Hoofdstuk 4

Beschrijvende statistiek

Statistiek is het verzamelen en bestuderen van numerieke gegevens om vervolgens conclusies te trekken uit deze data.

Zijn we bijvoorbeeld geïnteresseerd in de lengte van alle Belgische 21-jarigen, dan is de *populatie* de verzameling van al die lengtes. Dit is een zeer omvangrijke groep die een onderzoeker allicht niet ter beschikking heeft. Derhalve beperkt men zich vaak tot het onderzoek van een *steekproef*. Dit is een deelverzameling van de populatie. Een goede steekproef moet voldoende elementen bevatten en een goed beeld geven van de volledige populatie.

In de *beschrijvende statistiek* zal men de meetresultaten overzichtelijk weergeven in tabellen of grafische voorstellingen en samenvatten d.m.v. enkele kengetallen. We maken hierbij een onderscheid tussen centrummaten en spreidingsmaten.

Op basis van een steekproef zal men in de *verklarende statistiek* uitspraken doen over de ganse populatie. Dit komt aan bod in het volgende hoofdstuk.

4.1 Centrummaten

4.1.1 Het rekenkundig gemiddelde

Het (rekenkundig) gemiddelde van een reeks numerieke gegevens x_1, x_2, …, x_n is het getal $\frac{\sum_{i=1}^{n} x_i}{n}$.

We gebruiken de notatie μ voor het populatiegemiddelde en $\bar{x}$ voor een steekproefgemiddelde.

Voor de steekproefdata 8, -3, 0, 5, 1 , 4 ,-1 geldt : $\bar{x} = 2$.

4.1.2 Mediaan

Voor het berekenen van de mediaan van een reeks numerieke gegevens orden je ze eerst van klein naar groot.

De mediaan (notatie : *Med*), is het middelste getal voor een oneven aantal data en het gemiddelde van de twee middelste getallen voor een even aantal data.

De mediaan van de gegevens -3, -1, 0, 1, 4, 5, 8 is 1 en van -3, -1, 1, 4, 5, 8 is de mediaan 2,5.

4.2 Spreidingsmaten

4.2.1 De spreidingsbreedte

De spreidingsbreedte *R* (*range* in het Engels) van een reeks numerieke gegevens is het verschil tussen het grootste en het kleinste getal : $R = x_{\max} - x_{\min}$.

De spreidingsbreedte van de gegevens -3, -1, 0, 1, 4, 5, 8 is $R = 11$.

4.2.2 De interkwartielafstand

De mediaan verdeelt de geordende gegevens in een linker- en een rechterdeel met hetzelfde aantal getallen. De mediaan bepaalt de grens tussen deze twee delen en behoort tot geen van deze delen.

Het eerste kwartiel, Q_1, is de mediaan van het linkerdeel en het derde kwartiel, Q_3, de mediaan van het rechterdeel.

Voor -3, -1, 0, 1, 4, 5, 8 geldt : Q_1 = -1 en Q_3 = 5

Voor -3, -1, 0,1, 4, 5, 8, 9 geldt : Q_1 = -0,5 en Q_3 = 6,5.

De interkwartielafstand *IQR* meet de spreiding van de middelste helft van de geordende gegevens.

$IQR = Q_3 - Q_1$.

4.2.3 De standaardafwijking

In combinatie met het rekenkundig gemiddelde als centrummaat geeft men vaak de standaardafwijking als spreidingsmaat.

De *populatievariantie* σ^2 van een populatie van N getallen is het gemiddelde van de kwadraten van de afwijkingen van die getallen t.o.v. hun gemiddelde. Voor een *steekproefvariantie* s^2 van een steekproef van n getallen deel je door $n-1$ i.p.v. door n. We zullen later verklaren waarom.

populatievariantie	steekproefvariantie
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

De positieve vierkantswortel van de variantie noemt men de standaardafwijking. Dit geeft de populatiestandaardafwijking σ en de steekproefstandaardafwijking *s*.

