

Samenvatten van data

1.1 Inleiding

De term 'data' wordt overvloedig gebruikt bij wetenschappelijk onderzoek. Data zijn resultaten van een onderzoek. In het algemeen bestaat een verzameling data uit eigenschappen of kenmerken van individuen of experimentele eenheden. Vaak worden de data weergegeven in één of meer getallen.

1.2 Nominale, ordinale, interval-schaal

Voorbeeld 1.2.1 Van een groep studenten is gegeven

- de kleur van de ogen (donkerbruin, grijs, blauw, lichtbruin of groen)
- hun politieke activiteit (zeer gering, gering, gemiddeld, groot, zeer groot)
- de lengte (in cm.).

We zouden de volgende code kunnen hanteren: donkerbruin = 0, grijs = 1, blauw = 2, lichtbruin = 3, groen = 4, zeer gering = 1, gering = 2, gemiddeld = 3, groot = 4, zeer groot = 5. Dan is (2,4,169) een student met blauwe ogen, die grote politieke activiteit vertoont en die 169 cm. lang is. Het is niet noodzakelijk een codering met getallen te hanteren. We kunnen ook voor de hand liggende namen gebruiken, bijv. (blauw, groot, 169) i.p.v. (2, 4, 169). \square

De in voorbeeld 1.2.1 gehanteerde meetschalen zijn alle drie van een ander type. De kleur van de ogen is een voorbeeld van meting m.b.v. een **nominale schaal**. Bij gebruik van een dergelijke schaal worden de individuen in niet nader geordende categorieën ingedeeld. Aan elke categorie wordt een getal of een naam toegekend zodat verschillende categorieën geïdentificeerd kunnen worden door verschillende getallen of namen. Merk op dat rekenkundige bewerkingen op dit soort data van geen betekenis zijn. Hebben we bijvoorbeeld als gegevens de drietallen (2,4,169), (1,5,173), (4,2,170) dan heeft het geen zin het steekproefgemiddelde van de eerste component (dat is $\frac{2+1+4}{3}$) te bepalen, omdat dit geen enkele betekenis heeft. We hadden immers net zo goed kunnen afspreken donkerbruin = 100, grijs = 20 enz.

Een grootheid die wel betekenis heeft voor data op een nominale schaal is de **steekproefmodus**. De steekproefmodus is de waarde die in de steekproef het meest voorkomt. Zijn er verschillende waarden in een steekproef die allemaal in aanmerking komen voor de meest voorkomende waarde, dan is ieder van deze waarden een steekproefmodus.

Voorbeeld 1.2.2 Bekijk de volgende gegevensverzameling van gegevens uit voorbeeld 1.2.1.

(4,5,168)	(2,2,187)	(0,4,173)
(3,1,178)	(1,4,172)	(4,3,171)
(2,4,169)	(2,1,165)	(2,1,168)
(0,2,182)	(0,5,167)	(2,3,170)
(3,1,175)	(1,3,162)	(0,3,169)

De steekproefmodus van de ogenkleur is 2.

De steekproefmodi van de politieke activiteit zijn 1 en 3.

De steekproefmodi van de lengte zijn: 168 en 169. □

De politieke activiteit uit voorbeeld 1.2.1 is een voorbeeld van meting m.b.v. een **ordinaire** schaal. De categorieën worden hier niet alleen geïdentificeerd, er is ook sprake van een ordening. De onderlinge afstanden hebben echter geen betekenis: de afstand tussen zeer gering (1) en gering (2) is niet gelijk aan de afstand tussen gemiddeld (3) en groot (4). Er is uitsluitend een ordening gegeven. Omdat we net zo goed hadden kunnen afspreken zeer gering = 2, gering = 5, gemiddeld = 8, groot = 10, zeer groot = 15 heeft het steekproefgemiddelde opnieuw geen betekenis. Wel heeft de steekproefmodus betekenis en ook de **steekproefmediaan**.

De steekproefmediaan is de middelste waarneming, wanneer we de waarnemingen van beneden naar boven geordend hebben (bij even aantal waarnemingen: het gemiddelde van de middelste twee). Noteren we de waarnemingen met X_1, X_2, \dots, X_n , en de **geordende waarnemingen** met $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, dan is de steekproefmediaan

$$M_n = \begin{cases} X_{(\frac{n+1}{2})} & \text{als } n \text{ oneven} \\ \frac{1}{2} \{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\} & \text{als } n \text{ even.} \end{cases}$$

Bij waarnemingen op een **ordinaire** schaal is het nemen van het gemiddelde van de middelste twee waarnemingen bij even n eigenlijk niet altijd zinvol. Zijn de middelste twee gelijk dan is er geen probleem. Men vermijdt dit probleem wel eens door bij even n voor de kleinste of de grootste van de middelste twee te kiezen.

Voorbeeld 1.2.3 Kijk naar de tweede component van de gegevens uit voorbeeld 1.2.2: 5, 2, 4, 1, 4, 3, 4, 1, 1, 2, 5, 3, 1, 3, 3. We ordenen ze: 1, 1, 1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5. De middelste, dat is de achtste in deze rij, is 3. De steekproefmediaan van de politieke activiteit is 3. Dat betekent dat er tenminste 8 studenten zijn met politieke activiteit ≤ 3 en tenminste 8 met politieke activiteit ≥ 3 . De steekproefmediaan van de lengte bedraagt 170 cm. Het heeft nauwelijks zin de steekproefmediaan van de ogenkleur te bepalen omdat ordenen hier geen betekenis heeft. \square

Het meten van de lengte tenslotte is een voorbeeld van meting m.b.v. een **interval-schaal**. Hierbij hebben de metingen meer betekenis dan alleen het aangeven van een volgorde. Verschillen zijn nu zinvol. Nu heeft het ook zin, naast steekproefmediaan en steekproefmodus, het steekproefgemiddelde te bepalen. Dit is voor de gegevens uit voorbeeld 1.2.1 gelijk aan 171.7 cm. Veelal zullen we te maken hebben met data gemeten m.b.v. een interval-schaal.

1.3 Histogram

Als we data ter beschikking hebben, is het eerste dat we doen enkele **samenvattingen** van deze data maken om een eerste indruk van de gegevens te verkrijgen. Bovendien geven zij ons een idee in welke richting we de analyse voort zullen zetten. Dit in kaart brengen van de gegevens kan ruwweg op twee manieren gebeuren: met **plaatjes** en met **getallen**.

