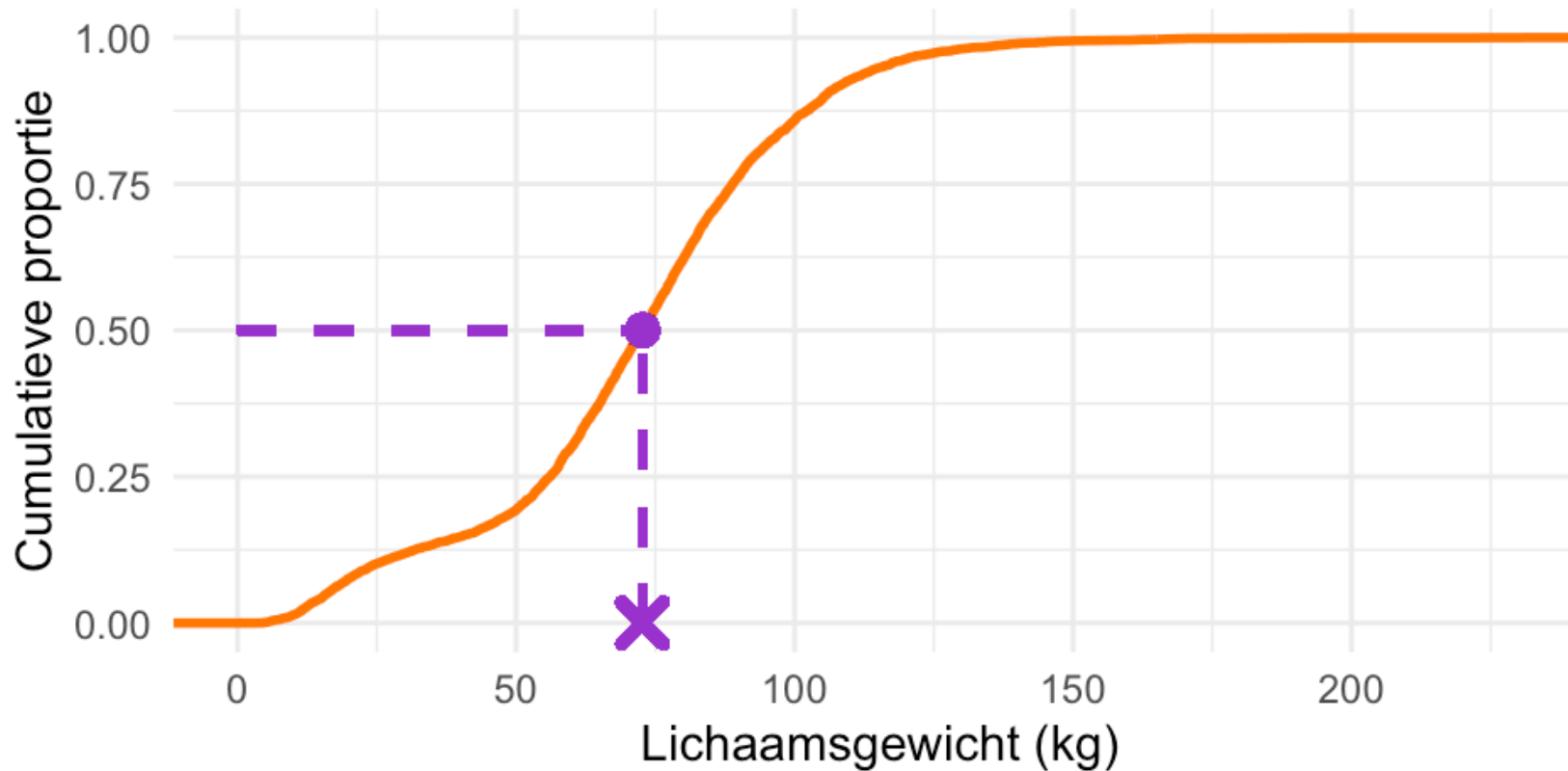
$$\tilde{x} = 7,5$$

Bij even aantal cijfer:

1 7 7,5 8,5 9 10.

$$\tilde{x} = \frac{7,5 + 8,5}{2} = 8.$$

Maten van ligging (centrumsmaten) 1: de mediaan



Maten van ligging (centrumsmaten) 2: het gemiddelde

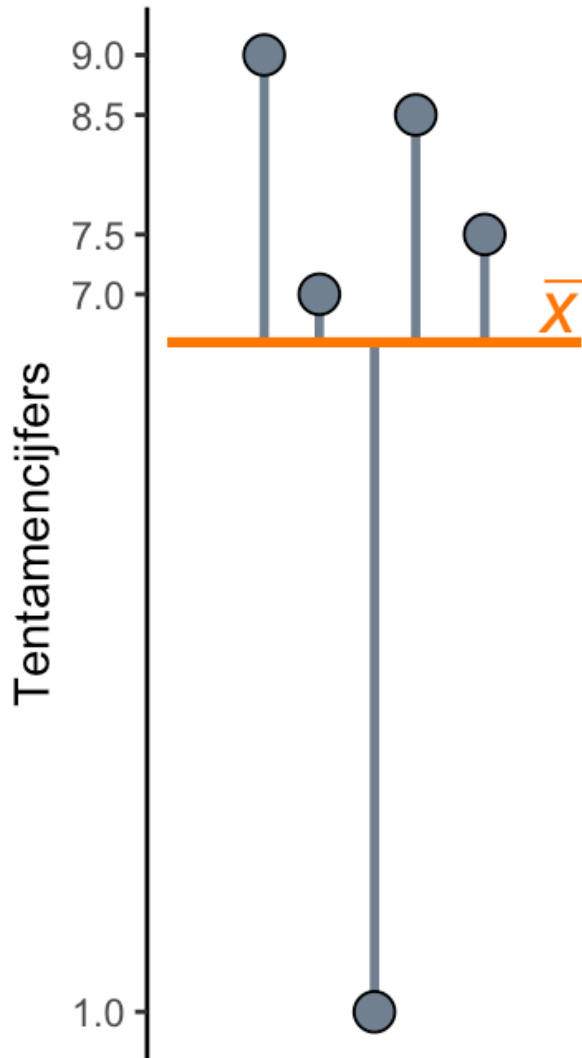
Som van de getallen gedeeld door het aantal getallen:

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

Zwaartepunt van de getallen:



Maten van ligging (centrummaten) 2: het gemiddelde



Grijze lijntjes: **residuen**.

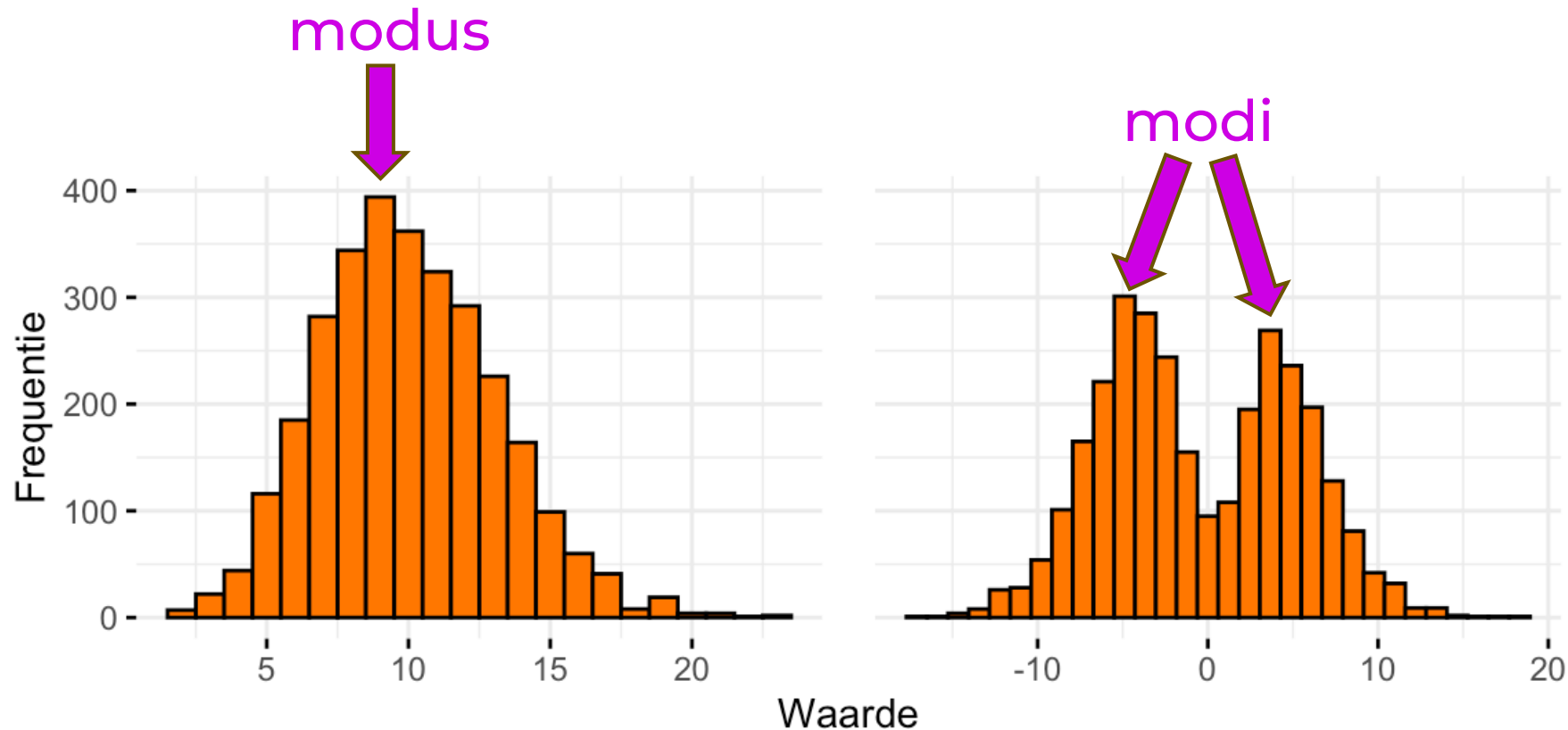
Gemiddelde minimaliseert de kwadratensom:

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

Alsof de grijze lijntjes elastiekjes zijn
die de oranje lijn naar zich toe trekken.

Maten van ligging (centrumsmaten) 3: de modus

De waarde (of klasse) die het vaakst voorkomt.



Maten van spreiding (spreidingsmaten) 1: het bereik

Het interval van het kleinste tot het grootste getal.

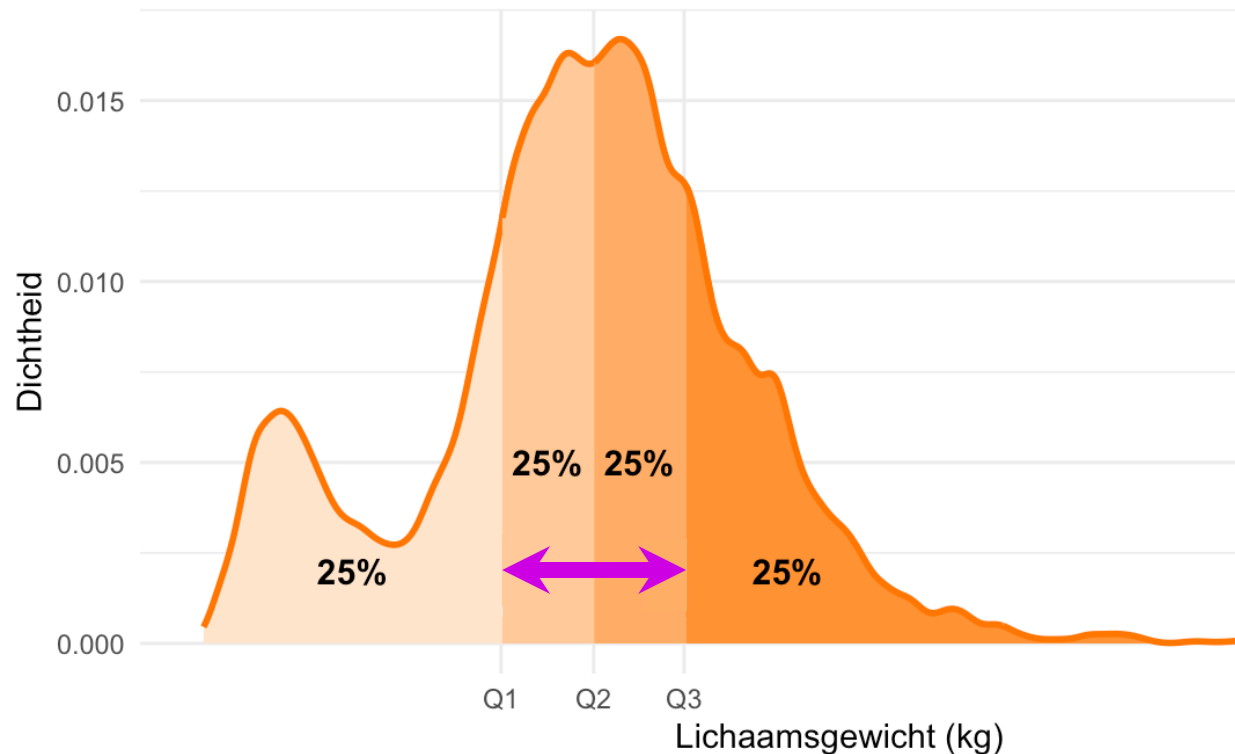
Tentamencijfers:

7 7,5 1 8,5 9.

