oktober 1991

experimentele versie

W 12 16

Freudenthal instituut

Histogram en boxplot

Leerlingentekst

Bij de omslagfoto:
Een levend histogram van studenten geordend naar lengte.
Het histogram heeft twee toppen.

Publikatie van het team W12-16
onder verantwoordelijkheid van de
Commissie Ontwikkeling Wiskundeonderwijs

Ontwerp: Heleen Verhage

© Freudenthal Instituut, RU Utrecht / SLO Enschede, oktober 1991

Toelichting bij het pakket Histogram en Boxplot

Het pakket *Histogram en Boxplot* is bedoeld voor klas 4 mavo, D-niveau. Het pakket geeft een mogelijke invulling van het onderdeel Boxplot uit het Trajectenboek (zie pag. 124 aldaar). Hiermee vervangt dit pakket Hoofdstuk 3 ('De boxplot') uit het pakket Statistiek voor de derde klas (versie november 1989). Omdat het onderwerp Boxplot in de vierde klas geplaatst is en alleen voor D-niveau, leek het handiger er een apart pakketje van te maken. De nieuwe uitwerking heeft naar ik hoop op enkele punten aan inhoud gewonnen. Aanleiding en inspiratiebron voor verbetering waren:
- observaties bij Astrid Ordelmans op Lunetten: het bleek dat de aanloop om tot het boxplot te komen wel erg compact was. Een rustiger opbouw in meer stappen leek beter.
- een verdere doordenking van wanneer boxplots zinvol zijn: vooral bij het vergelijken van groepen. In het pakket zou hier meer aandacht aan besteed moeten worden.

De opbouw van het pakket
Door de eerste helft van het pakket loopt als rode draad de context van de marathonloop. Eerst om enige oude bekenden op te halen (centrummaten, frekwentietabel en histogram), daarna om via de puntenband tot de boxplot te komen. En passant komen enkele aspecten van het histogram nader aan de orde: de keuze van klassebreedte en klassengrenzen(de paragrafen Frekwentietabel en histogram, Meer histogrammen; opg. 8 tm 13); de oppervlakte als maat van frekwenties (opg. 10 en 14).

Een nieuw idee is, om de boxplot te introduceren via de zogenaamde puntenband. De waarnemingen worden geplot op de juiste schaal boven een getallenlijn en dat plaatje wordt in kwarten verdeeld. Zo worden op een heel visuele manier de vijf boxplotgetallen gevonden (Alle punten op een rij, De boxplot, opg. 15 tm 23). Hopelijk is dit voor leerlingen een concretere aanpak dan die via mediaan en kwartielen en komt het misverstand om de mediaan op de getallenlijn precies tussen de laagste en de hoogste waarneming te willen tekenen nu minder voor.
Verder is van belang dat leerlingen de 25% stukken van de boxplot in redeneringen leren gebruiken. Dit wordt alvast aangezet in opgave 21b, maar vooral als het om het vergelijken van groepen gaat, is dit een belangrijk aspect (Salarissen, opg 24 tm 28).

Nadat de leerlingen op deze manier enige feeling voor de boxplot ontwikkeld hebben, komt een technisch detail aan de orde: het bepalen van de kwartielen. In het pakket is dit beperkt gebleven tot het geval van een even aantal waarnemingen, omdat dan het splitsen in twee helften makkelijk gaat. Bij een oneven aantal waarnemingen is de handigste oplossing om de mediaan zowel bij de eerste als bij de tweede helft te schrijven. (Anders is vervelend interpoleren onvermijdelijk.)

Vervolgens twee toepassingen over temperaturen (pag 11 en 12). Eerst een open opdracht om twee maanden met elkaar te vergelijken (mogelijke uitwerkingen: boxplots, maar ook stamblad-diagram, lijngrafieken, histogrammen), daarna twaalf boxplots in één plaatje verenigd. Hier blijkt echt de kracht van deze voorstellingswijze. In één oogopslag heeft de lezer een indruk van centrum en spreiding van de temperaturen in twaalf maanden.