De **klassieke** manieren om data samen te vatten zijn: histogram (plaatje), en de numerieke samenvatting met steekproefgemiddelde, steekproefvariantie en nog enkele grootheden. Een voorbeeld van een histogram is te zien in voorbeeld 1.3.1.

Voorbeeld 1.3.1 Van 39 Peruviaanse indianen is de polsslag gemeten. De gegevens zijn:

60 76 56 74 92 76 68 64 60 76 52 60 68
 60 76 64 72 88 64 72 72 68 72 80 64 64
 80 64 72 64 72 88 60 60 88 72 72 84 68.

We delen de waarnemingen in klassen in. Er zijn verschillende richtlijnen voor het aantal klassen om een behoorlijk plaatje te krijgen. Wij hanteren de regel: aantal klassen **niet meer** dan $10 \log n$, waarin n = aantal waarnemingen. In dit dictaat betekent \log de logaritme bij grondtal 10 en \ln de (natuurlijke) logaritme bij grondtal e . Voor $n = 39$ krijgen we als bovengrens $10 \log 39 = 15.9$. We nemen 15 klassen.

We ordenen de gegevens:

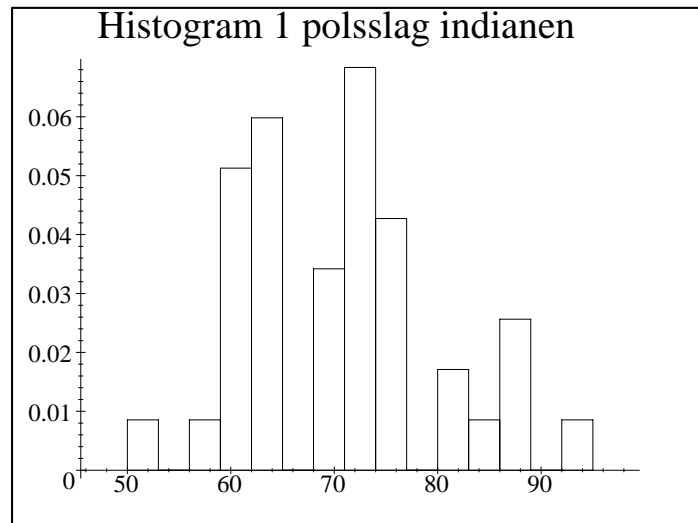
52 56 60 60 60 60 60 60 64 64 64 64 64
 64 64 68 68 68 68 72 72 72 72 72 72 72
 72 74 76 76 76 76 80 80 84 88 88 88 92.

I.4

We kiezen als klasse-indeling:

[50, 53), [53, 56), [56, 59), [59, 62), [62, 65), [65, 68),
[68, 71), [71, 74), [74, 77), [77, 80), [80, 83), [83, 86),
[86, 89), [89, 92), [92, 95).

Uit Statistiek & kansrekening weten we: hoogte \times klassebreedte = frequentiequotiënt, dus bijv. voor de klasse [50, 53): hoogte \times 3 = $1/39$ en dus hoogte = 0.0085. Deze regel zorgt er voor dat de totale oppervlakte gelijk wordt aan 1, nl. de som van de frequentiequotiënten. Immers, het histogram wordt gezien als een “schatting” van de kansdichtheid, die ook oppervlakte 1 heeft.



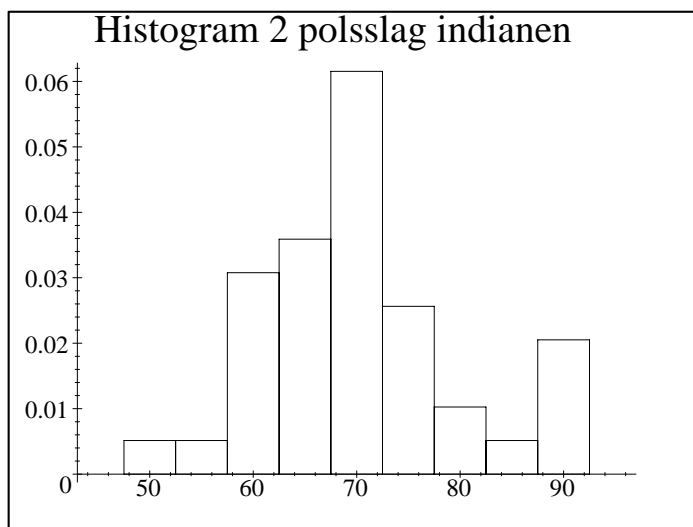
□

We herhalen nog even de regel voor het aantal klassen bij het histogram.

aantal klassen bij het histogram ten hoogste $10 \log n$

Een iets andere indeling in klassen doet het histogram er weer wat anders uitzien. Dit moet men zich wel realiseren bij het trekken van conclusies.

Voorbeeld 1.3.2 (vervolg van voorbeeld 1.3.1) We kiezen de klasse-indeling zó dat 50, 55, . . . , 85, 90 de middens van de intervallen zijn. Het resultaat is als volgt.



Computerpakketten geven een histogram soms in iets andere vorm dan hierboven. Een voorbeeld op grond van bovenstaande gegevens volgt hier:

midden van het interval	aantal waarnemingen
50	1 *
55	1 *
60	6 *****
65	7 *****
70	12 *****
75	5 *****
80	2 **
85	1 *
90	4 ****

Als je de bladzijde een kwartslag draait krijg je een soortgelijk plaatje als in “Histogram 2 polsslag indianen”. □

Terwijl de lijst van 39 getallen in voorbeeld 1.3.1 nauwelijks een idee geeft over de data, zien we met de histogrammen al heel wat. In het algemeen lezen we uit een histogram af:

- of de verzameling data al dan niet redelijk **symmetrisch** is
- hoe groot de **spreiding** van de data is
- of er enkele waarden ver van de rest affliggen, d.w.z. of er **uitschieters** zijn (een preciezere omschrijving van het begrip uitschieter komt in 1.4 aan bod)
- of er op bepaalde plaatsen **concentraties** van de data zijn
- of er “**gaten**” zitten in de data.