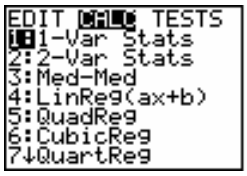
Zo geldt voor de steekproefdata 8, -3, 0, 5, 1 , 4 ,-1 met $\bar{x}=2$ dat

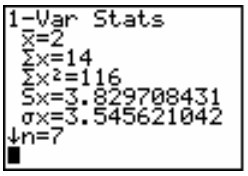
$$s^2 = \frac{6^2+(-5)^2+(-2)^2+3^2+(-1)^2+2^2+(-3)^2}{6} = \frac{44}{3}, \text{ zodat } s = 3.83.$$

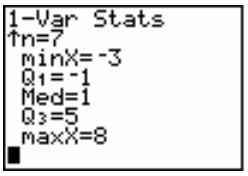
4.3 Statistische kengetallen en de TI-83

Om het berekenen van statistische kengetallen te illustreren, definiëren we eerst de lijst **L1** als volgt : **L1={8,-3,5,Ø,1,4,-1}**.

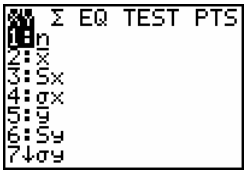
Na het indrukken van **STAT<CALC> 1:1Var Stats 2nd[L1]** verschijnen er veel kengetallen van de lijst **L1** op het basisscherm.







Na het uitvoeren van het commando **1:Var Stats** is het mogelijk deze kengetallen te gebruiken in verdere berekeningen. Je kan de kengetallen oproepen met **VARS 5:Statistics**.



4.4 Histogram en frequentietabel

Het is aangeraden om eerder gedefinieerde grafieken en statistische plots uit te zetten vooraleer we statistische plots construeren,

Dit doe je respectievelijk met de commando's **FnOff** en **PlotsOff**. Dit laatste commando vind je in het **2nd[STAT PLOTS]**-menu. **FnOff** kan je oproepen vanuit de cataloog (**2nd[CATALOG]**).

Voor we de constructie van een histogram behandelen, plaatsen we de onderstaande data, de schoenmaat 30 volwassen mannen, in **L2**.

42	39	42	41	40	44	43	41	40	40
42	40	39	38	43	40	39	44	42	40
41	46	40	41	42	42	38	39	44	41

Het tekenen van een histogram gaat als volgt :

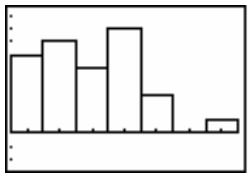
a. **2nd[STAT PLOT] 1:Plot1.**

b. Zet de cursor op **On** en druk op **ENTER**.

c. Selecteer, met ◀▶, voor **Type** het pictogram ⊓ van Histogram.

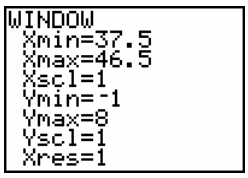
d. Tik achter **Xlist 2nd[L2]**.

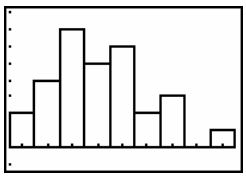
Standaard staat **Freq** op **1** (= maak geen gebruik van frequenties).



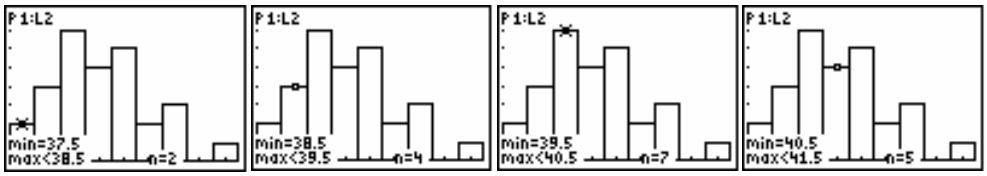
e. Een aan de data aangepast venster kies je d.m.v. **ZOOM 9:ZoomStat**.

f. Stel de vensterinstellingen in zoals hieronder aangegeven en druk op **GRAPH**. De *klassenbreedte*, d.i. de breedte van de rechthoekjes, wordt aangegeven door **Xscl**. Door de onderstaande vensterinstellingen worden de schoenmaten 38 t.e.m. 46 de *klassenmiddens* van het histogram.





Druk op **TRACE** als het grafisch venster actief is. Door te drukken op de ◀▶ toetsen kun je dan de *klassenfrequenties* aflezen van het scherm. Er zijn bijvoorbeeld 7 data x_i met $39.5 \le x_i < 40.5$, wat in dit voorbeeld neerkomt op 7 keer schoenmaat 40. De frequenties worden aangegeven door de hoogte van de rechthoekjes.