Bereik:

[1; 9]

Maten van spreiding (spreidingsmaten) 2: de interkwartielafstand



Kwartielen Q1, Q2, Q3:

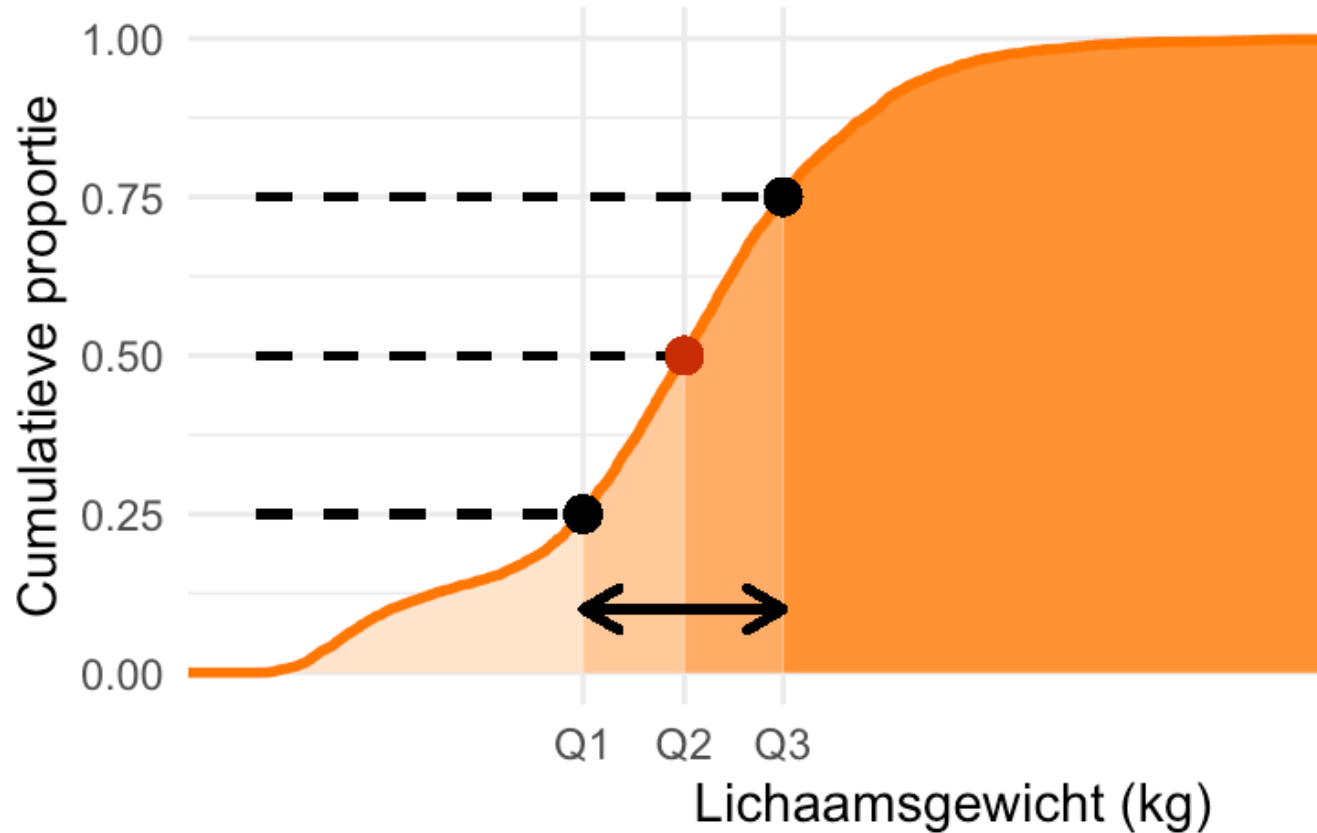
- Verdeel de gegevens in vier gelijke delen;
- Q1, Q2, Q3 zijn de grenzen tussen die delen.
- Q2 = de mediaan

Interkwartielafstand:

$$\text{IKA} = Q3 - Q1$$

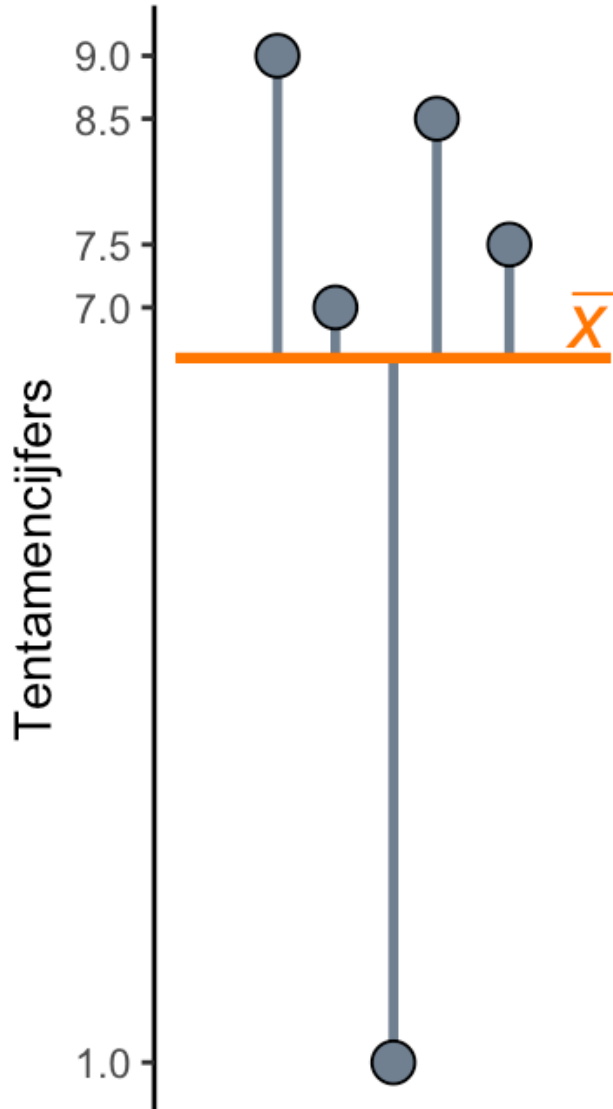
Breedte van het interval dat de middelste 50% van de gegevens bevat

Kwartielen, percentielen, kwantielen, en de cumulatieve verdeling



Kwartiel	Percentiel	Kwantiel
Q1	P25	0,25
Q2	P50	0,5
Q3	P75	0,75

Maten van spreiding (spreidingsmaten) 3: variantie en standaarddeviatie



Variantie van een reeks waarnemingen:

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

Gemiddelde van het kwadraat van de residuen.

Standaarddeviatie:

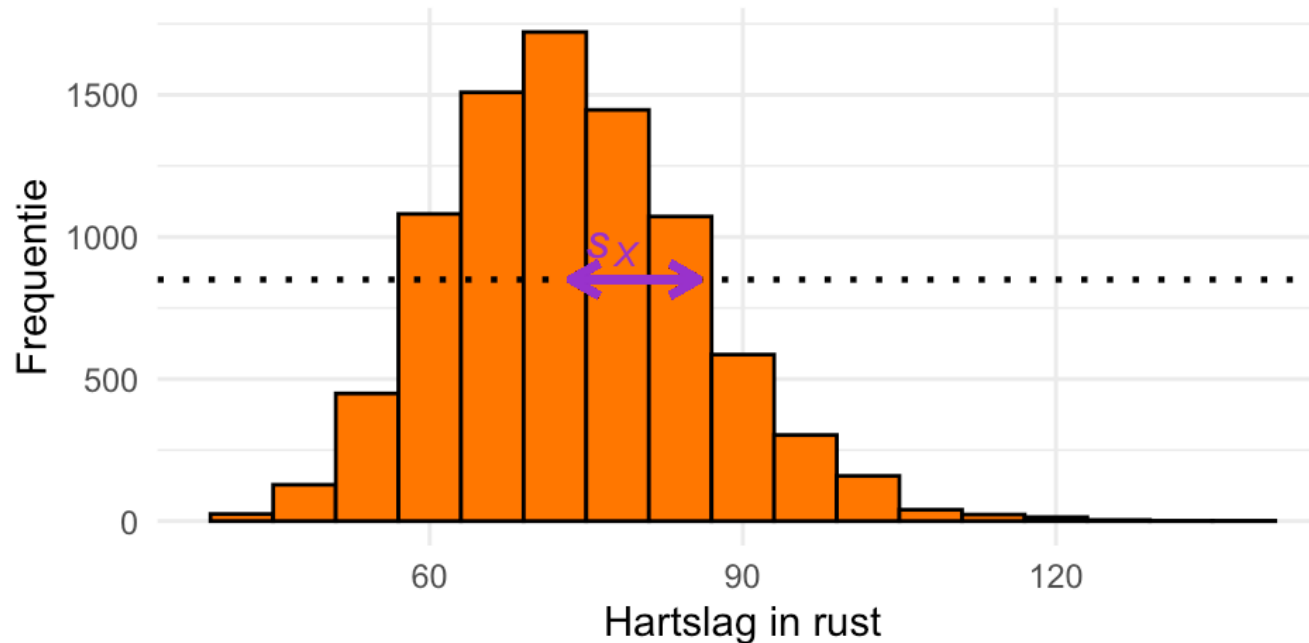
$$s_X = \sqrt{V_X}$$

Dus:

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

Interpretatie van de standaarddeviatie

- “Typische” afwijking van een waarde van het gemiddelde.
Een afwijking van $\sim s_x$ van het gemiddelde is niet verrassend.
- Bij een belvormige verdeling:



Her-tentamen



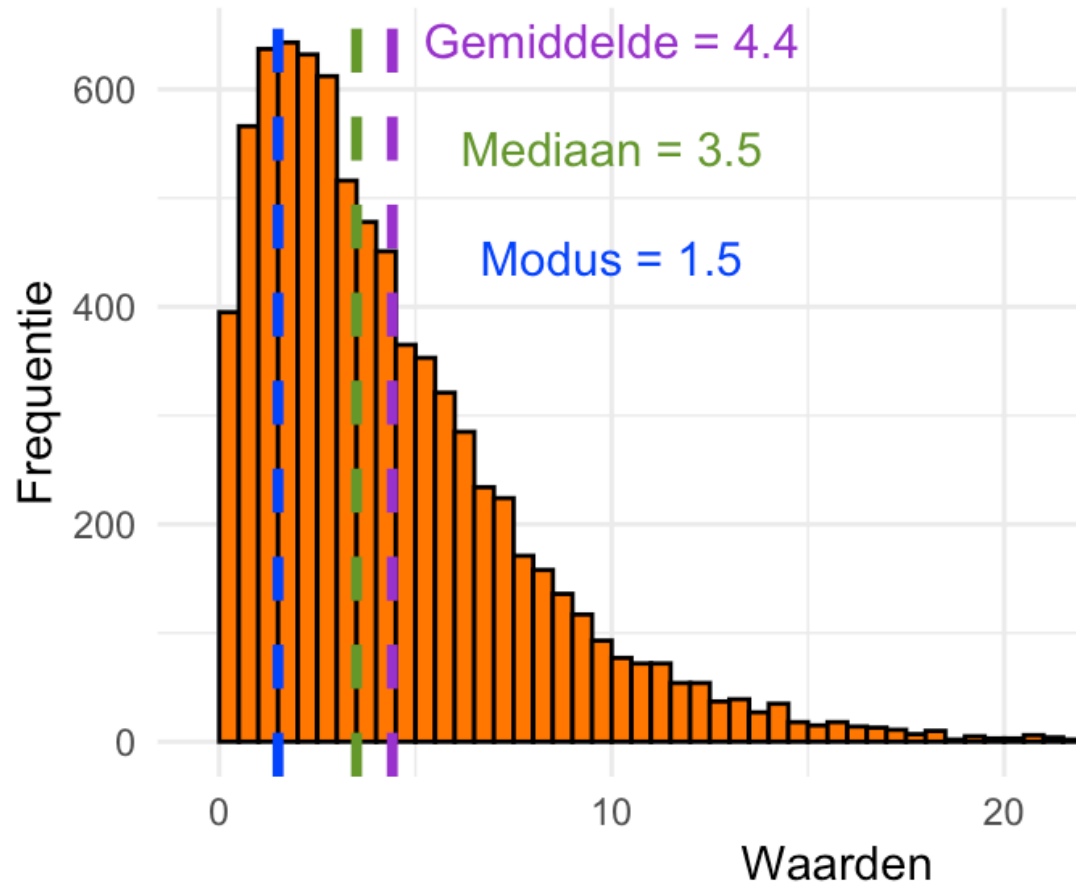
Je tentamencijfers zijn:



Je doet een hertentamen; de 1 wordt een 7,5.

- Wat gebeurt er met het gemiddelde?
- Wat gebeurt er met de mediaan?

Gevoeligheid voor extreme waarden / scheefheid



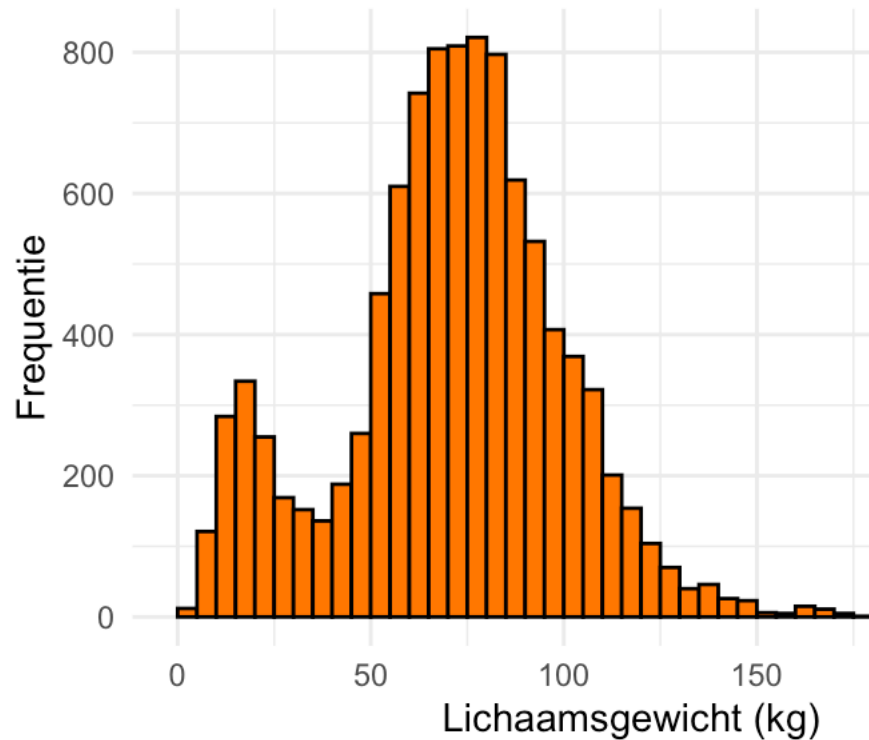
Het gemiddelde is veel **gevoeliger** dan de mediaan of de modus voor waarden die sterk afwijken van het gemiddelde.