De contextloze pagina 'Histogram en boxplot vergeleken' heeft iets van een puzzel. In feite gaat het erom dat leerlingen de karakteristieken (eventuele symmetrie, scheefheid, spreiding) van histogram en boxplot met elkaar in verband kunnen brengen. Het matchen op grond van symmetrie (a en d) gaat makkelijk. Bij de a-symmetrische figuren hebben de leerlingen in eerste instantie de neiging om het net verkeerd om te doen. Ze matchen bijvoorbeeld in eerste instantie c met 1, met impliciet de (foute!) redenering 'korte staart in boxplot = weinig waarnemingen = lage staaf in histogram' en 'lange staart in boxplot = veel waarnemingen = hoge staaf in histogram'. De volgende pagina gaat hier verder op door. Belangrijk is het inzicht dat bij het histogram de oppervlakte maat is voor de waarnemingen (eerder in 't pakket expliciet aan de orde geweest), terwijl de boxplot eigenlijk een verkorte en globale visualisatie van een puntenband is. De oppervlakte van de box heeft geen enkele betekenis!

Echte denk- en redeneeropgaven staan er op pagina 15 en 16. Bij 'Nogmaals salarissen' gaat het om de vraag of de boxplots voor twee deelgroepen (mannen en vrouwen apart) gecombineerd kunnen worden tot een boxplot voor de hele groep. De opgaven bij 'Inkomsten van scholieren' laten zien hoe uit een cumulatief polygoon (dit woord wordt in 't pakket niet genoemd) een boxplot geconstrueerd kan worden. Tot besluit van het pakket een waar gebeurd verhaal, zonder opgaven. Misschien een idee om dit op pakkende wijze te vertellen in de klas. Statistiek, waaronder het visualiseren van waarnemingen, maakt soms dingen zichtbaar die anders verborgen zouden blijven!

Een mogelijke indeling over de lessen, uitgaande van de vijf lessen die het Trajectenboek noemt:

Les 1:	pag 1 tm 5	(tm opg 14)	Marathonloop Frekwentietabel en histogram Meer histogrammen
Les 2:	pag 5 tm 8	(tm opg 23)	Alle punten op een rij De boxplot
Les 3:	pag 9 tm 11	(tm opg33)	Salarissen Kwartielen bepalen Temperaturen vergelijken
Les 4:	pag 12 tm 14	(tm opg 41)	Een heel jaar temperaturen Histogram en boxplot vergeleken Symmetrisch of scheef
Les 5:	pag 15 tm 18	(tm opg 47)	Nogmaals salarissen Inkomsten van scholieren Waar gebeurd verhaal

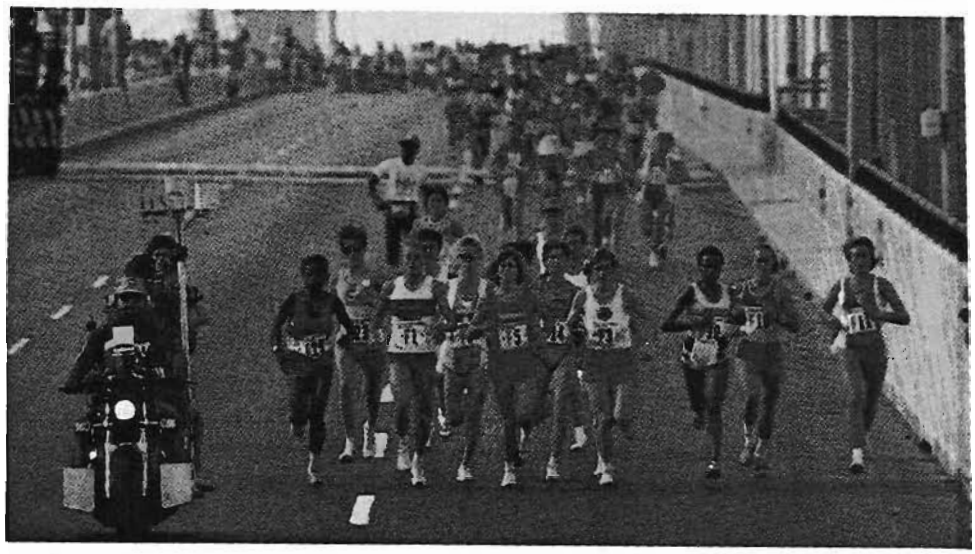
Heleen Verhage,
Utrecht, 14 oktober 1991.