1.4 Klassieke numerieke samenvatting

Naast het histogram zijn de tweede en derde manier van samenvatten die we bekijken een tweetal numerieke samenvattingen. De klassieke numerieke samenvatting van een verzameling data X_1, \dots, X_n bestaat uit de volgende grootheden:

steekproefgrootte	n
steekproefgemiddelde	$\bar{X} = n^{-1} \sum_{i=1}^n X_i$
steekproefvariantie	$S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$
steekproefstandaardafwijking	$S = \sqrt{S^2}$
steekproefscheefheidscoëfficiënt	$B_1 = n^{1/2} \sum_{i=1}^n (X_i - \bar{X})^3 / \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 \right\}^{3/2}$
steekproefkurtosis	$B_2 = n \sum_{i=1}^n (X_i - \bar{X})^4 / \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 \right\}^2$.

Merk op dat exploratieve technieken als methoden gehanteerd kunnen worden zonder dat er sprake is van een model. Echter, zodra we statistische eigenschappen zoals bijv. zuiverheid willen onderzoeken, vereist dit een stochastisch model. Vaak, maar niet altijd zullen we dan veronderstellen dat X_1, \dots, X_n o.o. s.v.-en zijn en identiek verdeeld als X .

Voorbeeld 1.4.1 Van 32 verschillende automodellen is het benzineverbruik weergegeven in mijlen (1609 meter) per gallon (3.79 l.). (De gegevens zijn van 1974.)

21.0	21.0	22.8	21.4	18.7	17.8	16.4	17.3
18.1	14.3	24.4	22.8	19.1	15.2	10.4	14.7
10.4	32.4	30.4	33.9	21.5	15.5	15.2	13.3
19.3	27.2	26.0	30.4	15.8	19.7	15.0	21.4

De klassieke numerieke samenvatting is:

$$n = 32 \quad \bar{x} = 20.09 \quad s^2 = 36.32 \quad s = 6.03 \quad b_1 = 0.64 \quad b_2 = 2.80. \quad \square$$

De eerste vier grootheden van de klassieke numerieke samenvatting, n , \bar{X} , S^2 , S , zijn bekend uit het college Statistiek & kansrekening: het steekproefgemiddelde is een zuivere schatter van de verwachting $\mu = EX$, de steekproefvariantie is een zuivere schatter van de variantie $\sigma^2 = E(X - EX)^2$, de steekproefstandaardafwijking is een (gewoonlijk onzuivere, maar meestal redelijke) schatter van σ , indien X_1, \dots, X_n o.o. s.v.-en zijn en identiek verdeeld als X .

De scheefheid van een **verdeling** kan gemeten worden met de **scheefheidscoëfficiënt**

$$\gamma_1 = \frac{E(X - EX)^3}{\{E(X - EX)^2\}^{3/2}} = \frac{E(X - EX)^3}{\sigma^3}.$$

Is X **symmetrisch** verdeeld dan is $\gamma_1 = 0$: waarden groter dan het midden $c = EX$ vallen weg tegen waarden kleiner dan het midden bij het berekenen van $E(X - EX)^3$. Immers, X is symmetrisch met **symmetriepunt** c dan en slechts dan als $X - c$ dezelfde verdeling heeft als $c - X$. In dat geval is dus

$$E(X - c) = E(c - X)$$

en

$$E(X - c)^3 = E(c - X)^3.$$

Derhalve geldt

$$E(X - c) = -E(X - c) =: E(X - c) = 0 =: EX = c$$

en

$$E(X - c)^3 = -E(X - c)^3 =: E(X - c)^3 = 0 =: \gamma_1 = 0.$$

Is daarentegen de kansverdeling van X **scheef naar rechts**, dan neemt X met grote kans grote waarden aan en is $\gamma_1 > 0$. Omgekeerd: is X **scheef naar links**, dan komen juist kleine waarden met grote kans voor, resulterend in $\gamma_1 < 0$. De begrippen “scheef naar rechts” en “scheef naar links” zijn niet precies omschreven. De vorige zinnen zijn dan ook meer een intuïtieve interpretatie van $\gamma_1 > 0$ en $\gamma_1 < 0$, dan strikte mathematische beweringen. Symmetrie is daarentegen exact gedefinieerd, en impliceert $\gamma_1 = 0$. (Het omgekeerde is niet waar! Een stochastische variabele kan scheefheidscoëfficiënt 0 hebben en toch niet symmetrisch zijn; ga na!).

Merk op dat γ_1 **plaats-** en **schaalinvariant** is, d.w.z. als $Y = aX + b$ met $a > 0$ dan is $E(Y) = E(aX + b) = aE(X) + b$ en dus $Y - E(Y) = aX + b - \{aE(X) + b\} = a\{X - E(X)\}$, zodat

$$E(Y - EY)^2 = E\{a(X - EX)\}^2 = a^2 E(X - EX)^2$$

$$E(Y - EY)^3 = E\{a(X - EX)\}^3 = a^3 E(X - EX)^3$$

resultierend in

$$\begin{aligned} \gamma_1(Y) &= \frac{E(Y - EY)^3}{\{E(Y - EY)^2\}^{3/2}} = \frac{a^3 E(X - EX)^3}{\{a^2 E(X - EX)^2\}^{3/2}} \\ &= \frac{a^3 E(X - EX)^3}{a^3 \{E(X - EX)^2\}^{3/2}} = \frac{E(X - EX)^3}{\{E(X - EX)^2\}^{3/2}} = \gamma_1(X). \end{aligned}$$

Dit is een zeer gewenste eigenschap van de scheefheidscoëfficiënt: immers, als we de scheefheid van bijv. de inkomensverdeling van een bepaalde populatie willen beschrijven, is het niet wenselijk dat we er een ander getal uit krijgen als we de inkomens in guldens uitdrukken dan wanneer we ze in duizenden guldens uitdrukken. Dit is een voorbeeld van schaalinvariantie. Het meten van de temperatuur in graden Celsius of graden Kelvin betreft plaatsinvariantie, terwijl meten in graden Celsius of graden Fahrenheit betrekking heeft op plaats- en schaalinvariantie.

Voor de duidelijkheid brengen we hier in herinnering de conceptuele verschillen tussen **parameters**, **schatters** en **schattingen**. De parameter $\mu = EX$ functioneert in het kansmodel. Het is de voor ons altijd **onbekend** blijvende verwachting van de s.v. X . We kunnen deze parameter wel schatten met behulp van waarnemingen van X . De schatter \bar{X} geeft het **voorschrift** hoe we dat op grond van X_1, \dots, X_n doen. De schatter is dus een (waarneembare) s.v. Zodra we de realisaties x_1, \dots, x_n van X_1, \dots, X_n hebben, kunnen we de schatting bepalen. Dit is de **gerealiseerde waarde** van de schatter. Aan het begin van 1.4 is het voorschrift van de klassieke numerieke samenvatting gegeven. Vandaar dat er hoofdletters gebruikt zijn, duidend op s.v.-en in de context van een statistisch model. Voeren we een klassieke numerieke samenvatting daadwerkelijk uit, dan gebruiken we kleine letters om aan te geven dat het om schattingen gaat.