- de staart van de verdeling
- bijzonder afwijkende punten (**uitbijters / outliers**)

Invoerfout



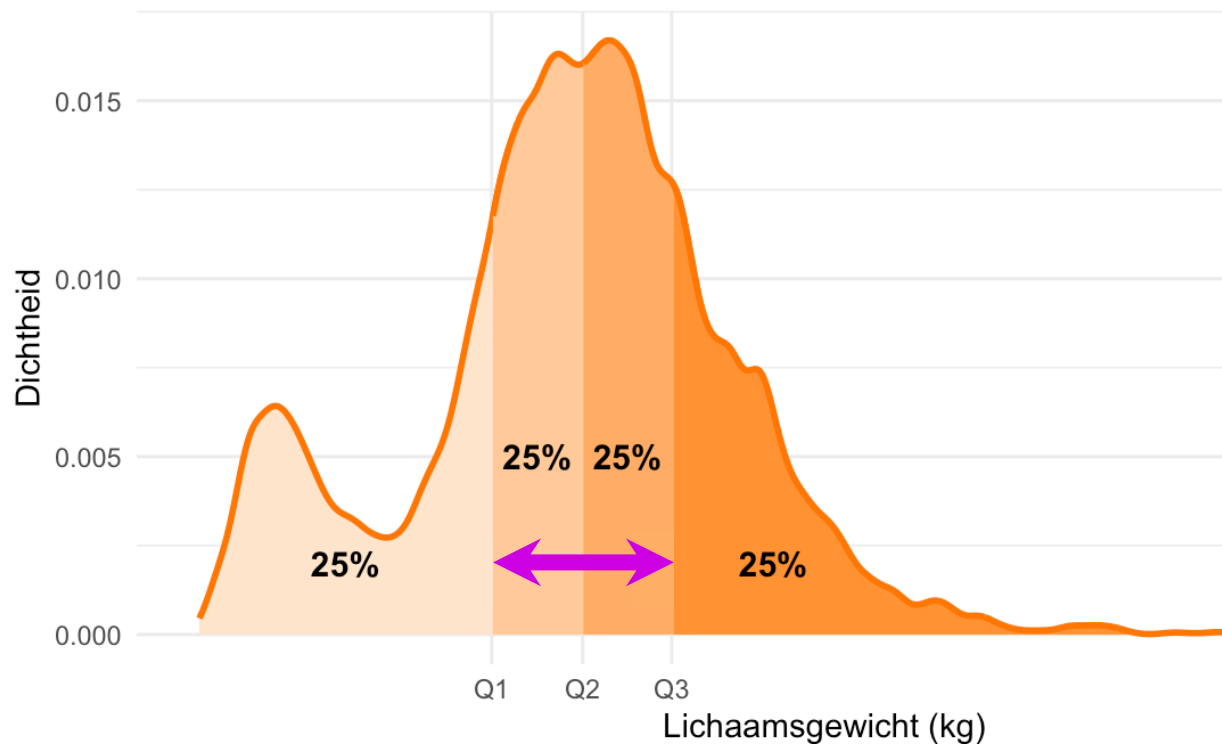
Zie de verdeling hieronder van lichaamsgewicht in de dataset NHANES. Er wordt een fout ontdekt:



bij één persoon is 170 kg ingevuld, terwijl dat **4000kg** had moeten zijn!

Wat gebeurt er met de IKA?

Gevoeligheid voor extreme waarden



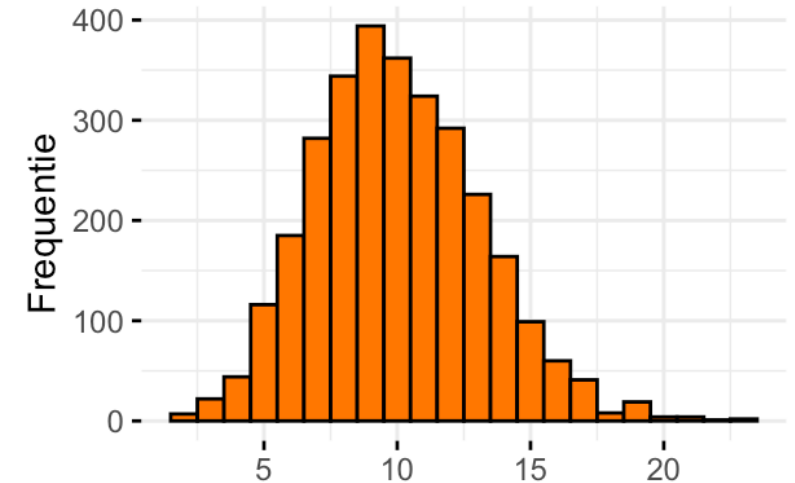
Standaarddeviatie en variantie zijn veel **gevoeliger** dan de **IKA** voor waarden die sterk afwijken van het gemiddelde.

- de staart van de verdeling
- bijzonder afwijkende punten (**uitbijters / outliers**)

Welke maten gebruiken? Richtlijn:

Situatie 1: De verdeling van gegevens is grofweg symmetrisch en er zijn geen opvallende uitbijters.

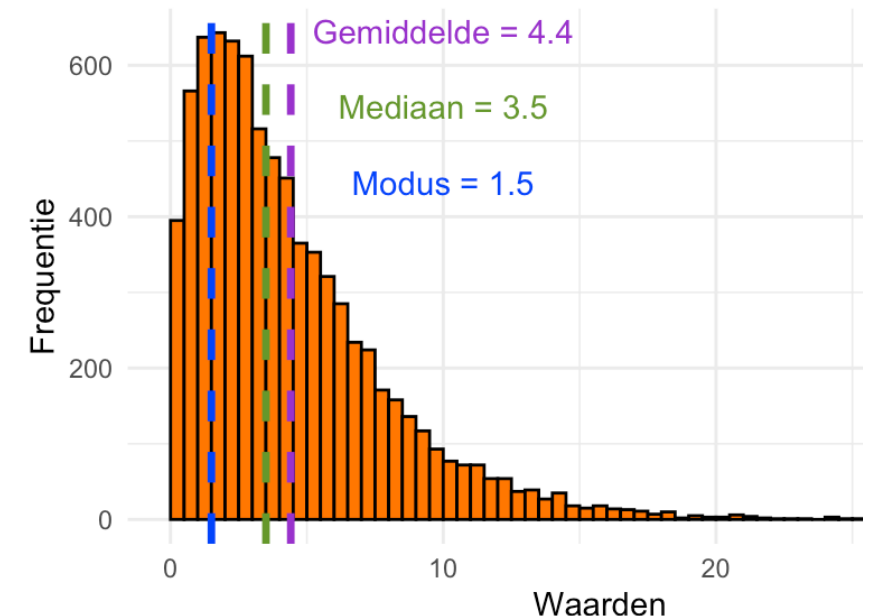
Rapporteer gemiddelde en standaarddeviatie.



Situatie 2: De verdeling van gegevens is scheef en/of er zijn opvallende uitbijters.

In dat geval kunnen het gemiddelde en de standaarddeviatie sterk beïnvloed worden door uitbijters en de staart van de verdeling.

Rapporteer mediaan (of modus/modi) en de IKA.



Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. Frequentietabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Frequentietabellen: categorische variabelen

Gezondheidsstatus	Absolute frequentie	Relatieve frequentie
Slecht	187	0.02
Matig	1010	0.13
Goed	2956	0.39
Zeer Goed	2508	0.33
Uitstekend	878	0.12

Frequentietabellen: numerieke variabelen

Klasse / bin

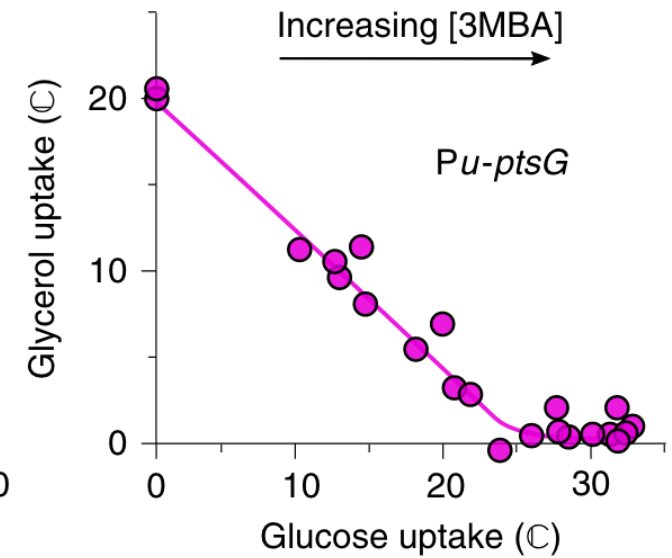
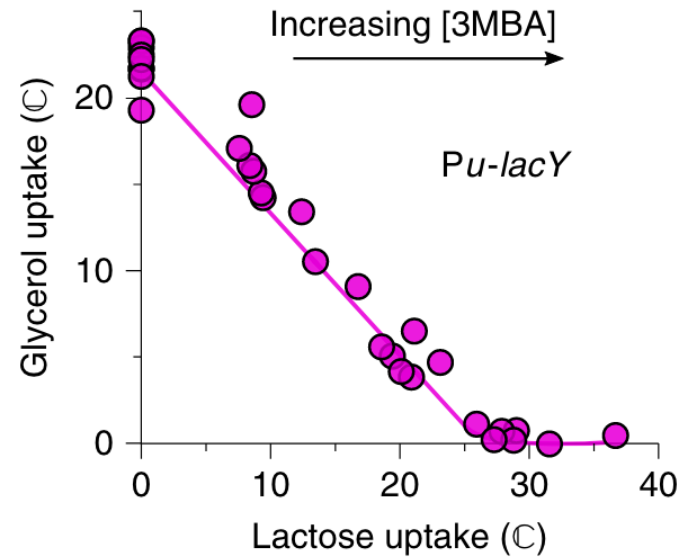
Lichaamsgewicht (kg)	Absolute frequentie
[0,10)	133
[10,20)	618
[20,30)	424
[30,40)	288
[40,50)	448
[50,60)	1068
[60,70)	1547
[70,80)	1630
[80,90)	1416
[90,100)	939
[100,110)	691

Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. Frequentietabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Waarom beschrijvende statistiek? (reprise)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
7																			
8																			
9																			
10																			
11																			
12																			
13																			
14																			
15																			
16																			
17																			
18																			
19																			
20																			
21																			
22																			
23																			
24																			
25																			
26																			
27																			
28																			
29																			
30																			
31																			
32																			
33																			
34																			
35																			
36																			
37																			
38																			
39																			
40																			
41																			
42																			
43																			
44																			
45																			
46																			
47																			
48																			
49																			
50																			
51																			
52																			
53																			
54																			
55																			
56																			
57																			
58																			
59																			
60																			
61																			
62																			
63																			
64																			
65																			
66																			
67																			
68																			
69																			
70																			
71																			
72																			
73																			
74																			
75																			
76																			
77																			
78																			
79																			
80																			
81																			
82																			
83																			
84																			
85																			
86																			
87																			
88																			
89																			
90																			
91																			
92																			
93																			
94																			
95																			
96																			
97																			
98																			
99																			
100																			



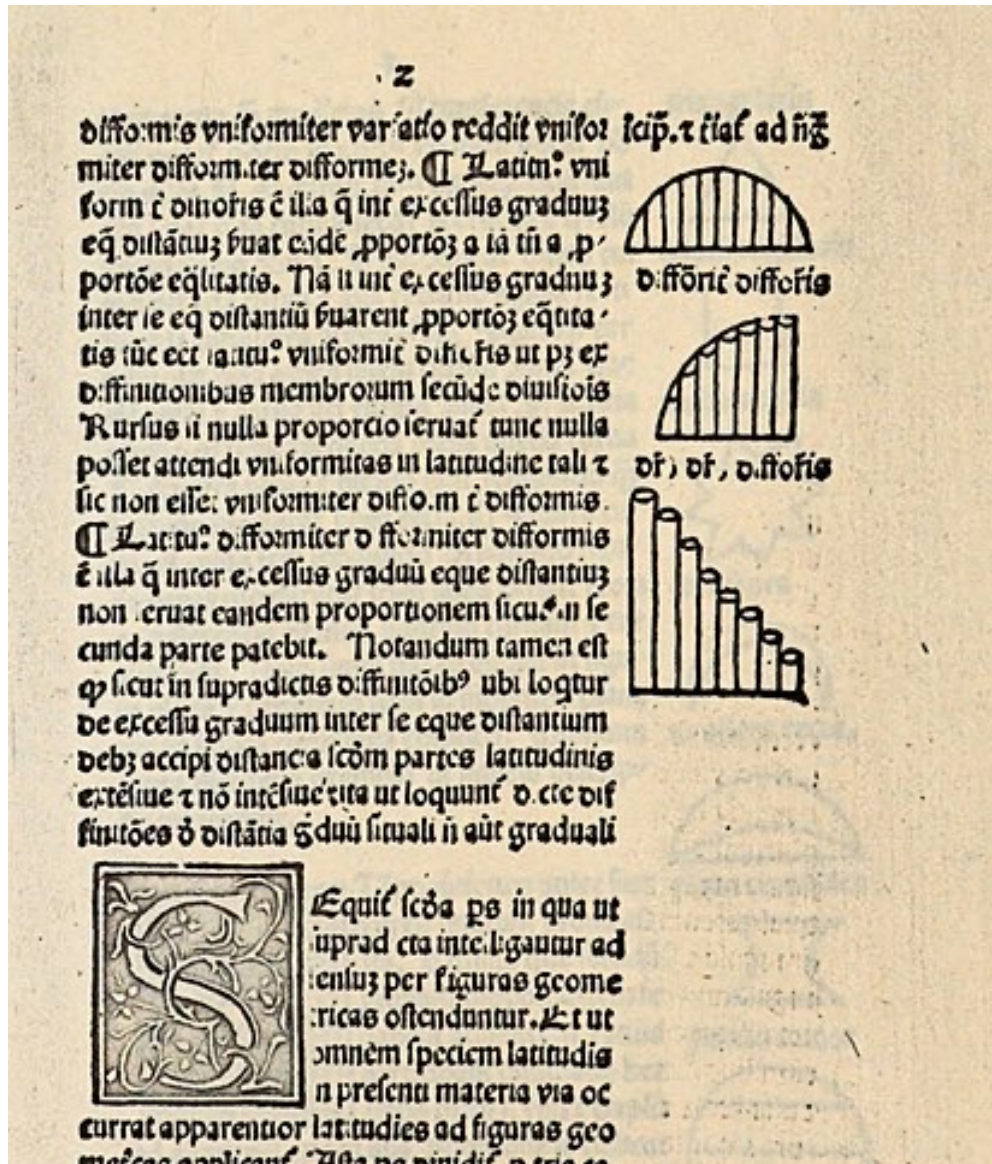
De truc: Nicole Oresme (1325-1382)



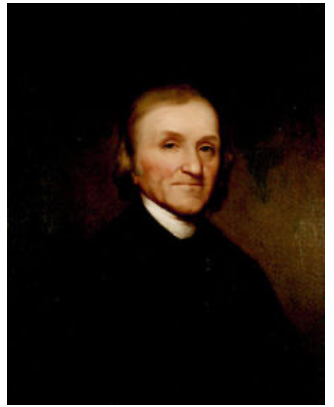
"Every measurable thing, except numbers, is imagined in the manner of continuous quantity. Therefore, for the mensuration of such a thing, it is necessary that points, lines and surfaces, or their properties be imagined. For in them, as the Philosopher has it, measure or ratio is initially found, while in other things it is recognized by similarity as they are being referred to by the intellect to the geometrical entities."