Inleiding

Statisiek gaat over het verwerken van gegevens en het trekken van conclusies daaruit. In dit boekje herhalen we enkele bekende begrippen zoals gemiddelde, modus en mediaan. Ook kom je de de frekwentietabel en het histogram weer tegen. Nieuw zijn de *puntenband* en de *boxplot*, twee bijzondere plaatjes. Bij het tekenen van een box-plot speelt de mediaan een belangrijke rol.

Marathonloop

Een marathon is een hardloopwedstrijd over ruim 42 km. Om die te kunnen lopen moet je goed getraind zijn!



1. Hoe lang zou jij doen over 42 km lopen? En op de fiets?

Bij een bepaalde marathon hebben de deelnemers de volgende tijden gelopen:

2.26	2.31	2.33	2.37	2.44	2.48	3.01	3.03	3.07	3.09
3.15	3.15	3.16	3.18	3.20	3.25	3.29	3.35	3.45	3.45
3.47	3.50	3.56	3.57	3.58	3.59	4.05	4.09	4.15	4.15
4.25	4.25	4.27	4.29	4.33	4.37	4.42	4.42	4.42	5.01
5.09	5.17	5.31	5.46	6.32					

2. a Hoeveel deelnemers waren er?
   b De winnaar liep de marathon in 2 uur en 26 minuten. Wat was zijn gemiddelde snelheid?
   c Hoelang deed de hekkesluiter (= degene die het laatste aan kwam) over de marathon? Hoeveel km per uur is dat gemiddeld?

3. Op een zeker moment is de helft van de lopers binnen.

a Welke tijd heeft de loper die op dat moment over de finish gaat?

Waarnemingen worden vaak geordend van klein naar groot. Bij de looptijden is de volgorde van binnenkomst de ordening. De middelste waarneming heeft een speciale naam, deze wordt de *mediaan* genoemd.

4. Wat is de mediaan van de marathontijden?

5. Er waren bij deze wedstrijd zes deelnemers van de sportclub Grande Vitesse. Hun tijden waren:

2.37 3.03 3.25 3.58 4.37 5.17

a Wat was de gemiddelde tijd van deze zes lopers?

Bij een oneven aantal waarnemingen is de *mediaan* de middelste waarneming. Als het aantal waarnemingen even is, zoals bij lopers van Grande Vitesse, is dit geen goede afspraak. Er is dan immers niet één middelste waarneming. In dat geval is de afspraak om naar de middelste *twee* waarnemingen te kijken. De mediaan is dan het gemiddelde van die twee getallen.

b Wat is de mediaan van de tijden van de lopers van Grande Vitesse?

6. De gemiddelde looptijd van *alle* 45 deelnemers aan de marathon is 3 uur en 59 minuten. Ga dat na met een berekening en probeer die overzichtelijk op te schrijven.

TIP: Doe deze opgave niet alleen, maar met z'n tweeën of zelfs met z'n vieren. Je kunt dan het werk verdelen. Bedenk voordat je aan het rekenen slaat eerst hoe je het aan wilt pakken!

Mediaan en gemiddelde worden wel *centrummaten* genoemd. Ze geven iets aan over het centrum (midden) van een serie waarnemingen.
Er is nog een derde centrummaat, de *modus*. Dat is de waarneming die het vaakst voorkomt.

---

7. Wat is de modus van de marathontijden?
Zegt die inderdaad iets over het centrum? Waarom wel/niet?

Frekwentietabel en histogram

In de vorige paragraaf is de uitslag van een marathonwedstrijd getypeerd met behulp van centrummaten. Maar er zijn meer manieren om naar de uitslag te kijken.
Je kunt de gegevens van de marathonwedstrijd ook verwerken in een *frekwentietabel*. Daar kun je dan vervolgens een *histogram* bij tekenen.
De waarnemingen worden dan ingedeeld in *klassen*. Dat heet het maken van een *klasse-indeling*.