Zo is in voorbeeld 1.4.1 het getal 20.09 de schatting van μ , gebaseerd op de schatter \bar{X} . Immers, het getal 20.09 is de realisatie van de schatter \bar{X} . De waarde van μ is nog altijd onbekend (hoewel we uiteraard vermoeden dat μ wel in de buurt van 20.09 zal liggen).

Evenzo is γ_1 een onbekende parameter van de kansverdeling van X en is B_1 bedoeld als schatter van γ_1 . In voorbeeld 1.4.1 is 0.64 de schatting van γ_1 . Een eerste indruk hieruit is dat de onderliggende verdeling redelijk symmetrisch is, wellicht enigszins scheef naar rechts (zie ook de tabel met scheefheidscoëfficiënten aan het einde van 1.4).

Waarom B_1 een geschikte schatter van γ_1 is, wordt duidelijk als we B_1 herschrijven. Er geldt

$$B_1 = \frac{n^{1/2} \sum_{i=1}^n (X_i - \bar{X})^3}{\{\sum_{i=1}^n (X_i - \bar{X})^2\}^{3/2}} = \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^3}{\{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2\}^{3/2}}.$$

We kunnen de teller beschouwen als een schatter van $E(X - EX)^3$: immers, schat EX met \bar{X} en schat eveneens de buitenste verwachting met het bijbehorende steekproefgemiddelde. De noemer is evenzo een schatter van σ^3 en daarmee is B_1 een voor de hand liggende schatter van γ_1 .

Analoog is B_2 een voor de hand liggende schatter van de **kurtosis**

$$\gamma_2 = \frac{E(X - EX)^4}{\{E(X - EX)^2\}^2} = \frac{E(X - EX)^4}{\sigma^4}.$$

Ook hier geldt weer dat γ_2 een **parameter** van de kansverdeling van X is. Verder is γ_2 eveneens plaats- en schaalinvariant. Het bewijs is als volgt. Laat $Y = aX + b$ zijn met $a > 0$. Dan is $Y - EY = a(X - EX)$ en dus

$$E(Y - EY)^4 = E\{a(X - EX)\}^4 = a^4 E(X - EX)^4.$$

Omdat $E(Y - EY)^2 = a^2 E(X - EX)^2$, is

$$\gamma_2(Y) = \frac{E(Y - EY)^4}{\{E(Y - EY)^2\}^2} = \frac{a^4 E(X - EX)^4}{\{a^2 E(X - EX)^2\}^2} = \frac{E(X - EX)^4}{\{E(X - EX)^2\}^2} = \gamma_2(X).$$

Merk op dat ook nog geldt $\gamma_2(-X) = \gamma_2(X)$. De parameter γ_2 beschrijft de dikte van de staart van de verdeling. Legt een kansverdeling veel kansmassa in de staart dan wordt de teller flink groot door de 4^e macht en is γ_2 groot. De referentiewaarde is de kurtosis van de normale verdeling. Deze is gelijk aan 3.

In de volgende tabel staan voor enkele verdelingen (zie ook de appendix) de scheefheidscoëfficiënt γ_1 en de kurtosis γ_2 .

verdeling	γ_1	γ_2
uniform	0	1.8
normaal	0	3
exponentiële verdeling	2	9
Erlang (n, λ)	$\frac{2}{\sqrt{n}}$	$3 + \frac{6}{n}$
gamma (p, λ)	$\frac{2}{\sqrt{p}}$	$3 + \frac{6}{p}$
lognormaal $\omega = \exp(\sigma^2)$	$(\omega - 1)^{1/2}(\omega + 2) > 0$	$\omega^4 + 2\omega^3 + 3\omega^2 - 3 > 3$

Omdat voor de normale verdeling $\gamma_2 = 3$, wordt in sommige boeken $\gamma_2 - 3$ de kurtosis genoemd (die dan geschat wordt met $B_2 - 3$). Wij noemen (met vele anderen) γ_2 de kurtosis met 3 als referentiewaarde voor de normale verdeling. Zo concluderen we voor de gegevens van voorbeeld 1.4.1 dat de dikte van de staarten niet uitzonderlijk is.

1.5 Exploratieve samenvatting

1.5.1 EDA samenvatting

Naast de klassieke numerieke samenvatting kennen we de EDA-samenvatting. Ook nu vatten we de data samen in enkele getallen en wel in de volgende: (noteer

de waarnemingen met X_1, X_2, \dots, X_n en de geordende waarnemingen met $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

steekproefgrootte	n
steekproefmediaan	M_n
steekproefkwartielen	
kleinste waarde	$X_{(1)}$
grootste waarde	$X_{(n)}$

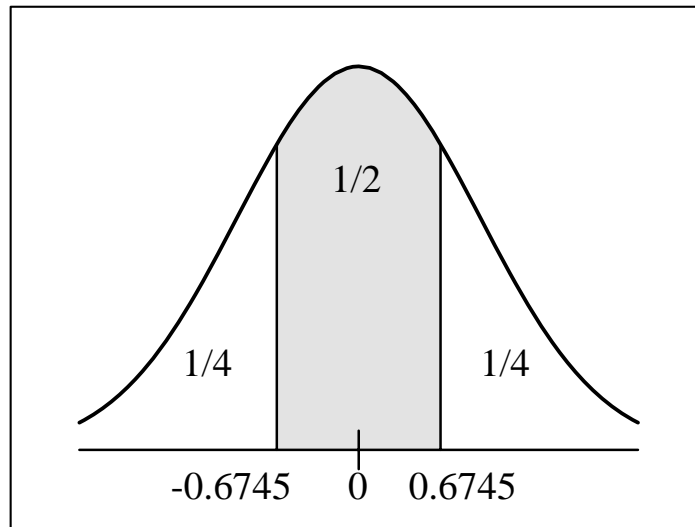
De steekproefgrootte, de kleinste en grootste waarde spreken voor zich. De steekproefmediaan (de 'middelste' waarde) hebben we al eerder gezien (zie 1.2). De twee **steekproefkwartielen** zijn die waarden waar beneden ongeveer een kwart van de waarnemingen ligt (**laagste steekproefkwartiel**), respectievelijk waar boven ongeveer een kwart van de waarnemingen ligt (**hoogste steekproefkwartiel**). Preciezer:

bij een **even** aantal waarnemingen is het laagste steekproefkwartiel gelijk aan de steekproefmediaan van de $\frac{1}{2}n$ kleinste waarnemingen en het hoogste steekproefkwartiel is gelijk aan de steekproefmediaan van de $\frac{1}{2}n$ grootste waarnemingen; bij een **oneven** aantal waarnemingen delen we de waarnemingen in in de $\frac{1}{2}(n+1)$ kleinste en de $\frac{1}{2}(n+1)$ grootste en nemen weer van beide 'helften' de steekproefmediaan.