Tractatus de Configurationibus et Qualitatibus et Motuum (c. 1350)

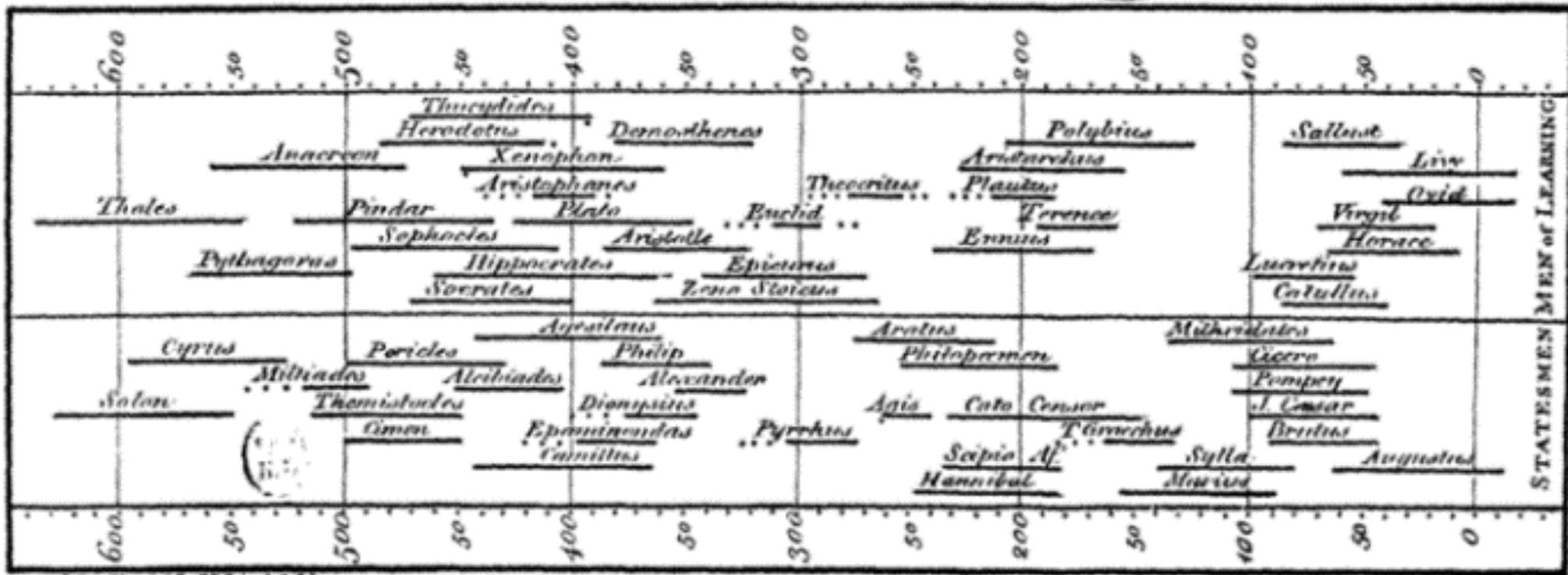
BMS



Joseph Priestley (1733-1804)



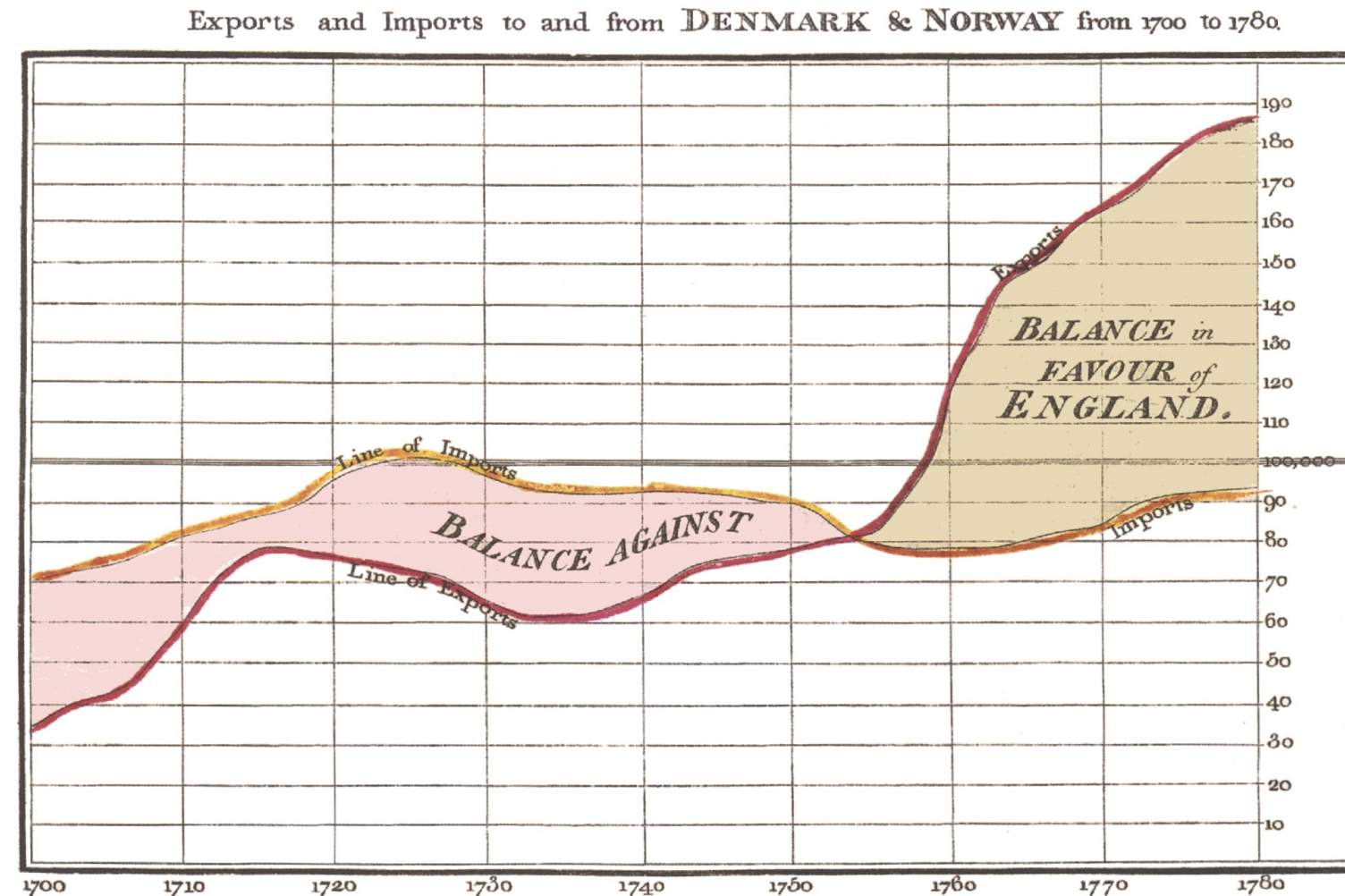
A Specimen of a Chart of Biography.



J. Priestley L. Ed. Phil. Soc. at del.



William Playfair (1759-1823)



The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 1st May 1786, by W^m Playfair

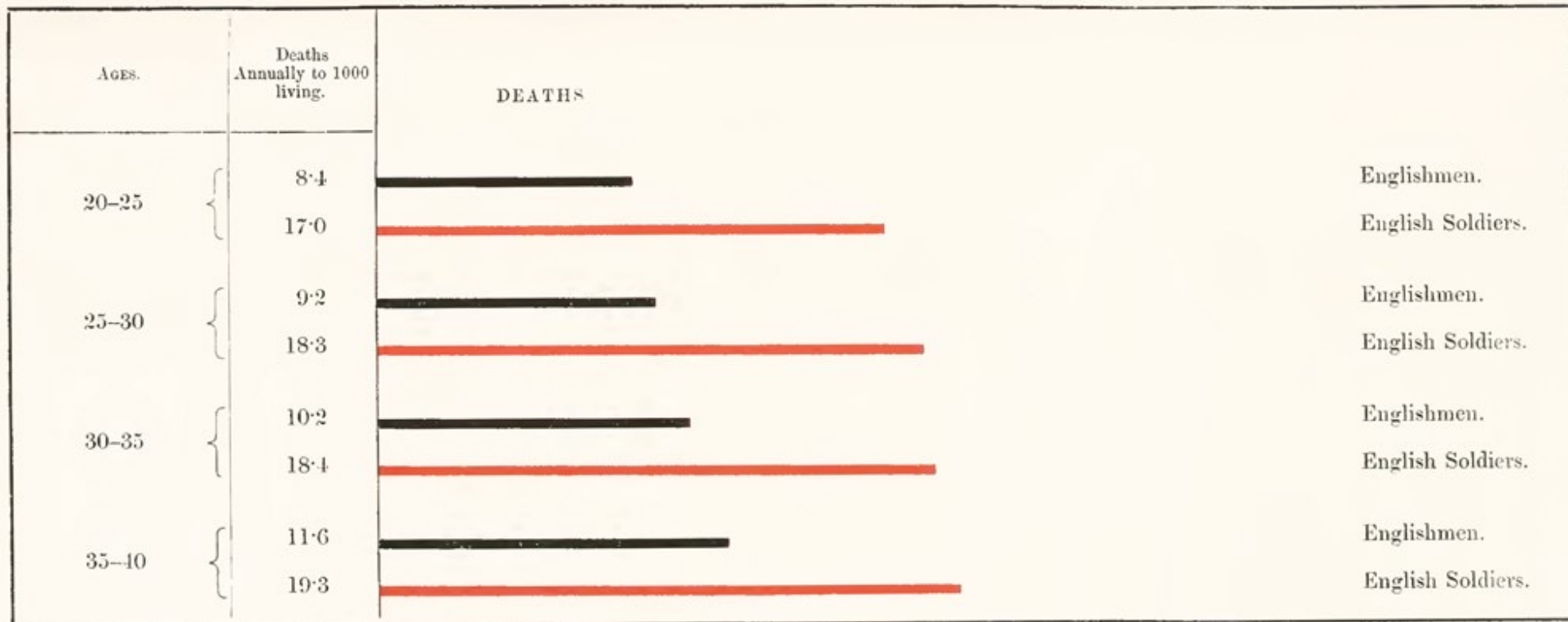
Neele sculpt 352, Strand, London.

The Commercial and Political Atlas, William Playfair (1768)
(Image: Wikimedia Commons)

Florence Nightingale (1820-1910)



Representing the Relative Mortality of the Army at Home and of the English Male Population at corresponding Ages.