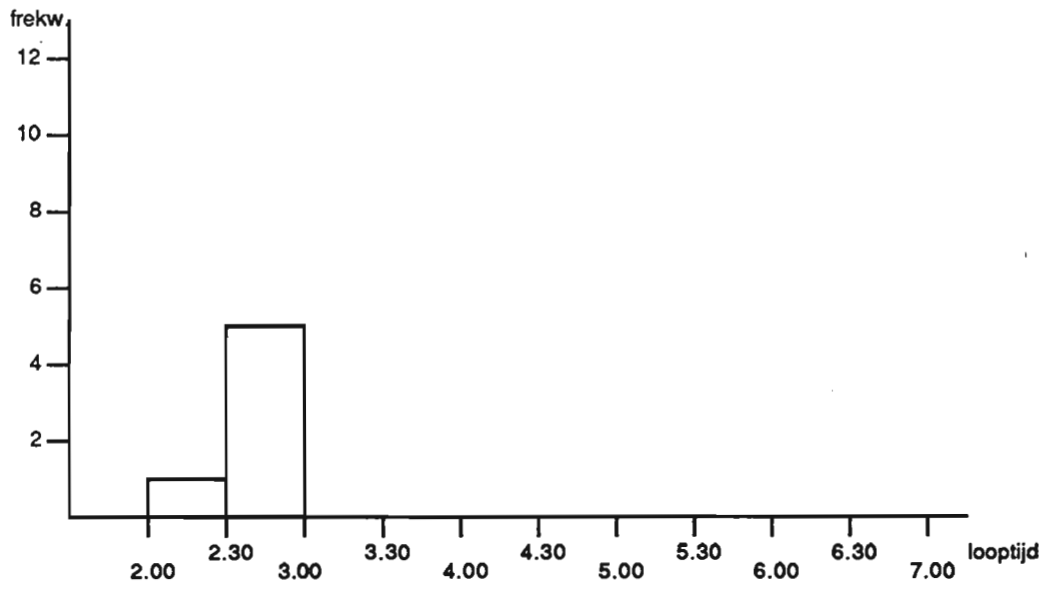
8.a Hoeveel klassen krijg je als de klassebreedte 30 minuten is en de laagste klasse begint bij 2.00 uur?

b Maak (in je schrift) de frekwentietabel af:

tijd	aantal lopers
2.00 - 2.29	1
2.30 - 2.59	5
....	

c In welke klasse zitten de meeste lopers?

Bij een frekwentieverdeling heet de klasse met het hoogste aantal waarnemingen de *modale klasse*. Als je de frekwentietabel eenmaal hebt, kun je er een histogram bij tekenen. Hieronder is alvast een begin gemaakt:



9. a Maak het histogram verder af.

b Hoe lees je uit het histogram de modale klasse af?

10.a Geef met een kruisje op de horizontale as aan waar de mediaan ligt.Teken ook een verticale lijn door dit punt. Het histogram is nu in twee stukken verdeeld.

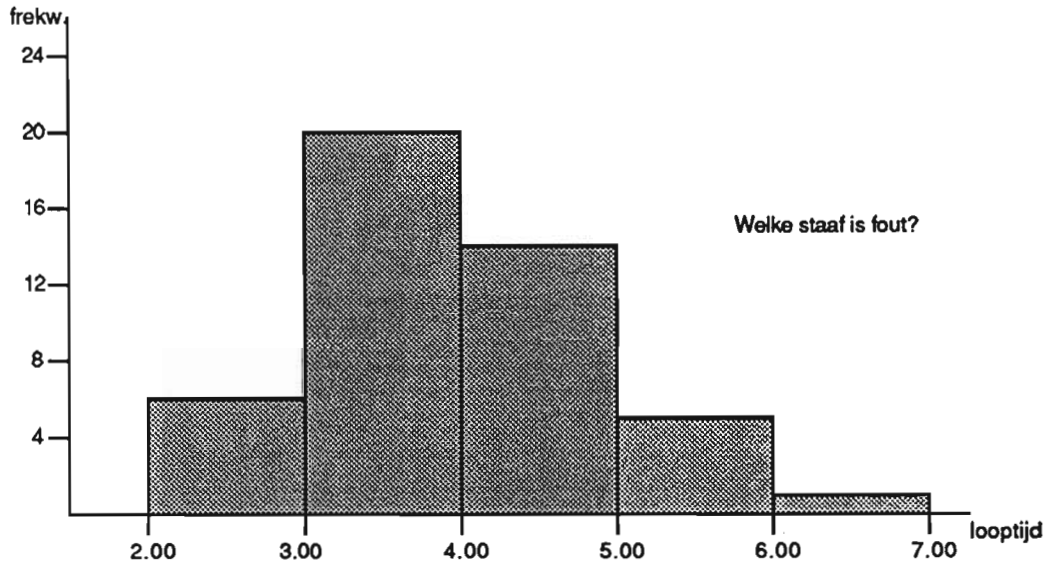
b Wat kun je zeggen over de grootte van deze twee stukken?