A.h.v. de **TRACE**-functie kunnen we de onderstaande *frequentietabel* opstellen. We plaatsen deze data in de lijsten **L**1 en **L**2.

Schoenmaat	38	39	40	41	42	43	44	45	46
Aantal	2	4	7	5	6	2	3	0	1

Het bepalen van de kengetallen en het tekenen van een histogram a.h.v. de frequenties verloopt zoals aangegeven op de onderstaande plaatjes.

Het tekenen van een "staafdiagram" is mogelijk door **Xscl** gelijk te stellen aan bv. 0.5 in de vensterinstellingen.

Een echt staafdiagram bestaat uit lijnstukken ter hoogte van de *verschillende* discrete gegevens. De lengte van de lijnstukken correspondeert met de frequentie van de verschillende data.

Een staafdiagram is in dit voorbeeld meer aangewezen dan een histogram, aangezien je bij een histogram niet kunt zien waaraan de data van een bepaalde klasse gelijk zijn, je kent enkel de grenzen.

In de oefeningen geven we een programma waarmee je rechtstreeks de frequenties kan berekenen van discrete data, zonder gebruik te maken van een histogram.

4.5 De frequentiepolygoon

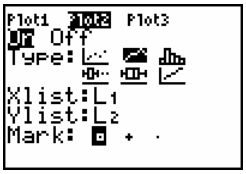
De *frequentiepolygoon* verbindt de opeenvolgende punten (x_i, f_i) met x_i een klassenmidden en f_i de klassenfrequentie. We voegen links en rechts een lege klasse toe (een klasse met frequentie nul) om te starten en te eindigen op de *x*-as.

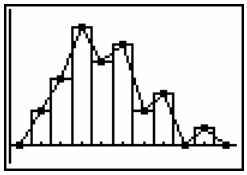
Het plotten van een frequentiepolygoon, bovenop een histogram, verloopt als volgt. Pas de data in de lijsten **L1** en **L2** aan zoals hieronder aangegeven :

Schoenmaat	37	38	39	40	41	42	43	44	45	46	47
Aantal	0	2	4	7	5	6	2	3	0	1	0

Stel **Xmin = 36.5** en **Xmax = 47.5** in de vensterinstellingen. Definieer **Plot1** als het hierbij horende histogram.

De frequentiepolygoon definiëren we als **Plot2** zoals hieronder aangegeven en we plotten beide statistische plots.





Indien we alleen de frequentiepolygoon willen tekenen, zetten we **Plot1** uit. Met **Plot2** verbindt de **TI-83** gewoon de opeenvolgende koppels uit de lijsten **L1** en **L2**. Dit levert inderdaad de frequentiepolygoon.

Om de *cumulatieve frequentiepolygoon* te tekenen voeren we de volgende stappen uit.

- Zet eerst alle statistische plots uit en pas de data in lijst **L1** en **L2** aan.

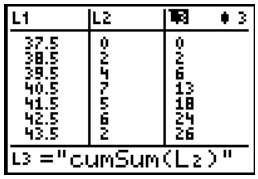
L1	37.5	38.5	39.5	40.5	41.5	42.5	43.5	44.5	45.5	46.5
L2	0	2	4	7	5	6	2	3	0	1

  Merk op dat we nu in **L1** de bovenste klassengrens van elke klasse noteren.

- Definieer de lijst **L3** als de cumulatieve som van **L2**

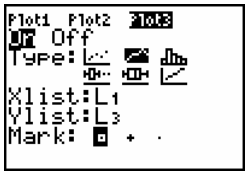
  2nd[LIST]<OPS> 6:cumSum(2nd[L2]).

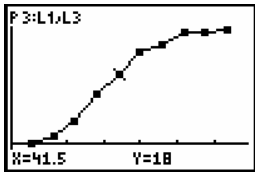
  Dan bevat **L3** de cumulatieve frequentie van de klassen.



- Definieer **Plot3** als de frequentiepolygoon van **L3** door het venster van **Plot3** zoals hieronder in te vullen.

- Voer het commando **9:ZoomStat** uit.