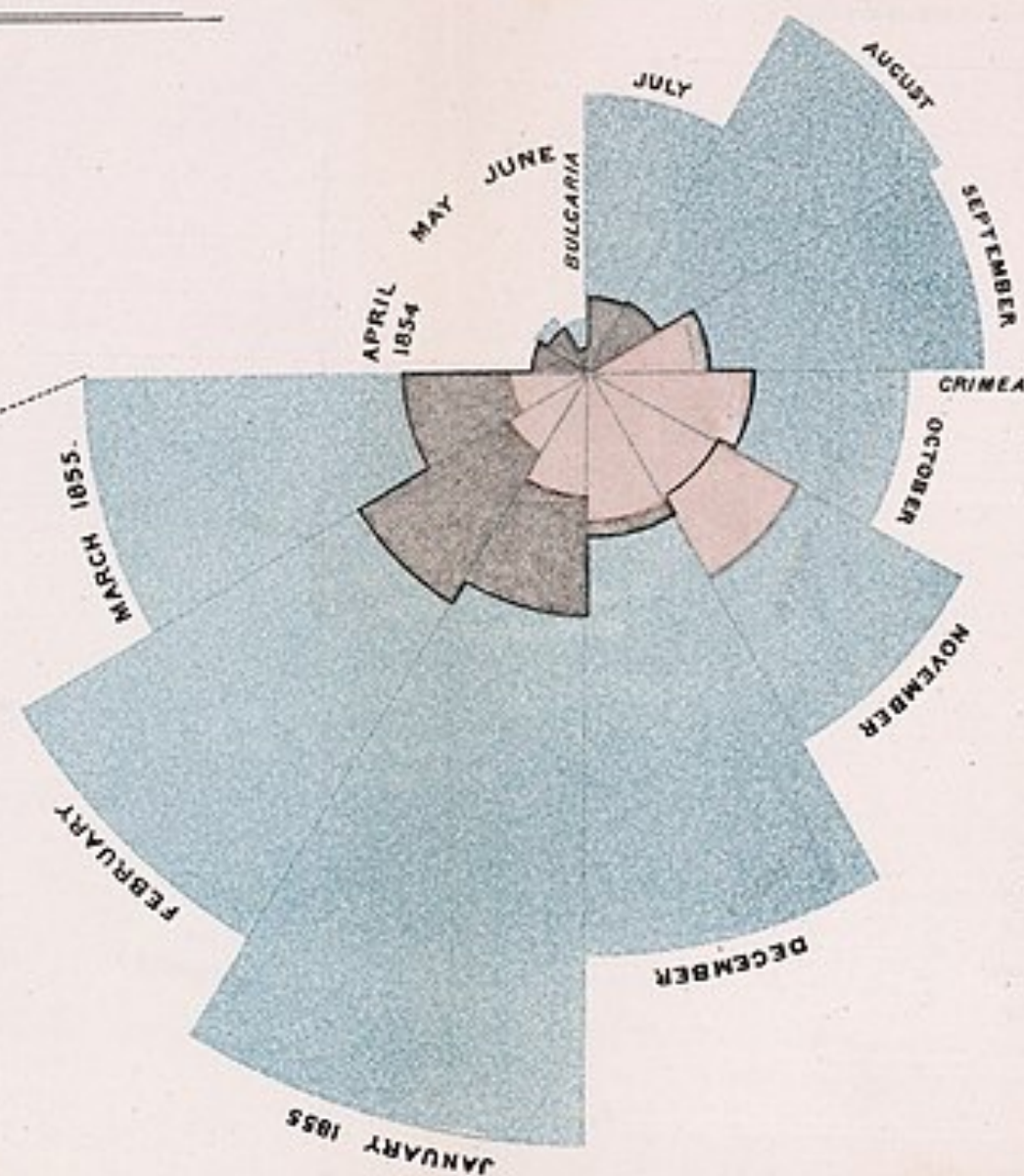
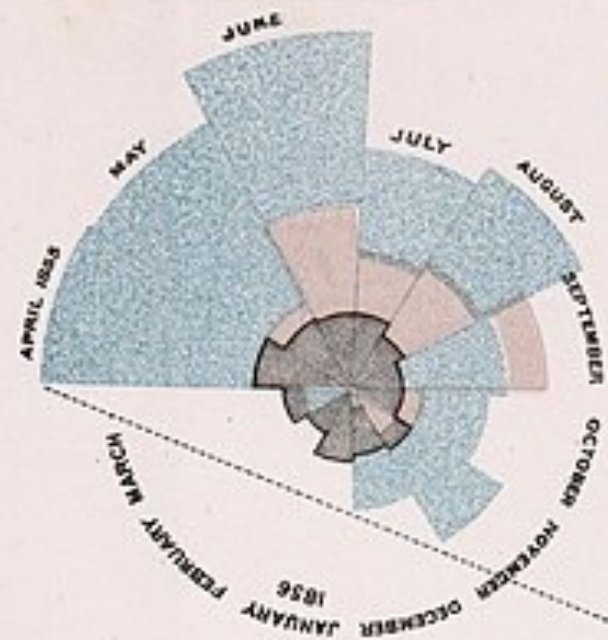


JAMES LEWIS, del.

DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

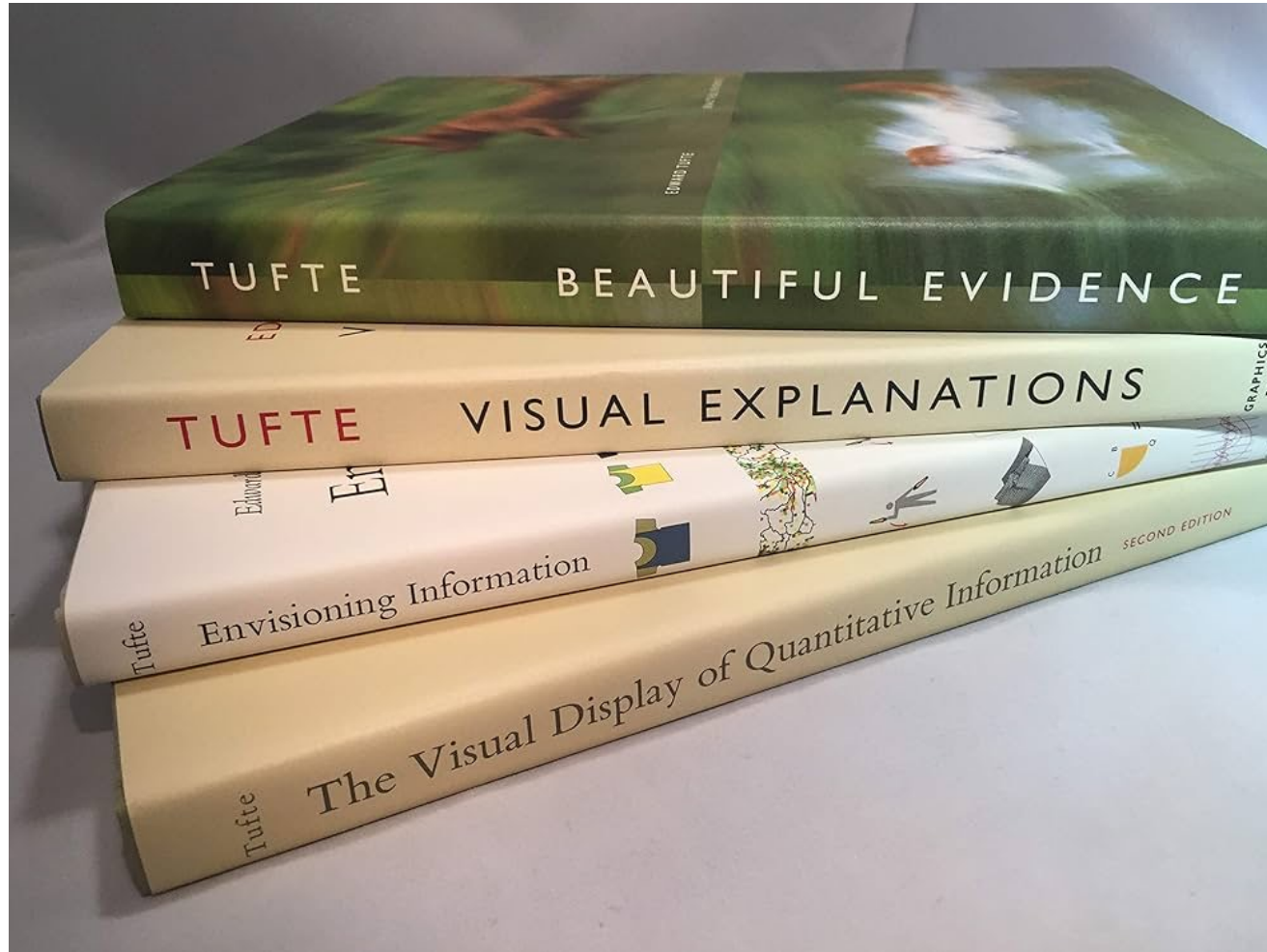
2.
APRIL 1855 to MARCH 1856.

1.
APRIL 1854 to MARCH 1855.



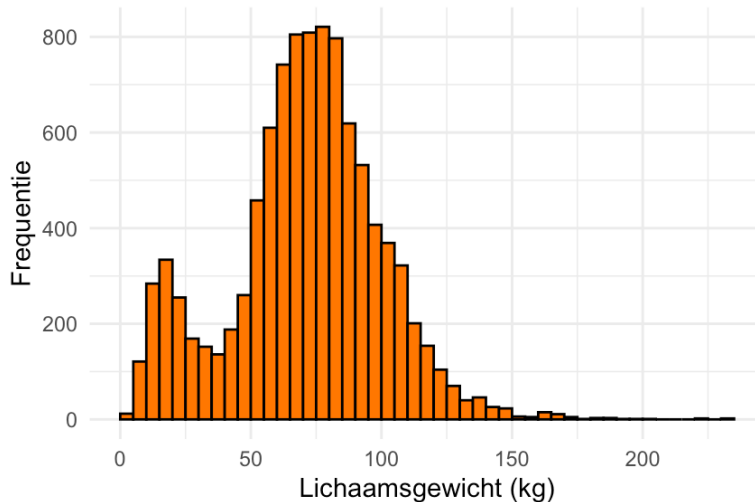
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Novr 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red;
in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Tip: boeken van Edward Tufte



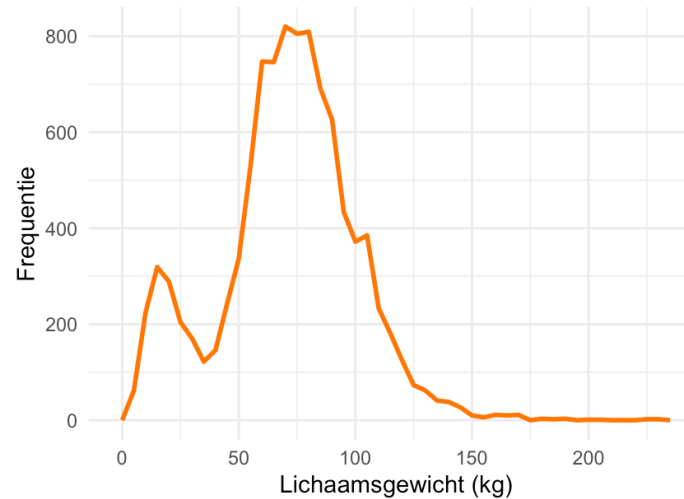
Het visualiseren van verdelingen van continue variabelen

Histogram



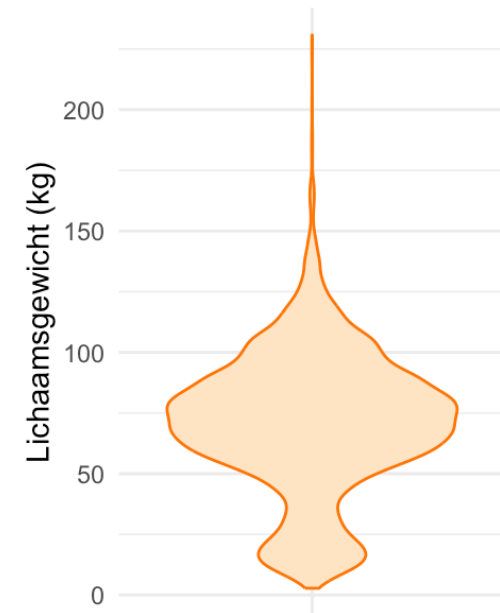
- Oppervlakte staaf representeert frequentie
- Geen ruimte tussen staven