Meer histogrammen

Bij de tabel en het histogram is een klassebreedte van 30 minuten als uitgangspunt genomen. De laagste klasse begon bij 2.00 uur. Dit kan ook anders.

11. Neem een klassebreedte van 30 min. en laat de laagste klasse beginnen bij 2.15 uur.
    a Hoeveel klassen krijg je dan?
    b Maak de frekwentietabel en teken het histogram.
    c Wat is nu de modale klasse?
    d Wijkt dit histogram erg af van het histogram van opgave 9?

12. Het histogram hieronder hoort bij de marathontijden, maar nu is een klassebreedte van 60 min genomen. Helaas heeft één staaf de verkeerde hoogte gekregen. Welke staaf is dat? Verbeter die.

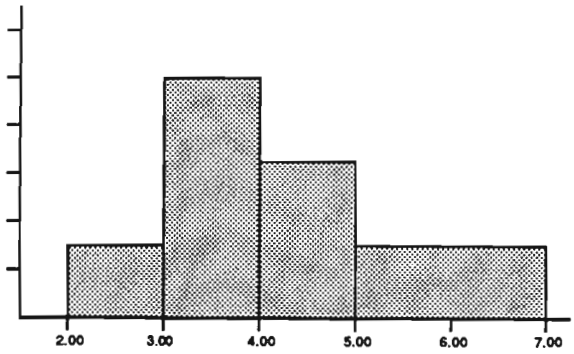


13. Vergelijk de histogrammen van opgave 9 en opgave 12 met elkaar.
    De twee histogrammen hebben op de horizontale as dezelfde schaalverdeling.
    De schaalverdeling op de verticale as is echter *niet* hetzelfde.
    a Wat is het verschil? Waarom zou dat gedaan zijn?
    b Welk histogram vind je duidelijker, het histogram met klassebreedte van 30 min of het histogram met een klassebreedte van 60 min? Waarom?

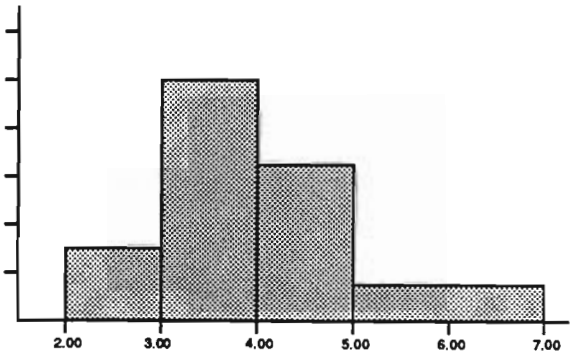
Bij één serie waarnemingen zijn dus verschillende histogrammen mogelijk.
Als je een histogram tekent, moet je eerst beslissingen nemen over:
- de klassebreedte, rekening houdend met het aantal klassen dat je dan krijgt,
- de klassengrenzen, in het bijzonder de ondergrens van de eerste klasse.
Hoe meer klassen het histogram heeft, hoe gedetailleerder informatie je eruit kunt aflezen. Vanwege de overzichtelijkheid zal een histogram meestal tussen de 5 en de 10 klassen hebben. Een enkele keer kan het aantal klassen groter zijn, maar meer dan 20 klassen zul je niet vaak tegen komen.