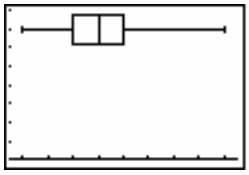
Er zijn 18 data kleiner dan 41.5.

4.6 Box-plots

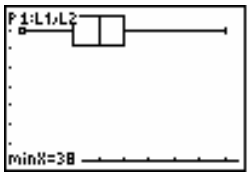
Een box-plot is een grafische voorstelling van de volgende 5 getallen als samenvatting van de gegeven data :

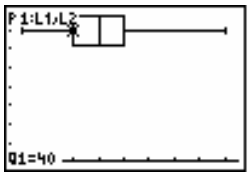
Het minimum, het eerste kwartiel, de mediaan, het derde kwartiel, het maximum.

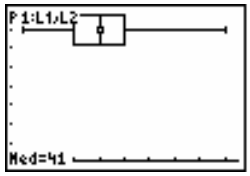
Het genereren van een box-plot () verloopt analoog aan het genereren van een histogram. Voor de data uit 4.4 wordt het **STAT PLOT**–venster in dit geval als volgt ingevuld :

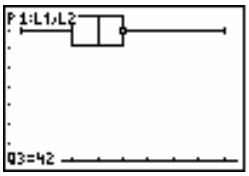


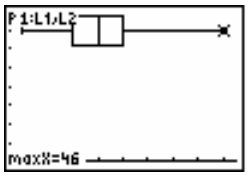
Het uitvoeren van de **TRACE**-functie op de box-plot levert de volgende vijf kengetallen op : **minX**, **Q1**, **Med**, **Q3**, **maxX**.









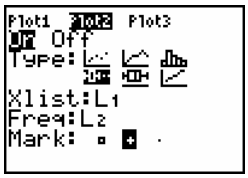


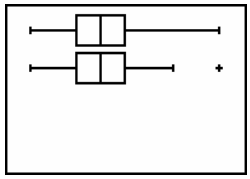
Om tijdens het uitvoeren van **TRACE** geen commentaar, zoals zo-even (**P1:L1,L2**), te bekomen, selecteer je in het **FORMAT**-venster de optie **ExprOff**. Indien je bovendien geen assen wil, selecteer je ook de optie **AxesOff.**

In de box-plots hierboven lopen de snorren door tot aan het minimum en het maximum. Een andere lay-out is de box-plot met uitschieters. Deze box-plot wordt getekend zoals een gewone, uitgezonderd de punten die verder dan 1.5 keer de interkwartielafstand *IQR* links van Q_1 of rechts van Q_3 liggen.

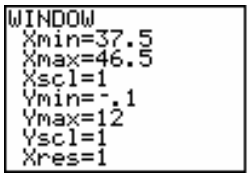
Deze punten worden afzonderlijk geplot en noemt men *uitschieters*. Dit zijn opvallend grote of kleine data in vergelijking met de andere data. De uiteinden van de snorren zijn in dit geval het kleinste en grootste getal die geen uitschieters zijn.

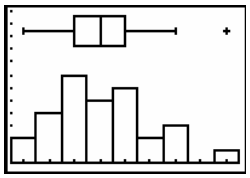
Definieer **Plot1** zoals hierboven en **Plot2** als de box-plot met uitschieters (⊡⋯) zoals hieronder aangegeven en teken beide box-plots.





De box-plot met uitschieters geeft meer informatie, het is dan ook aangeraden om deze laatste te gebruiken.
Bovendien is het mogelijk een histogram en een box-plot tegelijk op het scherm te plotten. Definieer **Plot1** als de box-plot en **Plot2** als het histogram. Indien de plots mekaar overlappen, moet je de *y*-as aanpassen.





Box-plots worden vaak gebruikt voor het vergelijken van verschillende reeksen van vergelijkbare data. Met de **TI-83** kan je tot 3 box-plots op één venster tekenen.

4.7 Analyse van een probleem

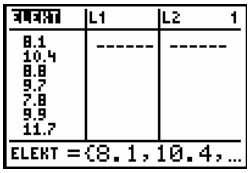
Een fabrikant van telecommunicatiemateriaal ontving klachten in verband met een zwak geluid bij telefoonverbindingen op lange afstand. Deze fabrikant had honderden elektronische versterkers geleverd die op regelmatige afstand langs de telefoonlijn werden gemonteerd.