Frequentiepolygoon



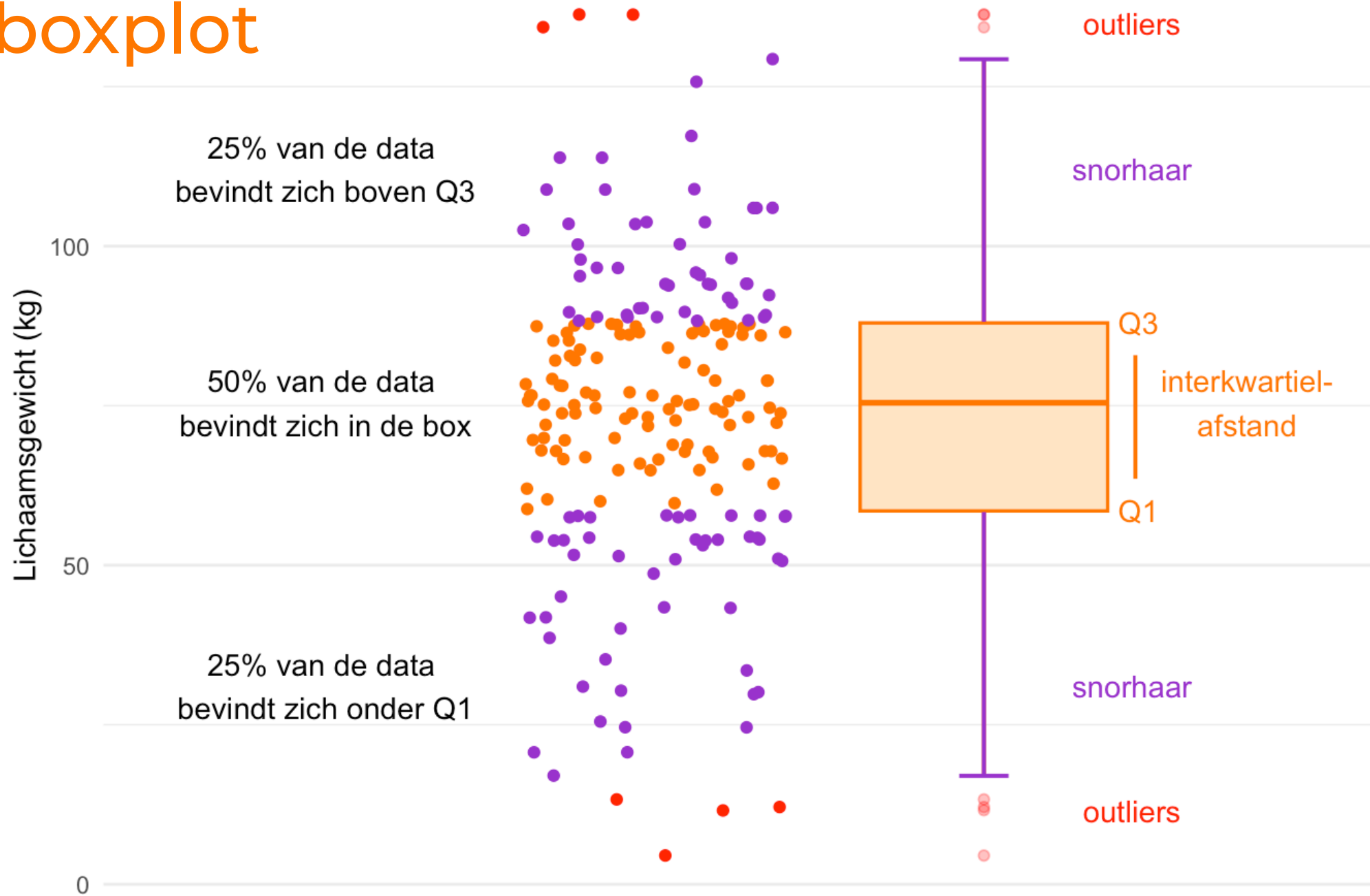
- Hoogte lijn representeert frequentie

Vioolplot

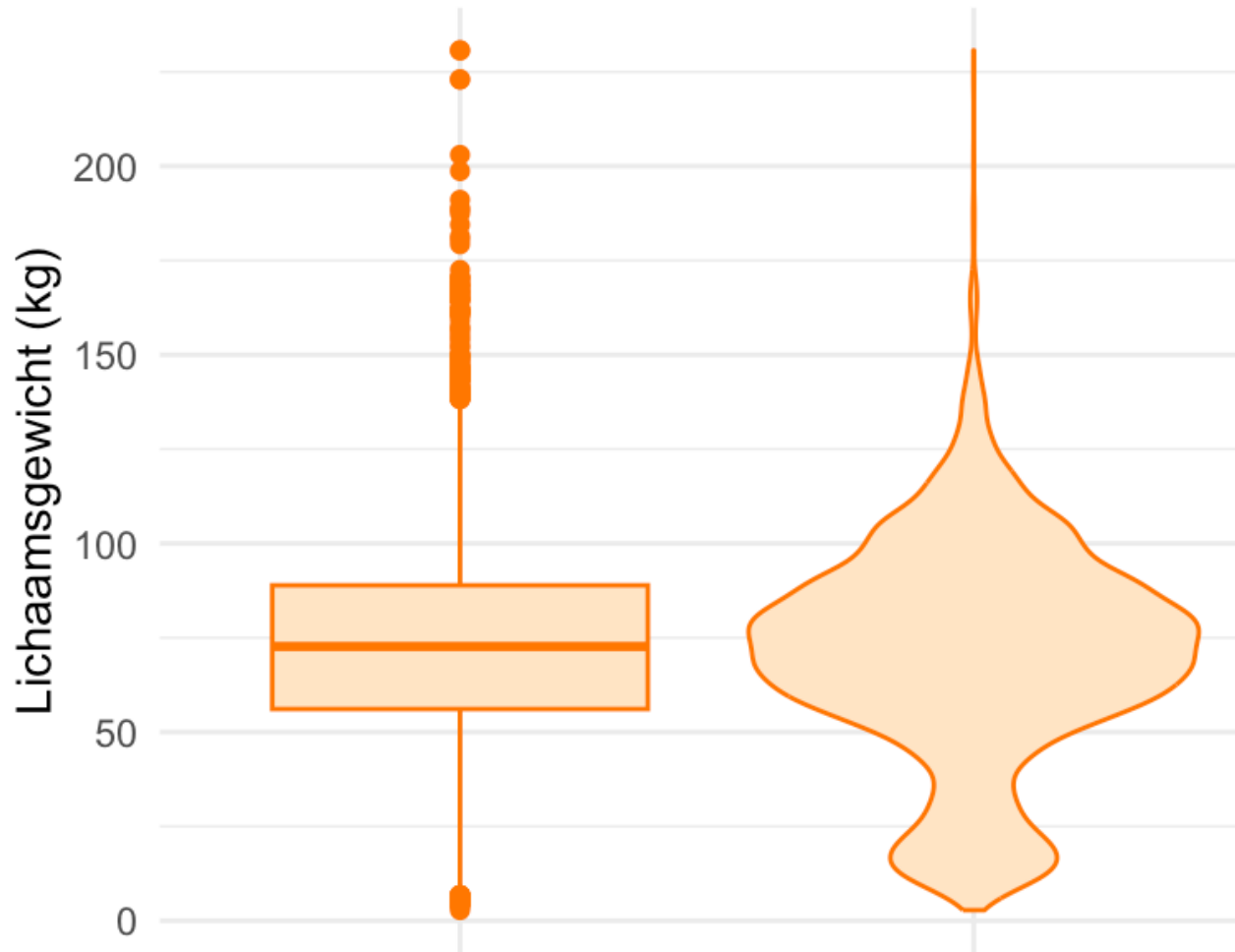


- Breedte “viool” representeert frequentie

De boxplot



Voor- en nadelen van boxplots



Voordelen:

- overzichtelijk
- scheefheid goed zichtbaar

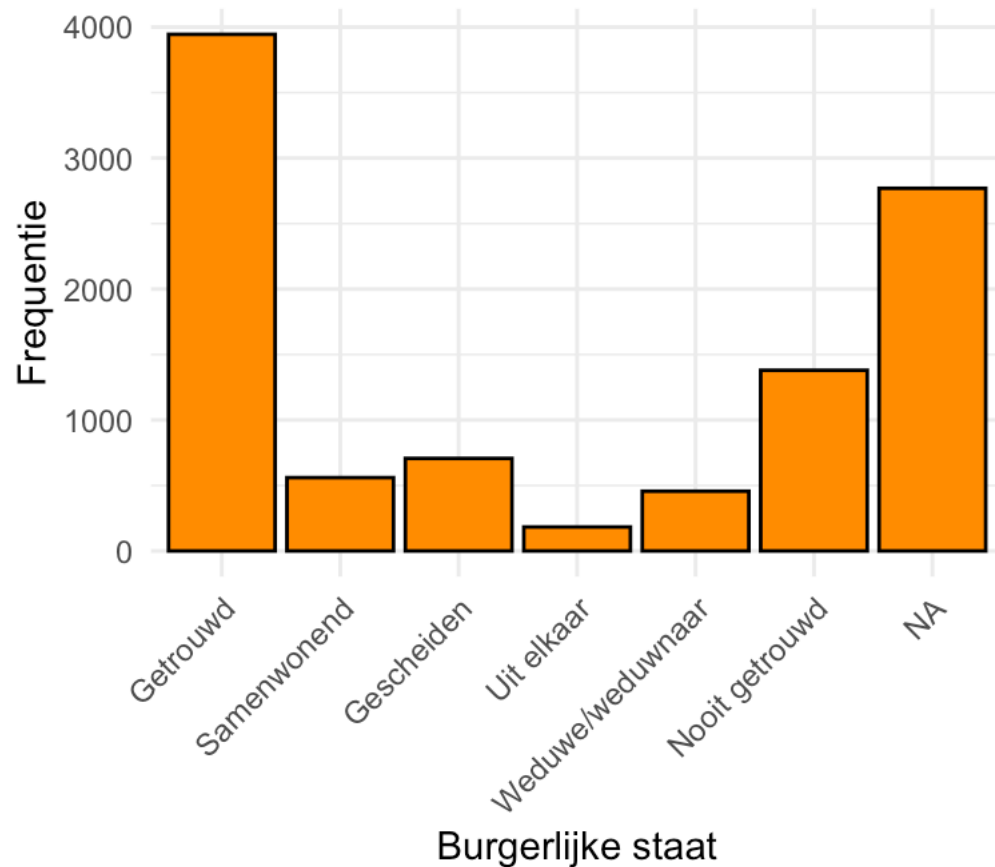
Nadeel:

- kan belangrijke eigenschappen verbergen

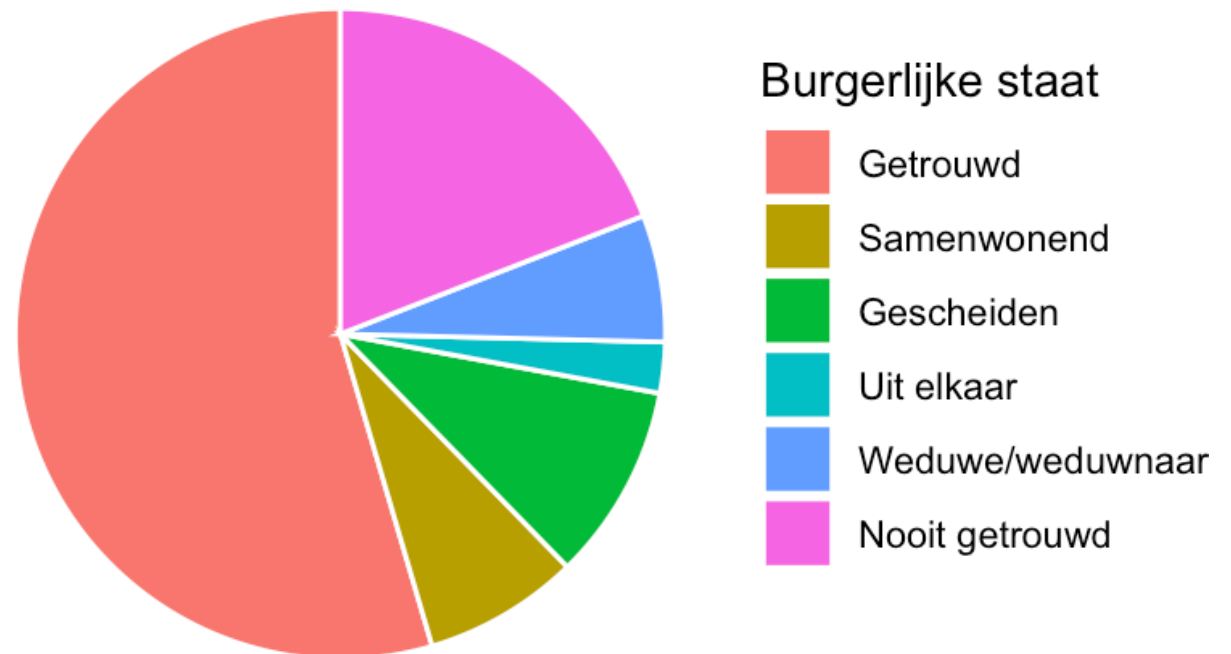
Bimodaliteit hier niet te zien!

Het visualiseren van verdelingen van nominale variabelen

Staafdiagram



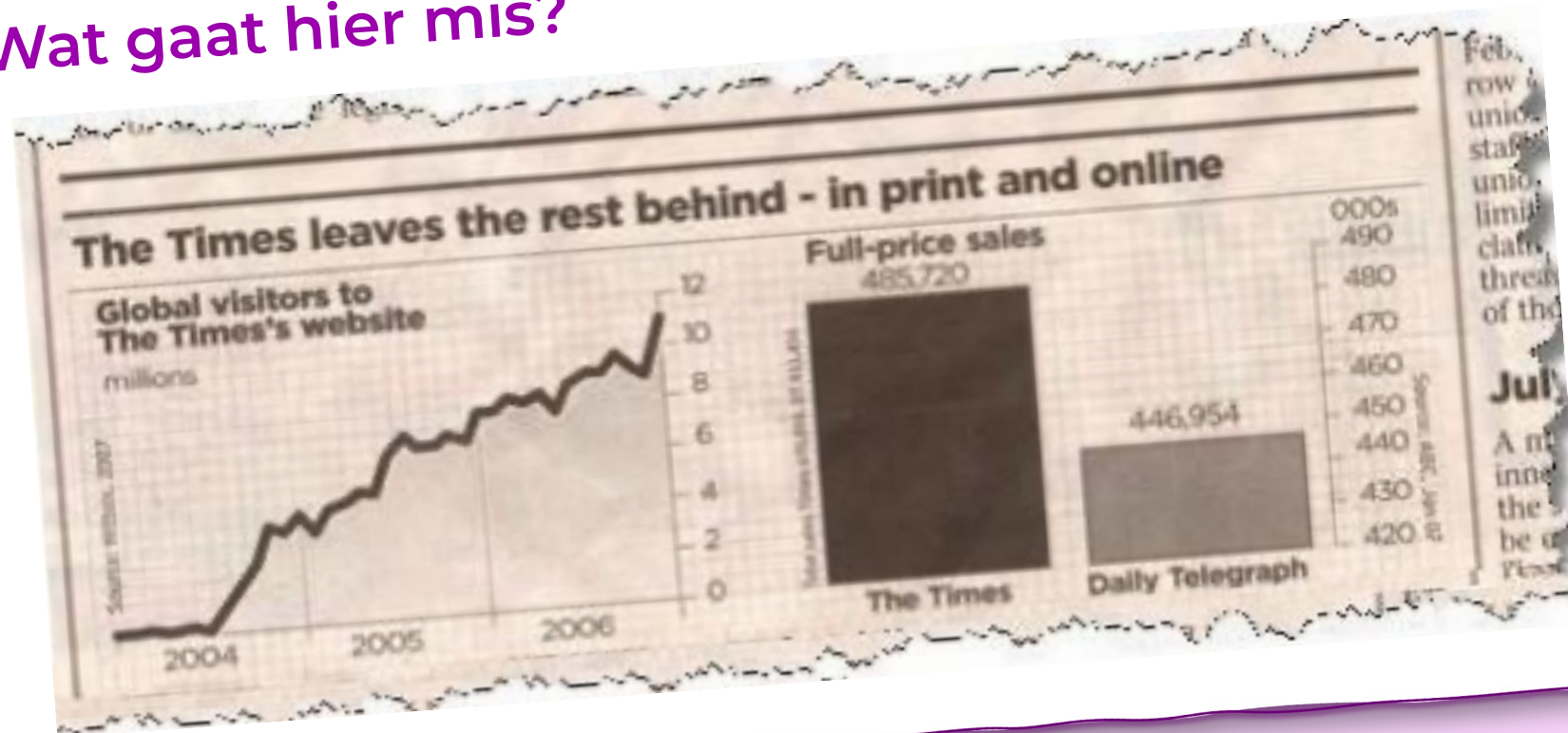
Taartdiagram



The Times versus de Daily Telegraph



Zie het staafdiagram hieronder (rechts).
Wat gaat hier mis?