Soms kom je wel eens een histogram tegen waarbij de klassen niet allemaal even breed zijn. Hieronder staan er twee, weer bij de marathontijden.

a.



b.



14 a Vergelijk de twee histogrammen met elkaar. Wat is het verschil?

b Schrijf in de staven van de histogrammen het bijbehorende aantal waarnemingen.

c Welk histogram vind je het beste? Waarom?

Bij histogrammen met ongelijke klassebreedtes is de afspraak dat de *oppervlakte* van de staven de maat is voor het aantal waarnemingen. Het is dan niet meer mogelijk frekwenties bij de verticale as te zetten. Je kunt de frekwenties wel noteren in of vlak boven de staven.

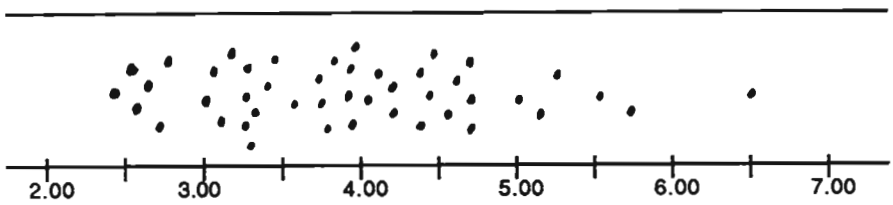
d Welk histogram is volgens deze afspraak het goede? Had jij dat ook bij c?

Alle punten op een rij

Een goed gekozen histogram geeft een duidelijk beeld van een serie waarnemingen.
Maar: de losse waarnemingen zijn er niet meer precies in terug te vinden.
In deze paragraaf komt een plaatje aan bod waarin wél alle waarnemingen afzonderlijk staan.
Nogmaals de tijden van de 45 marathonlopers:

2.26	2.31	2.33	2.37	2.44	2.48	3.01	3.03	3.07	3.09
3.15	3.15	3.16	3.18	3.20	3.25	3.29	3.35	3.45	3.45
3.47	3.50	3.56	3.57	3.58	3.59	4.05	4.09	4.15	4.15
4.25	4.25	4.27	4.29	4.33	4.37	4.42	4.42	4.42	5.01
5.09	5.17	5.31	5.46	6.32					

In het volgende plaatje staan alle looptijden van de marathonwedstrijd ingetekend:



Zo'n plaatje heet wel een *puntenband.*
De punten zijn op verschillende hoogtes getekend om ze niet over elkaar heen te laten vallen. Verder heeft die hoogte geen betekenis.

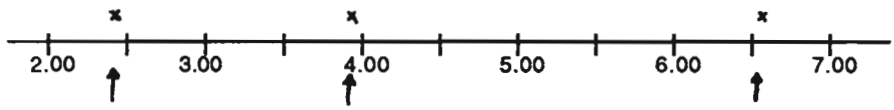
15.a Zoek in de puntenband de snelste, de langzaamste en de middelste loper op. Zet een rondje om hun looptijd.

b Hoeveel lopers waren er volgens het plaatje tussen de drie en vier uur onderweg? Klopt je antwoord met de oorspronkelijke gegevens?

16.a Na hoeveel tijd is een kwart van de lopers binnen?

b Na hoeveel tijd is driekwart van de lopers binnen?

De drie omcirkelde tijden (laagste waarneming, hoogste waarneming en mediaan) geven samen al een aardig beeld van de wedstrijd.
Je kunt er een plaatje van maken door ze op een 'looptijdenlijn' te plaatsen, zo:



17. Iemand zegt: "Hè? De middelste waarneming moet toch midden tussen de kleinste en de grootste liggen?"
Probeer uit te leggen hoe dit zit.

Oefenopgave

18. Een rijtje proefwerkcijfers:
6.8, 8.2, 5.5, 7.4, 6.5, 5.2, 7.3, 9.1, 8.4, 4.9, 6.7,
7.4, 6.6, 6.3, 7.8, 9.6, 4.1, 5.3, 7.5, 6.6, 5.7, 4.5

a Teken een puntenband van de proefwerkcijfers.

b Zet de cijfers op volgorde van grootte.