Het laagste steekproefkwartiel wordt gehanteerd als schatter van het laagste **kwartiel** van de kansverdeling van X , d.w.z. van l gegeven door $P(X \leq l) = 1/4$ (aangenomen dat l bestaat en uniek is). Het getal l is een, veelal onbekende, **parameter** van de kansverdeling van X , het laagste steekproefkwartiel is een **schatter** van l .

Evenzo wordt de **mediaan** m van de kansverdeling gedefinieerd door $P(X \leq m) = 1/2$ (aangenomen dat m hierdoor uniek bepaald is). De mediaan m is een veelal onbekende parameter van de kansverdeling met als **schatter** de steekproefmediaan.

Tenslotte wordt het hoogste steekproefkwartiel gehanteerd als **schatter** van het hoogste kwartiel h van de kansverdeling, gedefinieerd door $P(X \geq h) = 1/4$ (aangenomen dat h hierdoor uniek bepaald is). Het getal h is een veelal onbekende **parameter** van de kansverdeling. Voor een $N(\mu, \sigma^2)$ -verdeling is $l = \mu - 0.6745\sigma$, $m = \mu$ en $h = \mu + 0.6745\sigma$. Immers, als U een $N(0, 1)$ -verdeling heeft, geldt $P(U < -0.6745) = 1/4$ (zie onderstaand plaatje en de tabel van de $N(0, 1)$ -verdeling).



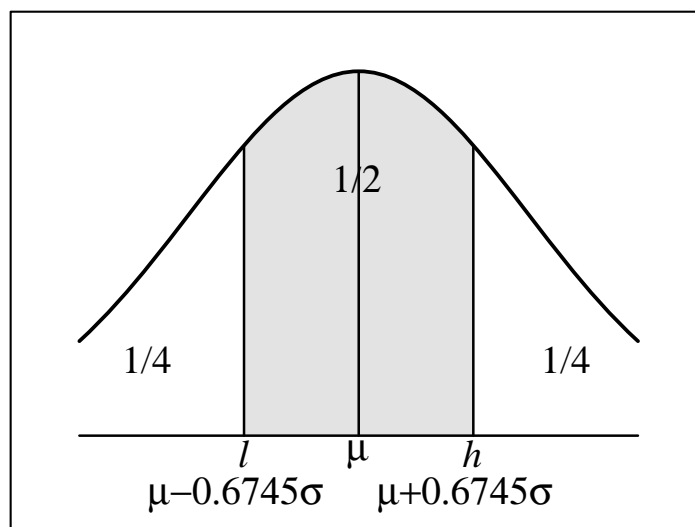
Als $X \sim N(\mu, \sigma^2)$, dan is $(X - \mu)/\sigma \sim N(0, 1)$ en dus

$$P\left(\frac{X - \mu}{\sigma} \leq -0.6745\right) = 1/4.$$

Hieruit volgt

$$P(X \leq \mu - 0.6745\sigma) = 1/4$$

en dus is $l = \mu - 0.6745\sigma$. Op analoge wijze volgt $h = \mu + 0.6745\sigma$.



Voorbeeld 1.5.1 (vervolg van voorbeeld 1.4.1) We sorteren eerst de data van voorbeeld 1.4.1:

10.4	10.4	13.3	14.3	14.7	15.0	15.2	15.2
15.5	15.8	16.4	17.3	17.8	18.1	18.7	19.1
19.3	19.7	21.0	21.0	21.4	21.4	21.5	22.8
22.8	24.4	26.0	27.3	30.4	30.4	32.4	33.9

De steekproefmediaan is het gemiddelde van de middelste twee (n is even) en levert als realisatie op $\frac{1}{2}(19.1 + 19.3) = 19.2$. Het laagste steekproefkwartiel is de steekproefmediaan van de 16 kleinste en geeft $\frac{1}{2}(15.2 + 15.5) = 15.4$ (afgerond). Het hoogste steekproefkwartiel neemt als waarde aan $\frac{1}{2}(22.8 + 22.8) = 22.8$. De kleinste waarde is 10.4, de grootste 33.9.

We geven de numerieke EDA-samenvatting vaak weer in de vorm van een tabel:

steekproefgrootte : 32					
	diepte	laag	hoog	centrum	afstand
steekproefmediaan	16.5	19.2	19.2	19.2	0
steekproefkwartiel	8.5	15.4	22.8	19.1	7.4
extreem	1	10.4	33.9	22.2	23.5

□

De kolom diepte geeft de plaats aan in de rij van geordende waarnemingen, van buiten naar binnen geteld. Zo is 1 de grootste of kleinste waarneming. Verder betekent 16.5 dat de steekproefmediaan het gemiddelde is van de 16-de en 17-de waarneming in de rij van geordende waarnemingen.

De waarde van de steekproefmediaan in de kolommen laag, hoog en centrum is **altijd** dezelfde, want er is maar één steekproefmediaan (ook als n even is). Dus krijgen we in de kolom afstand = "hoog" – "laag" voor de steekproefmediaan **steeds** 0. (In voorbeeld 1.5.1 krijgen we dus op de regel "steekproefmediaan" zowel bij "laag", "hoog" als "centrum" de waarde 19.2 en bij de kolom "afstand" de waarde 0.)

De kolom centrum is het gemiddelde van de kolommen laag en hoog. Wanneer de kolom centrum stabiel is, wijst dit op **symmetrie**. Immers, dan liggen beide steekproefkwartielen ongeveer even ver van de steekproefmediaan en beide extremen eveneens ongeveer symmetrisch t.o.v. de steekproefmediaan.