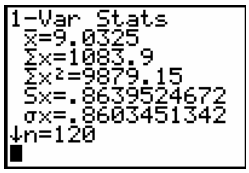
Als hoofdverdachte werd de versterkingsfactor van de versterkers aangeduid. In de verkoopsvoorwaarden werd als streefwaarde of *nominale waarde* een versterkingsfactor van 10 dB (decibel), een toelaatbare minimale versterkingsfactor van 7.75 dB en maximaal 12.25 dB vermeld. We noemen 7.75 en 12.25 dB de *tolerantiegrenzen*.

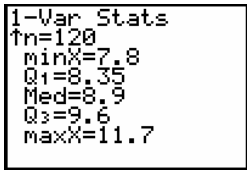
Om de klacht te onderzoeken werden er lukraak 120 versterkers verzameld die tot hetzelfde productielot of populatie behoorden als diegene waarover de klacht werd uitgebracht. Het resultaat van de 120 gemeten versterkingsfactoren vind je in volgende tabel.

8.1	10.4	8.8	9.7	7.8	9.9	11.7	8.0	9.3	9.0
8.2	8.9	10.1	9.4	9.2	7.9	9.5	10.9	7.8	8.3
9.1	8.4	9.6	11.1	7.9	8.5	8.7	7.8	10.5	8.5
11.5	8.0	7.9	8.3	8.7	10.0	9.4	9.0	9.8	10.7
9.3	9.7	8.7	8.2	8.9	8.6	9.5	9.4	8.8	8.3
8.4	9.1	10.1	7.8	8.1	8.8	8.0	9.2	8.4	7.8
7.9	8.5	9.2	8.7	10.2	7.9	9.8	8.3	9.0	9.6
9.9	10.6	8.6	9.4	8.8	8.2	10.5	9.7	9.1	8.0
8.7	9.8	8.5	8.9	9.1	8.4	8.1	9.5	8.7	9.3
8.1	10.1	9.6	8.3	8.0	9.8	9.0	8.9	8.1	9.7
8.5	8.2	9.0	10.2	9.5	8.3	8.9	9.1	10.3	8.4
8.6	9.2	8.5	9.6	9.0	10.7	8.6	10.0	8.8	8.6

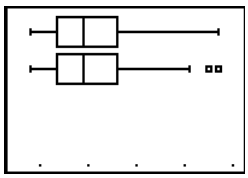
De data worden eerst ingevoerd in de lijst ELEKT en met **1-Var Stats ʟELEKT** verkrijgen we dan de statistische kengetallen van de data.





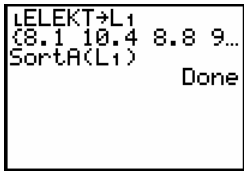


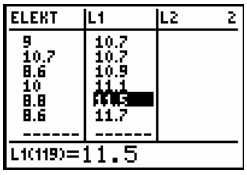
De minimale waarde is groter dan 7,75 en het maximum is kleiner dan 12,25 zodat al de data binnen de vooropgestelde specificatiegrenzen vallen. Maar het gemiddelde 9,03 ligt duidelijk lager dan de ideale waarde 10. Daar de mediaan kleiner is dan het gemiddelde, verwachten we een *positief scheve verdeling*, d.w.z. met de langste staart naar rechts. Een box-plot en een histogram verduidelijken de situatie.

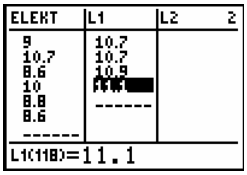


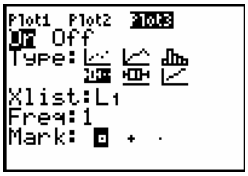
De gewone box-plot en de box-plot met uitschieters werden onder elkaar getekend. Traceren van de tweede plot levert 11.1 als grootste niet-uitschieter. De uitschieters zijn 11.5 en 11.7 (= **maxX**). De box-plot met uitschieters levert een beter beeld van de scheefheid van de verdeling.