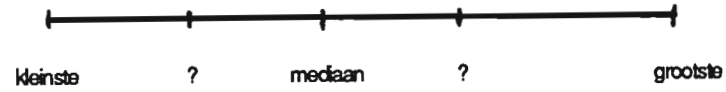
c Teken het laagste cijfer, het hoogste cijfer en de mediaan op een getallenlijn.

d Bereken het gemiddelde. Is het gemiddelde groter of kleiner dan de mediaan?

De boxplot

In de statistiek zoekt men naar simpele plaatjes die veel informatie geven.
Een puntenband geeft veel informatie, maar is nogal bewerkelijk om te tekenen.
Het lijnstukje kleinste-mediaan-grootste is snel getekend, maar bevat minder informatie.

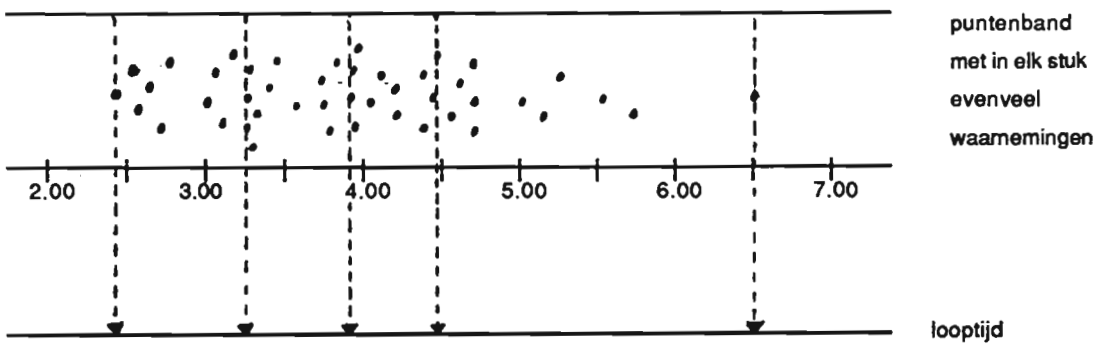
Als uitbreiding op het lijnstukje met drie punten, is een plaatje met vijf punten bedacht. Om te beginnen het drietal kleinste - mediaan - grootste, en dan nog twee andere. Zo iets:



19. Heb je al een vermoeden welke die andere twee punten zullen zijn?

De afspraak is om de reeks waarnemingen in vier stukken te verdelen, met in elk stuk evenveel getallen. We nemen de puntenband van de marathonlooptijden om te kijken hoe dit werkt.

- deel de puntenband in vieren, met in elk stuk evenveel waarnemingen:



- markeer de vijf getallen die je zo vind op de getallenlijn.

Er zijn dus vijf getallen van belang:

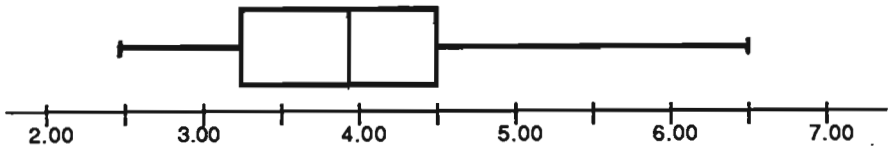
- de snelste tijd	:	2:26	(kleinste)
- de tijd waarop een kwart binnen is	:	3:15	(1e kwartiel)
- de middelste tijd	:	3:56	(de mediaan)
- de tijd waarop driekwart binnen is	:	4:29	(3e kwartiel)
- de langzaamste tijd	:	6:32	(grootste)

20.a Je ziet hier boven (in de rechter kolom) een nieuw woord staan. Welk woord is dat en wat zal het betekenen?

b Het 2e kwartiel bestaat ook. Wat zal dat zijn?

Niet vergissen: een *kwartiel* is één getal, dus niet een interval of zo.