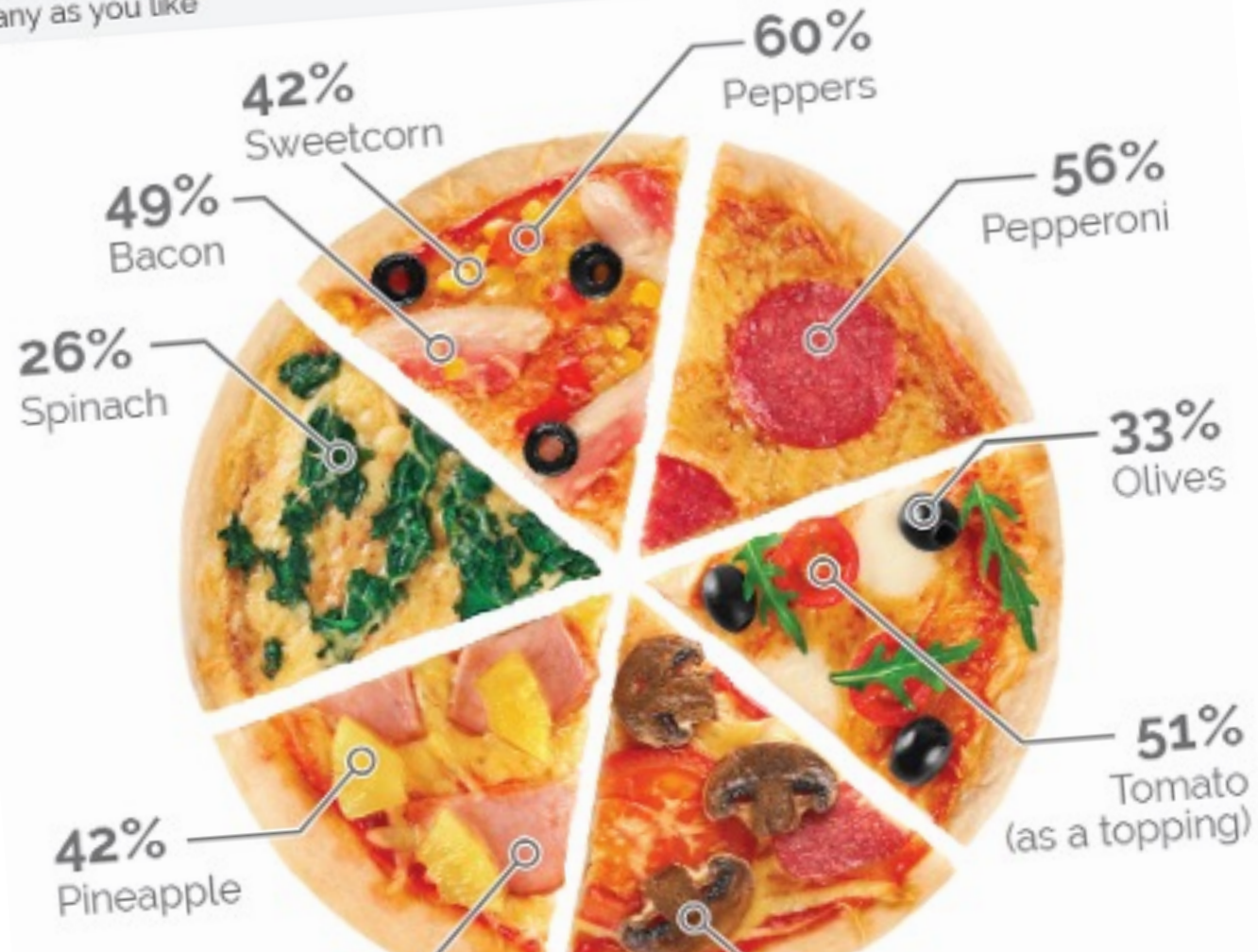


Even wachten...

Wat gaat hier mis?

Mushroom is the UK's most liked pizza topping

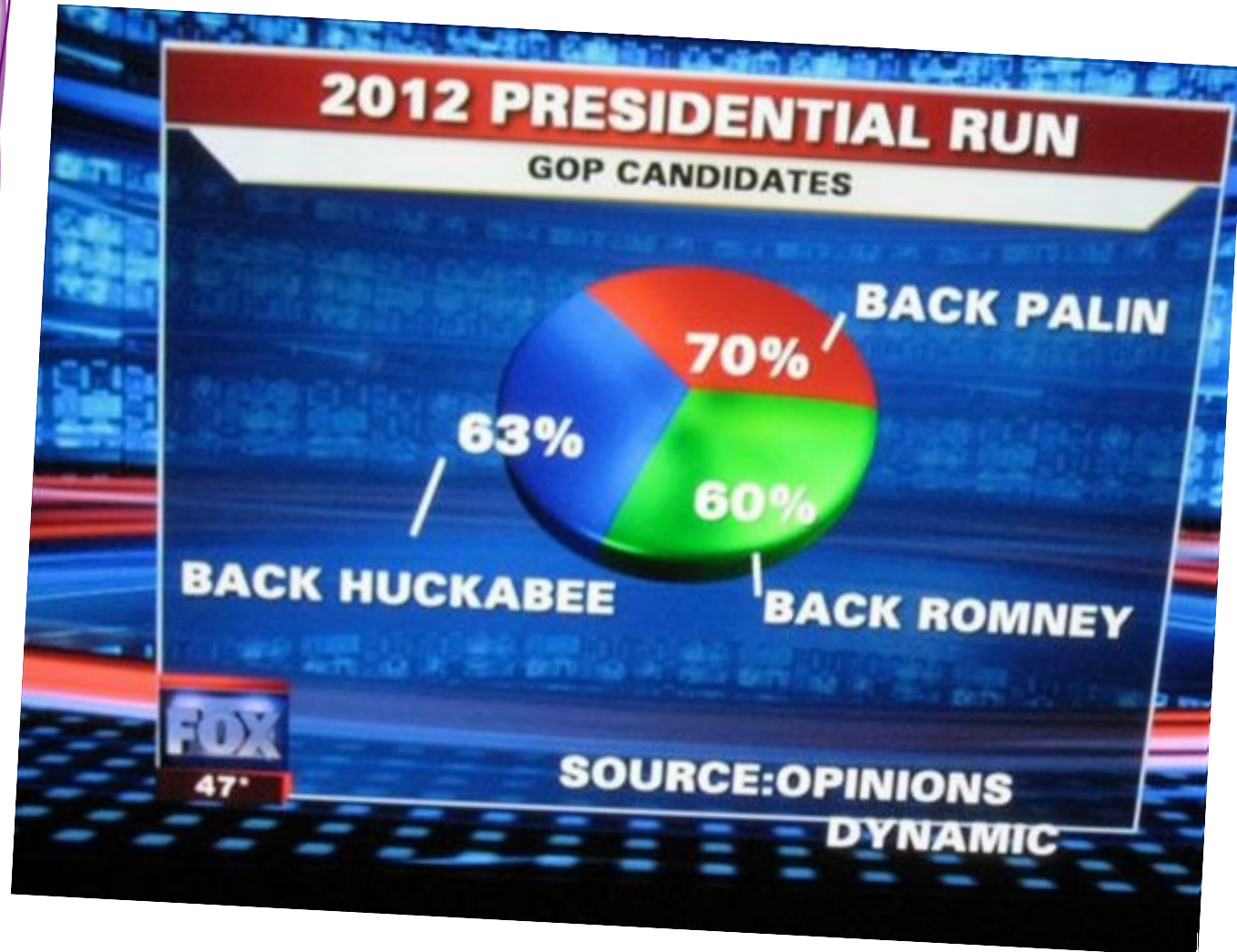
Generally speaking, which of the following toppings do you like on a pizza? Select as many as you like



Welke kandidaat ondersteun jij?

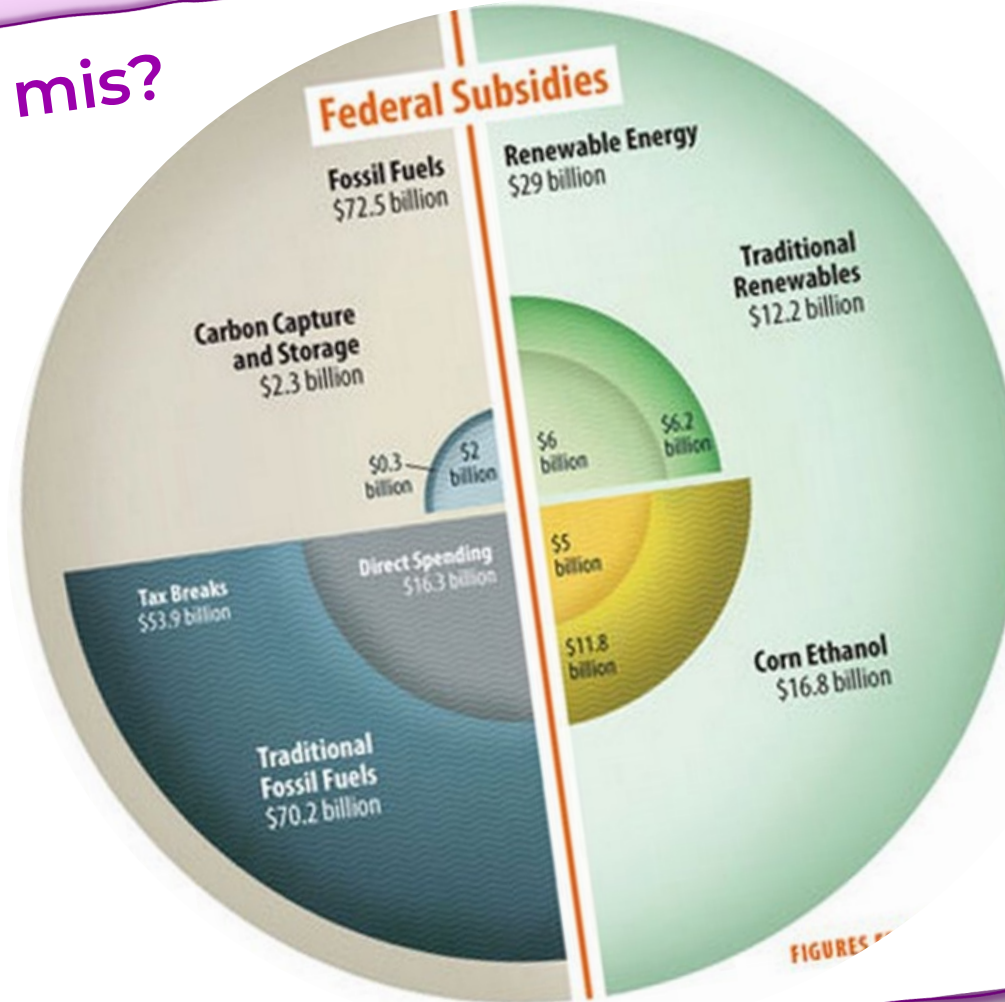


Wat gaat hier mis?



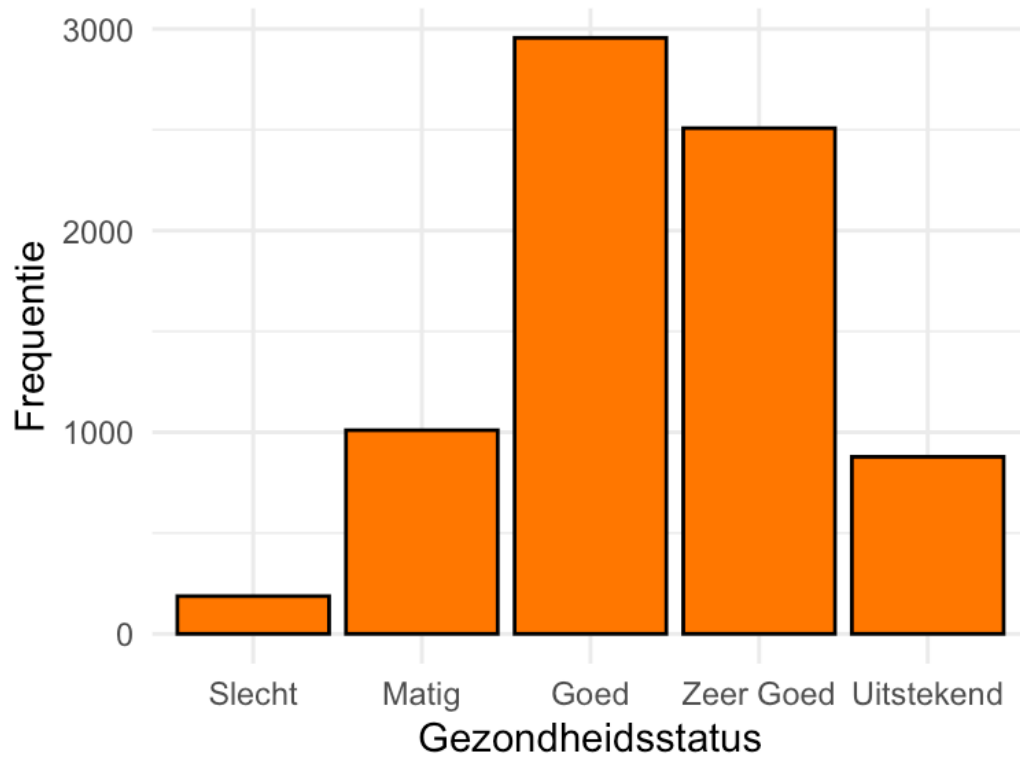
Maar wacht... wat?!?

Wat gaat hier mis?