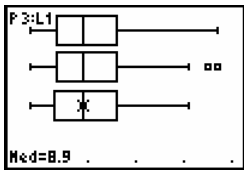
Om het effect van de uitschieters na te gaan kopiëren we de lijst ELEKT naar de lijst **L1** en vervolgens rangschikken we **L1** van klein naar groot. Na het verwijderen van de uitschieters uit **L1** berekenen we opnieuw de kengetallen en voegen we de box-plot met uitschieters (**Plot3**) toe aan de bovenstaande box-plots.

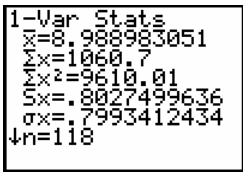


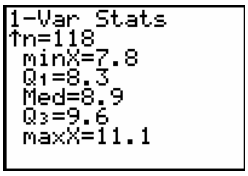






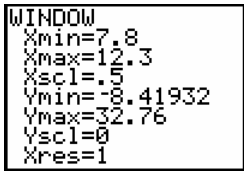


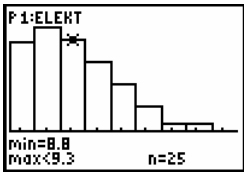


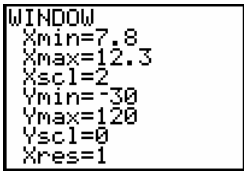


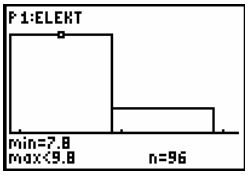
We zien dat het eerste kwartiel, de mediaan en tweede kwartiel nagenoeg ongewijzigd blijven bij weglaten van de uitschieters. Het rekenkundig gemiddelde echter daalt van 9.03 naar 8.99 en de steekproefstandaardafwijking *s* daalt van 0.86 naar 0.80. Voor de steekproefstandaardafwijking is dit een daling van 7% !

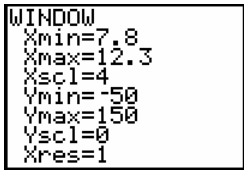
In de volgende plaatjes maken we een histogram van de lijst ELEKT en bestuderen we de invloed van de klassenbreedte op het histogram:

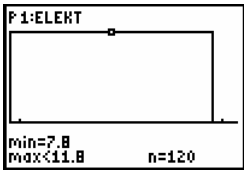


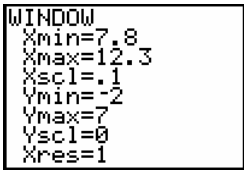


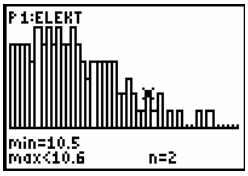




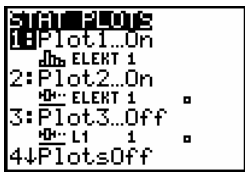


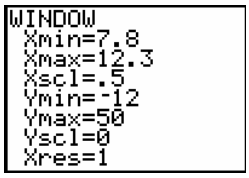


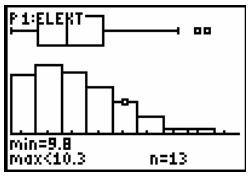




We zien dat een klassenbreedte van 0.5 een goed beeld geeft. Met breedte 2 gaat te veel informatie verloren en met breedte 4 krijg je zelfs de indruk van een uniforme verdeling. Breedte 0.1 daarentegen geeft te veel detailinformatie. Gebruik je ogen en zie wat goed is.







Een box-plot met uitschieters samen met een geschikt histogram geven goede visuele informatie. De klacht was terecht. De verdeling is niet symmetrisch. Er zijn te veel versterkers met een versterkingsfactor kleiner dan 10 dB.

Merk op dat de klassieke steekproefstandaardafwijking *s* hier niet zo betekenisvol is. Ze wordt sterk beïnvloed door de twee uitschieters en één getal kan onmogelijk de spreiding links en rechts t.o.v. het gemiddelde vertegenwoordigen bij een asymmetrische verdeling. Het rekenkundig gemiddelde en *s* zijn wel goede kengetallen voor symmetrische verdelingen, zeker bij normale verdelingen waar dan de 68-95-99.7-regel van toepassing is (zie hoofdstuk 6).