De **steekproefkwartielafstand** = hoogste steekproefkwartiel – laagste steekproefkwartiel is een maat voor de **spreiding** van de getallen. Het hoogste steekproefkwartiel is een schatter van het hoogste kwartiel h van de kansverdeling; het laagste steekproefkwartiel is een schatter van l , het laagste kwartiel van de kansverdeling. De steekproefkwartielafstand hanteren we als schatter van $h - l$.

Omdat bij een $N(\mu, \sigma^2)$ -verdeling de afstand tussen de kwartielen van de verdeling gelijk is aan $h - l = \mu + 0.6745\sigma - (\mu - 0.6745\sigma) = 1.349\sigma$, wordt

$$\frac{\text{de steekproefkwartielafstand}}{1.349}$$

gehanteerd als schatter voor $\sigma = (h - l)/1.349$. De keuze van de constante 1.349 wordt gemotiveerd vanuit **normaliteit**. Het idee is dat we hopen dat normaliteit een redelijke modelveronderstelling is, maar dat we ons door gebruik te maken van de steekproefkwartielafstand willen indekken tegen kleine afwijkingen van normaliteit, bijv. in de vorm van uitschieters. Ligt om één of andere reden een andere verdeling meer voor de hand, dan is het verstandiger 1.349 te vervangen door een bij die verdeling passende contante.

Voorbeeld 1.5.2 (vervolg van voorbeeld 1.4.1 en 1.5.1) Voor het benzineverbruik van de 32 verschillende automodellen is $s = 6.03$ en de steekproefkwartielafstand/1.349 = 5.49. Beide schattingen komen redelijk overeen. \square

De steekproefkwartielen zijn weinig gevoelig voor **uitschieters**. Daarmee is ook de schatter van σ gebaseerd op de steekproefkwartielafstand minder gevoelig voor uitschieters dan S . We illustreren dit aan de hand van het volgende voorbeeld.

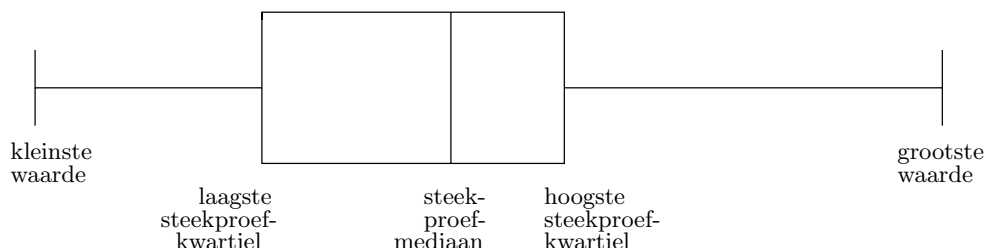
Voorbeeld 1.5.3 De oorspronkelijke verzameling data is:

31.357 46.514 53.525 11.719 70.882.

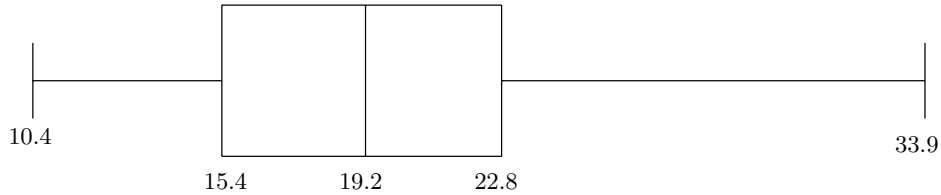
In dit geval is $s = 22.44$. Het laagste steekproefkwartiel is 31.357, het hoogste steekproefkwartiel 53.525, zodat de steekproefkwartielafstand/1.349 gelijk is aan 16.43. Door een typefout is in het laatste getal de decimale punt vergeten, waardoor het getal 70882 is geworden. Nu is $s = 31683$ terwijl de steekproefkwartielafstand/1.349 niet verandert. Uit dit voorbeeld blijkt duidelijk de resistentie van de schatter gebaseerd op de steekproefkwartielafstand en de gevoeligheid van s . De klassieke samenvatting geeft een verkeerd beeld van de data. Er is niet sprake van een enorme spreiding, maar er is een uitschieter en de spreiding in het merendeel van de data is in de orde van ongeveer 20 (en niet 30000). \square

1.5.2 Boxplot

Veelal wordt de EDA-samenvatting weergegeven in een plaatje, de zogenaamde boxplot. Schematisch ziet dat er zo uit



Voorbeeld 1.5.4 (vervolg van voorbeeld 1.4.1, 1.5.1 en 1.5.2) De EDA samenvatting van het benzinegebruik van de 32 auto's ziet er zo uit



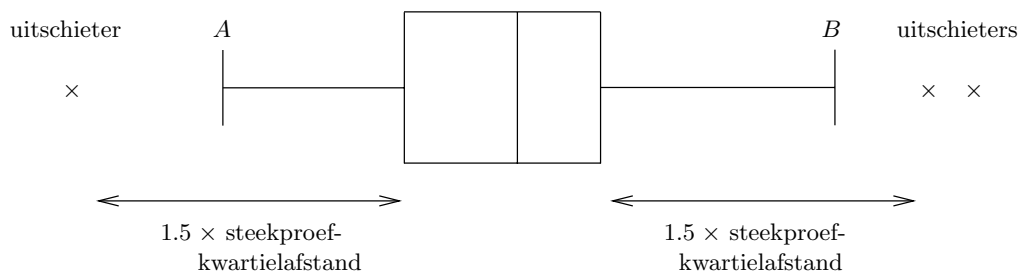
□

Aan de boxplot kunnen we, vaak duidelijker dan aan de getallen, snel een eerste indruk van de data ontleen. Zo zien we in voorbeeld 1.5.4 dat ongeveer de helft van de data ligt tussen 15 en 23 en dat de data in dat gebied redelijk symmetrisch zijn (de steekproefmediaan ligt mooi in het midden van de box). Wat eerder wellicht nog niet opgevallen was, is dat de waarde 33.9 tamelijk ver van de rest affligt. Dit is ook te zien aan de laatste regel van de EDA-samenvatting in tabelvorm, waar de kolom centrum 22.2 geeft. De waarde is daar een stuk hoger dan de eerdere waarden 19.2 en 19.1. Dit wordt veroorzaakt doordat de grootste waarde 33.9 een uitschieter is.