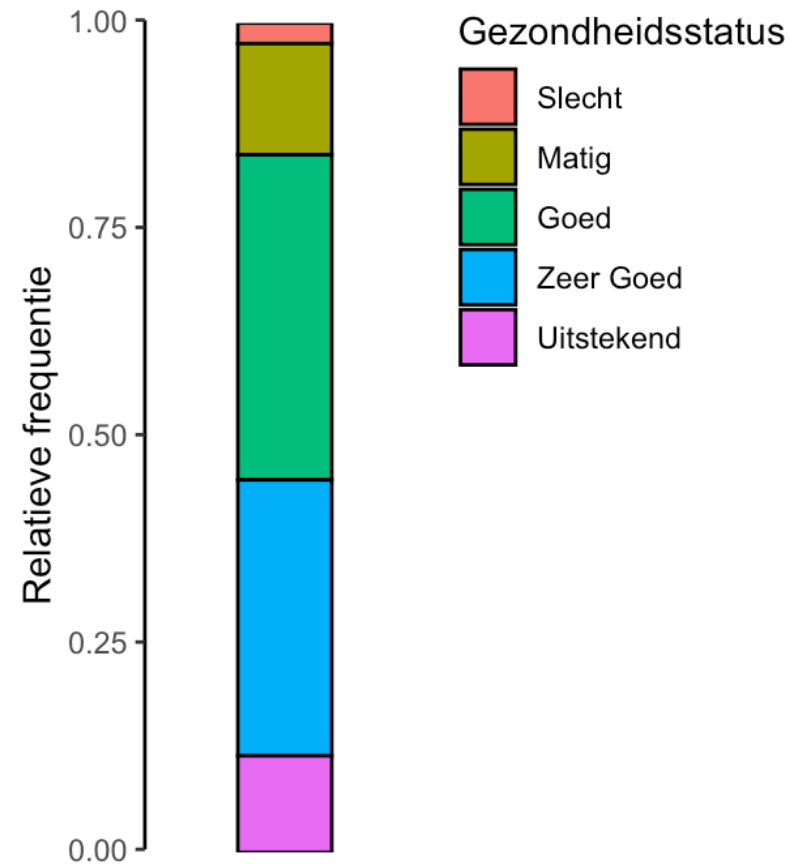


Het visualiseren van verdelingen van ordinale variabelen

Staafdiagram



Stapeldiagram



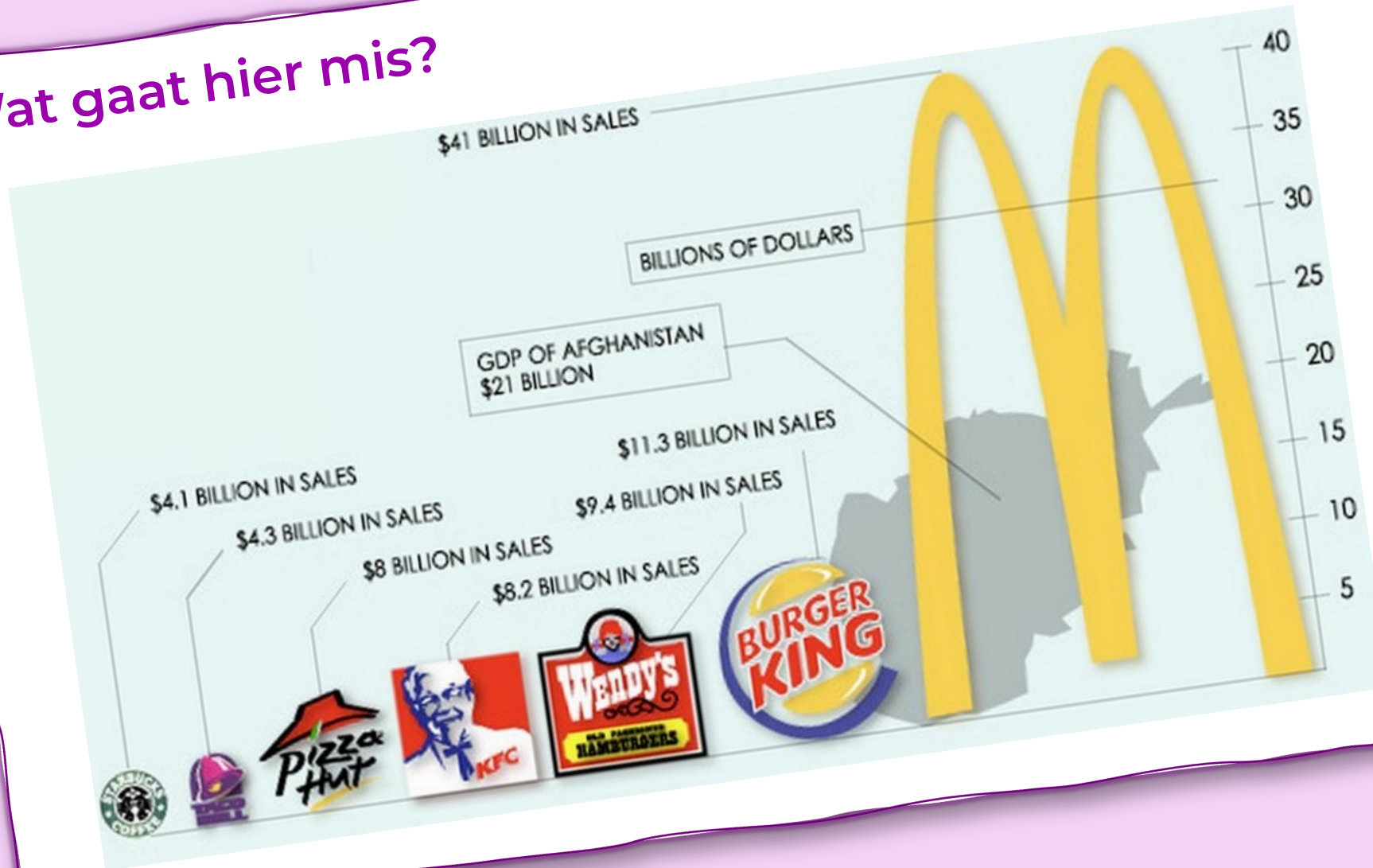
TRANSPORT

Wat gaat hier mis?



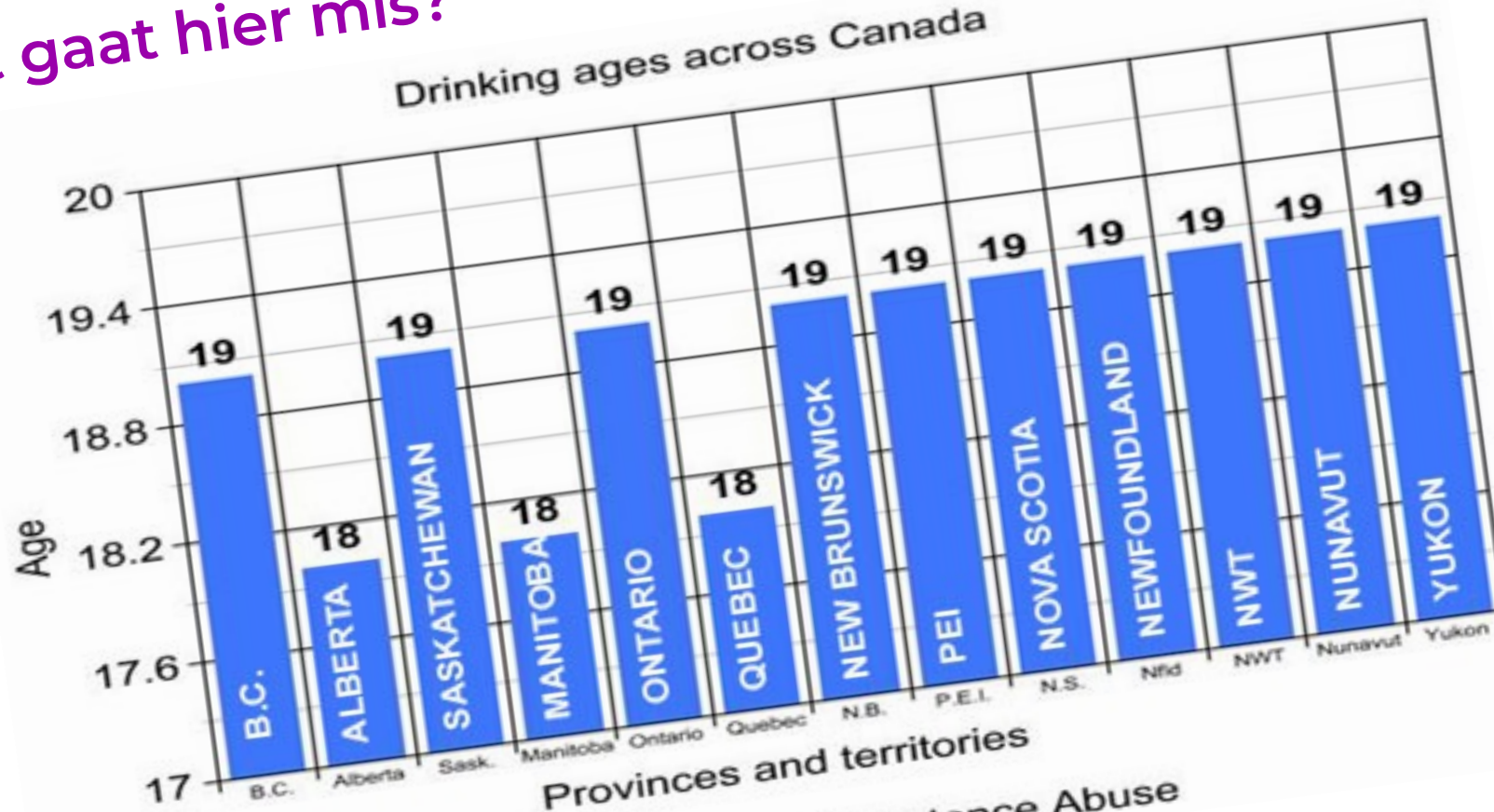
Fast food

Wat gaat hier mis?



Drinking age

Wat gaat hier mis?



Canadian Centre on Substance Abuse



De essentie in het kort:

- Grafieken representeren waarden als geometrische eigenschappen: lengtes, oppervlakten, afstanden, hoeken, ...
- Kies een representatie die aansluit bij de variabele.
- Zorg dat de representatie waarheidsgetrouw en consistent is.
- Een grafiek moet inzicht geven; laat de *data* zien!

Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. Frequentietabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Typen variabelen in R



Type variabele		Datatype in R
Numeriek	Continu	<code>numeric</code>
	Discreet	<code>numeric</code> / <code>integer</code>
Categorisch	Nominaal	<code>factor</code>
	Ordinaal	<code>factor</code>

Factoren



```
# Vector met de voedingsstrategie van 5 organismen:  
voedingsstrategie <- c("herbivoor", "carnivoor", "omnivoor", "detrivoor", "herbivoor")  
voedingsstrategie
```

```
[1] "herbivoor" "carnivoor" "omnivoor"  "detrivoor" "herbivoor"
```

```
# Omzetten naar een factor-variabele:  
factor.voedingsstrategie <- factor(voedingsstrategie)  
factor.voedingsstrategie
```

```
[1] herbivoor carnivoor omnivoor detrivoor herbivoor  
Levels: carnivoor detrivoor herbivoor omnivoor
```

```
levels(factor.voedingsstrategie)
```

```
[1] "carnivoor" "detrivoor" "herbivoor" "omnivoor"
```

Statistieken



```
mean(v)      # gemiddelde  
median(v)    # mediaan  
summary(v)   # verschillende statistieken voor vector v
```

```
range(v) # bereik  
sd(v)   # standaarddeviatie  
var(v)  # variantie  
IQR(v)  # interkwartielafstand (IKA)
```

```
quantile(v, prob = 0.25)
```

Frequentietabellen en figuren



```
table(NHANES$MaritalStatus)
```

Divorced	LivePartner	Married	NeverMarried	Separated	Widowed
707	560	3945	1380	183	456

Frequentietabellen en figuren



```
x <- c(1, 2, 3)
y <- c(2, 1, 5)
plot(x, y) # plot alleen punten
```

```
plot(x, y, type = "l") # plot alleen lijnen
```

```
plot(x, y, type = "b") # plot punten én lijnen
```

```
hist( NHANES$Age, breaks = 20 ) # histogram van leeftijden in NHANES
```

```
# Maak een frequentietabel
tabel <- table(NHANES$HealthGen)

# Teken het staafdiagram
barplot(tabel)
```

Inhoud

1. Waarom beschrijvende statistiek?
2. Populaties, eenheden, variabelen, en verdelingen
3. Kengetallen / statistieken
4. Frequentietabellen
5. Het visualiseren van gegevens
6. Beschrijvende statistiek in **R**
7. Wrap-up

Alles bij elkaar: een voorbeeld

“In de dataset NHANES heeft de variabele lichaamsgewicht een breed bereik van 2,7 tot 230,7 kg.

De verdeling is bimodaal (zie Fig. 1), met een modus bij 20 kg en een modus bij 76 kg. De interkwartielafstand is 32,8 kg. “

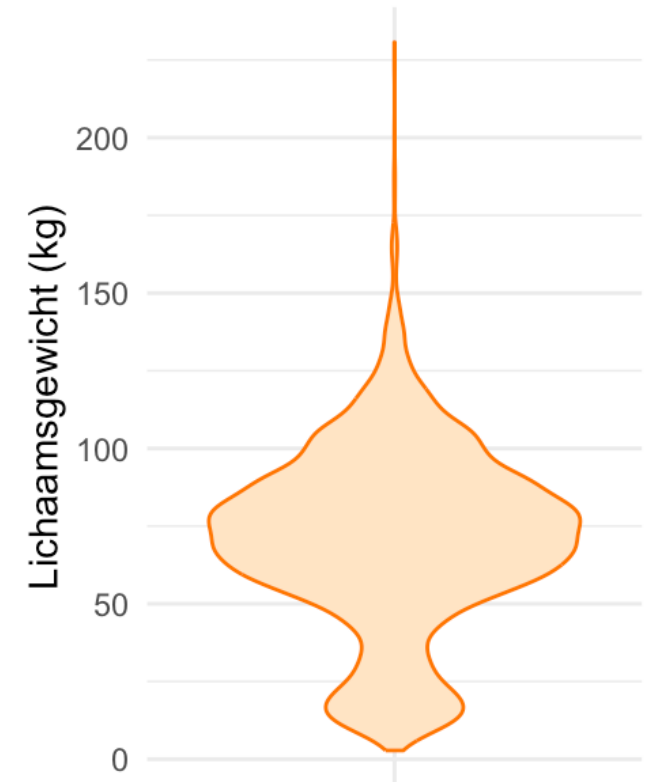


Fig. 1: Vioolplot van lichaamsgewicht. (Data: NHANES.)