4.8 Opdrachten

1. Bepaal het gemiddelde, de mediaan, de standaardafwijking, de kwartielen en de extreme waarden van de onderstaande data (in cm) i.v.m. de lengte van kinderen bij hun geboorte.

53	53	51	53	51	54	48	43	48	53
53	51	52	53	53	53	56	53	53	53
52	51	53	53	55	53	53	52	53	55

2. De volgende tabel bevat de data van de leeftijd van de moeders bij de geboorte van de kinderen uit opdracht 1. Bepaal een histogram van de leeftijd van de moeders bij de geboorte van hun kind.

21	34	37	23	22	29	25	29	27	22
29	33	27	30	24	26	32	28	26	25
23	26	29	26	41	19	27	22	20	32

   Bepaal een histogram waarmee je een frequentietabel kan opstellen per leeftijd.

3. Maak een box-plot van de gegevens over de lengte van de kinderen uit opdracht 1 als ze tien jaar worden. Deze gegevens (in cm) vind je hieronder.

120	157	141	132	145	142	128	121	130	128
128	142	121	133	142	144	146	128	127	124
139	138	128	137	136	133	141	136	133	136

   Bepaal met de **TRACE**-functie een vijf-getallensamenvatting (min, Q_1, mediaan, Q_3, max).

4. De geboortegewichten (in kg) van de kinderen uit opdracht 1 zijn :

JONGENS

3.58	3.76	4.31	3.22	3.99	3.76	3.13
3.13	3.76	2.77	4.08	3.86	3.36	3.76

MEISJES

3.54	2.72	3.58	3.67	3.22	4.08	3.13	3.4
3.49	4.13	3.36	3.22	3.22	4.58	3.04	1.63

   Voer een statistische analyse uit per geslacht voor deze geboortegewichten. Maak o.a. op één scherm twee box-plots met uitschieters, opgesplitst per geslacht.

5. De gegevens i.v.m. de schoenmaat van 100 personen vind je hieronder.

34	37	38	39	40	41	42	43	44	45	48
1	3	5	5	9	20	24	16	14	2	1

Bepaal de frequentiepolygoon van deze gegevens.

6. Beschouw de 5 data 1, 2, 3, 4, 1000. Bestudeer de invloed van het weglaten van de extreme waarde 1000 op het rekenkundig gemiddelde en de mediaan.

Teken een box-plot met uitschieters. Is 1000 hier een uitschieter ? Verklaar.

7. We meten de lengte van 27 personen waaronder 15 mannen en 12 vrouwen.

MANNEN					**VROUWEN**			
172	183	180	174	173	160	168	167	162
179	172	179	168	175	198	173	167	172
176	177	170	178	181	176	168	168	169

Maak in één venster 3 box-plots met uitschieters voor de vrouwen, de mannen en de volledige groep van 27 personen.

Bij nader onderzoek blijkt dat de lengte van de dame van 198 cm verkeerd werd genoteerd. Het moest 168 cm zijn. Bekijk de box-plots opnieuw met de juiste waarde.

Teken ook een histogram van de ganse groep en bestudeer de invloed van de klassenbreedte.

8. Gegeven een populatie van data $x_1, x_2, x_3, \ldots, x_n$ met gemiddelde μ en standaardafwijking σ.

Bewijs dat

(a) $\sum_{i=1}^{n} (x_i - \mu) = 0$

(b) $\sigma^2 = \frac{\sum_{i=1}^{n} x_i^{\,2}}{n} - \mu^2$

Dit is een handige formule voor het berekenen van de variantie.

(c) Controleer (a) en (b) aan de hand van concrete data.

9. Het onderstaande programma voor de **TI-83** genereert een frequentietabel van de data die zich bevinden in lijst **L**1.

   De *verschillende* data komen terecht in **L**2 en hun frequenties in lijst **L**3.

```
PROGRAM:FREQTAB
SortA(L )
ClrList L,,Lf
1üI:1üJ:1üT
While I÷dim(L ):L (I)üL,(J)
 While L (I)=L (min({I+1,dim(L )})) and I<dim(L )
 I+1üI
 T+1üT
 End
TüLf(J)
J+1üJ
1üT
I+1üI
End
```

   a) Bepaal met dit programma een frequentietabel van de data in paragraaf 4.4.

   b) Simuleer 300 worpen met een dobbelsteen en bepaal de frequentietabel.