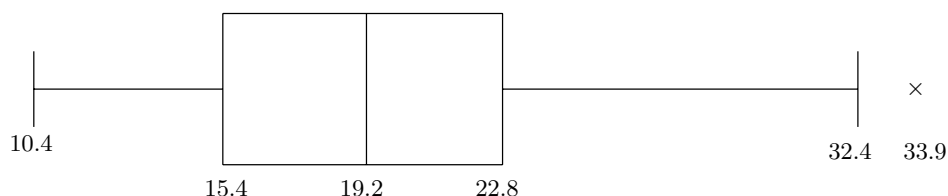
Veelal wordt de basisboxplot zoals we die hierboven gezien hebben nog iets uitgebreid met name om uitschieters op te sporen. We gaan als volgt te werk: we maken eerst de box, zetten dan $1.5 \times$ steekproefkwartielafstand naar links en rechts uit vanaf de box en noteren de uiterste waarden uit de steekproef die **daar binnen** liggen (A en B). De waarden die op $1.5 \times$ steekproefkwartielafstand of meer van de box affliggen worden als uitschieters beschouwd en worden aangegeven met \times . We hanteren dus als definitie voor uitschieters het volgende

uitschieter: $\text{waarde} \leq \text{laagste steekproefkwartiel} - 1.5 \times \text{steekproefkwartielafstand}$ of $\text{waarde} \geq \text{hoogste steekproefkwartiel} + 1.5 \times \text{steekproefkwartielafstand}$
--

Schematisch ziet het er dan zo uit.



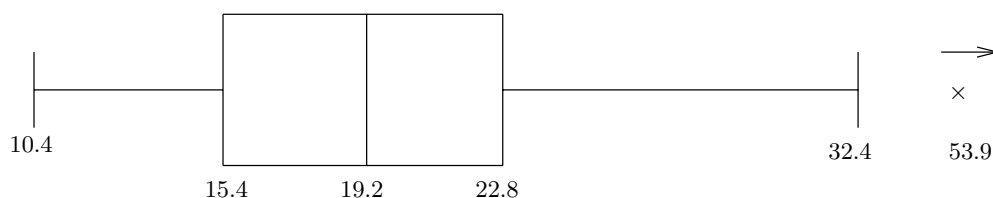
Voorbeeld 1.5.5 (vervolg van voorbeeld 1.4.1, 1.5.1, 1.5.2 en 1.5.4) De steekproefkwartielafstand is 7.4; de benedengrens voor uitschieters is $15.4 - 1.5 \times 7.4 = 4.3$; waarden ≤ 4.3 zijn er niet; er zijn dus geen lage uitschieters; de kleinste waarde > 4.3 is $A = 10.4$; de bovengrens voor uitschieters is $22.8 + 1.5 \times 7.4 = 33.9$; er is dus één hoge uitschieter 33.9 (net op de rand!); de grootste waarde < 33.9 is $B = 32.4$. Hiermee hebben we alle benodigde gegevens en kunnen we de boxplot tekenen.



□

Let bij het maken van boxplots erop dat de getallen op de **juiste afstand** van elkaar getekend worden. Ook uitschieters moeten precies op de juiste plaats gezet worden en niet schematisch met ergens ver weg een \times . Ligt een uitschieter heel ver van de rest af, dan wordt deze aangegeven aan de rand van het plaatje met een \times , de getalwaarde en een pijltje in de richting waarin de uitschieter ligt.

Voorbeeld 1.5.6 (vervolg van voorbeeld 1.4.1, 1.5.1, 1.5.2, 1.5.4 en 1.5.5) Als we dezelfde gegevens hebben als in voorbeeld 1.5.5 behalve dat 33.9 vervangen wordt door 53.9, dan ziet de boxplot er zo uit:



□

De keuze van de factor 1.5 impliceert dat als de waarnemingen een grote steekproef uit een $N(\mu, \sigma^2)$ -verdeling zijn, ongeveer 0.7% als uitschieters wordt gekwalificeerd. Immers, de kwartielen van de $N(\mu, \sigma^2)$ -verdeling zijn $\mu - 0.6745\sigma$ en $\mu + 0.6745\sigma$; $1.5 \times$ kwartielafstand = $1.5 \times 1.349\sigma$ zodat de grenswaarden voor uitschieters bij benadering zijn: $\mu - 0.6745\sigma - 1.5 \times 1.349\sigma = \mu - 2.698\sigma$ en $\mu + 2.698\sigma$. De kans dat een $N(\mu, \sigma^2)$ -verdeelde stochastische grootte buiten het interval $(\mu - 2.698\sigma, \mu + 2.698\sigma)$ terecht komt is gelijk aan de kans dat een $N(0, 1)$ -verdeelde stochastische grootte buiten het interval $(-2.698, +2.698)$ terecht komt. Uit de tabel van de $N(0, 1)$ -verdeling vinden we voor deze kans $0.007 = 0.7\%$. Voor grote n zullen bij normaal verdeelde waarnemingen dus ongeveer $0.007n$ waarnemingen als uitschieter bestempeld worden.

Uit de boxplots lezen we iets af over

- de **ligging** van de **data** (via de **steekproefmediaan**)
- de **spreiding** (via de lengte van de box = **steekproefkwartielafstand**)
- de **scheefheid** (steekproefmediaan midden in de box wijst op symmetrie; is het linkergedeelte van de box groter dan het rechtergedeelte, dan zijn er dus relatief (t.o.v. symmetrie) veel kleine waarnemingen, d.w.z. dit wijst op scheefheid naar **links**; is het rechtergedeelte groter, dan zijn er relatief veel grote waarnemingen, d.w.z. er zijn aanwijzingen voor scheefheid naar **rechts**)
- **uitschieters**.

Speciaal ook om datasets te vergelijken kunnen boxplots heel geschikt zijn.

Het is beslist niet zo dat de ene samenvattingsmethode altijd beter is dan de andere. Veeleer vullen ze elkaar aan. Het is daarom verstandig je niet te beperken tot één van de besproken methodes, maar verschillende soorten samenvattingen te maken om zo optimaal te ontdekken wat de data (in eerste instantie) te vertellen hebben.

1.6 Opgaven

1. Geef van ieder van de volgende variabelen aan of ze metingen zijn op een nominale, ordinale of intervallschaal.
 - a. De concentratie van DDT in een steekproef van melk, in milligram per liter.
 - b. De soort insect bij een steekproef in akkerbouwland.
 - c. De postcode van de proefpersonen bij een schriftelijke enquête.
 - d. De plaats van F.C. Twente op de ranglijst van de eredivisie.
 - e. De huisnummers in een straat.
 - f. De cijfers van een tentamen.
2. Het Bruto Nationaal Product (BNP) per inwoner in dollars voor de groep armste landen in 1986 (d.w.z. landen, waarvoor het BNP per inwoner ten hoogste 420 dollar is) is hieronder weergegeven.

120, 150, 150, 150, 160, 160, 160, 180, 200, 210, 230, 230,
 240, 250, 250, 260, 270, 280, 290, 290, 290, 300, 300, 300,
 310, 320, 330, 350, 370, 390, 400, 420, 420.

 - a. Maak een histogram van deze gegevens.
 - b. Bereken de steekproefmodi.

- c. Met een lokatiemaat proberen we in één getal de ligging van de waarnemingen te karakteriseren. Is in deze data-set de steekproefmodus een geschikte lokatiemaat? Argumenteer uw antwoord.
3. Gegeven zijn de volgende omzetcijfers (in duizenden guldens per week) van supermarkten gedurende twee perioden.

Periode 1

97	153	168	90	64	206	75	113	100
102	181	94	132	86	140	123	106	249

Periode 2

176	80	185	115	115	118	228	130	129
92	81	105	102	127	177	191	277	194
144	77	285	86					

Maak 2 boxplots (van iedere periode één) van deze gegevens, zodanig gepresenteerd dat de resultaten van beide periodes gemakkelijk vergeleken kunnen worden.

4. Men wenst het koopgedrag van klanten in een winkel te beschrijven. Van 50 mensen is het bestede bedrag hieronder weergegeven.

6, 17, 17, 20, 24, 26, 28, 29, 32, 32, 34, 34,
 36, 36, 36, 38, 38, 41, 43, 46, 46, 47, 49, 49,
 51, 52, 52, 53, 60, 65, 68, 72, 73, 76, 80, 82,
 83, 85, 87, 91, 94, 98, 102, 110, 114, 131, 154, 160,
 161, 174.

Er geldt $\bar{x} = 64.64$, $\sum_{i=1}^{50}(x_i - \bar{x})^2 = 26867$, $\sum_{i=1}^{50}(x_i - \bar{x})^3 = 3440039$,
 $\sum_{i=1}^{50}(x_i - \bar{x})^4 = 453007146$.

- a. Geef de klassieke numerieke samenvatting van deze data.
- b. Maak een boxplot in uitgebreide zin van deze gegevens.
- c. Welk type kansverdeling lijkt u geschikt bij de modelbeschrijving: een symmetrische verdeling, een verdeling scheef naar rechts of een verdeling scheef naar links? Argumenteer uw antwoord.
- d. Bepaal de steekproefkwartielafstand en bereken de steekproefkwartielafstand/1.349.
- e. Vergelijk s en de bij d. gevonden schatting van σ en geef commentaar.
5. Maak een boxplot in uitgebreide zin van de volgende gegevens

5.83	1.55	-3.13	4.77	1.49	-3.78	4.39	12.54	-2.24
0.80	3.65	4.38	0.49	3.42	1.28	5.27	-2.32	

6. Van 5 juni 1879 tot 2 juli 1879 verrichtte Michelson metingen om de lichtsnelheid te bepalen. De klassieke numerieke samenvatting van deze gegevens (in km/sec.) luidt:

steekproefgrootte	100
steekproefgemiddelde	299852.4
steekproefvariantie	6243
steekproefstandaardafwijking	79.01
steekproefscheefheidscoëfficiënt	-0.018
steekproefkurtosis	3.26

Aan welke kansverdeling zou u in eerste instantie denken om te gebruiken in het kansmodel dat dit experiment beschrijft? Waarom? Zijn dit voldoende gegevens om tot een definitieve keuze te komen?

7. De Boeing-fabriek levert vliegtuigmaterialen aan de KLM. De geplande leveringstijd van de materialen is 70 dagen. De klassieke numerieke samenvatting van de feitelijke levertijden gedurende een bepaalde periode ziet er als volgt uit:

steekproefgrootte	60
steekproefgemiddelde	57.73
steekproefvariantie	2540.94
steekproefstandaardafwijking	50.41
steekproefscheefheidscoëfficiënt	2.81
steekproefkurtosis	12.34

Is een normale verdeling een goede modelkeuze voor dit experiment? Motiveer het antwoord.

8. De data van opgave 7 zijn als volgt:

59	28	59	33	48	84	20	64	23	29
72	37	30	30	52	40	177	54	214	70
167	70	55	76	29	46	75	74	57	56
32	21	22	18	39	126	103	303	7	38
26	44	32	80	29	77	65	77	36	65
6	32	37	25	30	31	23	56	29	27

Laat de waarneming 303 weg en bereken de steekproefvariantie gebaseerd op de overige waarnemingen.

Geef commentaar op deze uitkomst. Maak de EDA samenvatting voor alle data en geef deze ook weer in de boxplot. Geef commentaar op de gevonden resultaten, ook in samenhang met opgave 7. Bereken ook de steekproefkwartielafstand/1.349 en vergelijk deze uitkomst met de steekproefstandaardafwijking. Geef commentaar.

9. Hieronder staan van 31 praktijken van huisartsen de praktijkgroottes. Geef een exploratieve numerieke samenvatting (EDA-samenvatting) van deze gegevens.

2173	2209	1790	2070	1426	1915	1615	1089	1672	1757	1475
988	1591	1436	1549	1570	1404	1949	1740	3344	193	2960
2936	3012	1685	3125	2638	2848	2583	3436	1680		

10. Maak van de volgende gegevens, betreffende de aantallen mm. neerslag in juni in De Bilt gedurende 10 jaren, een boxplot in uitgebreide zin, d.w.z. een boxplot waarbij ook aandacht aan uitschieters wordt gegeven:

45 50 34 31 36 40 128 77 45 67.

11. Om de drukte bij een telefonische hulpdienst te meten heeft men gedurende 35 weken het aantal telefoontjes geteld. De resultaten staan hieronder

23	22	38	43	24	35	27	25	23	22	52	31
30	41	29	28	37	25	29	31	24	49	33	25
27	25	34	32	21	23	24	18	48	23	16.	

- Maak een boxplot in uitgebreide zin van deze gegevens.
- Suggereert de boxplot symmetrie, scheefheid naar rechts of scheefheid naar links? Argumenteer uw antwoord.