Bij de vijf getallen wordt de volgende tekening gemaakt:

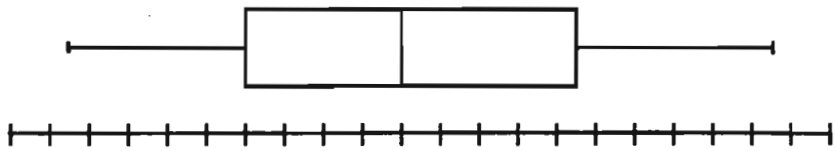


Zo'n tekening heet een *boxplot*. (De getallenlijn hoort er ook bij, anders kun je de boxplot immers niet aflezen.)

21. Aan de marathonwedstrijd deden 45 lopers mee.
   a Wat kun je als je alleen de boxplot hebt, zeggen over de looptijd van de loper die als 15e binnenkwam?
   b Verzin een uitspraak met: '50% van de lopers heeft een tijd gelopen ....'

22. a Kijk nog eens terug naar vraag 15a. Kun je die vraag beantwoorden met alléén de boxplot?
   b Kun je vraag 15b beantwoorden met alléén de boxplot?
   c Kun je vraag 16 beantwoorden met alléén de boxplot?

De algemene vorm van een boxplot is:



Een boxplot bestaat eigenlijk uit vier stukken:

streep halve doos halve doos streep

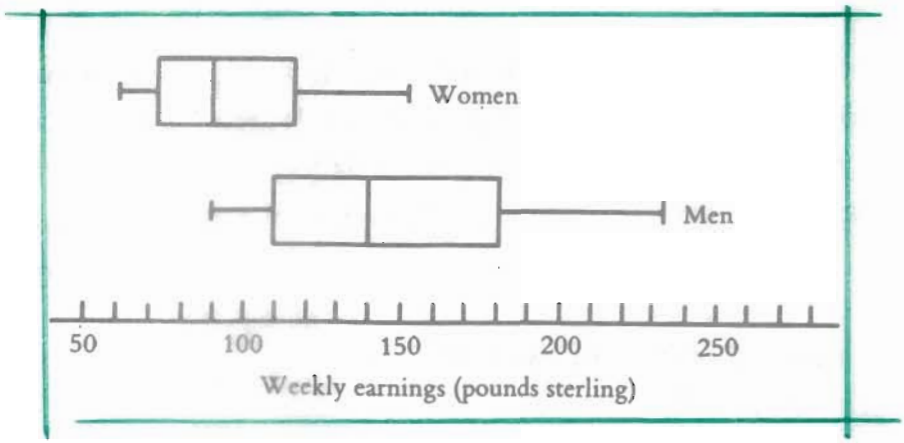
Elk stuk bevat een kwart (25%) van de waarnemingen. In de doos in het midden zit dus 50% van de waarnemingen.

---

23. Teken een box-plot bij de getallen:
kleinste 12, 1e kwartiel 15, mediaan 21, 3e kwartiel 23, grootste 40.

Salarissen

Boxplots worden vooral gebruikt om twee of meer *groepen met elkaar te kunnen vergelijken*. Bijvoorbeeld de inkomens van werkende mannen en vrouwen in Engeland:



De bedragen zijn per week en in pound sterling (Engelse ponden).
In de krant kun je opzoeken hoeveel een Engels pond ongeveer waard is.

24. Vul de juiste getallen in:
    a "Een Engelse vrouw verdient maximaal .... pond per week."
    b "Een Engelse man verdient ten minste .... pond per week."

25. Kijk naar de middelste 50% van beide groepen. Wat zijn de salarisgrenzen van die groepen?

26. Hieronder staan drie beweringen. Geef op elke bewering commentaar, uitgaande van de box-plots die hierboven gegeven zijn.

    a Bewering A: 'alle vrouwen verdienen in Engeland minder dan de mannen'
    b Bewering B: 'vrouwen verdienen in Engeland in het algemeen minder dan mannen'
    c Bewering C: 'vrouwen worden slechter betaald dan mannen'

27. Kloppen de volgende uitspraken met de box-plot?
    a. "50% van de mannen verdient meer dan het maximum weeksalaris van de vrouwen."
    b. Vrijwel alle mannen verdienen meer dan de 50% laagst betaalde vrouwen."

28. Kun je zelf nog twee uitspraken bij de box-plots verzinnen?

Kwartielen bepalen

Nogmaals het rijtje proefwerkcijfers (zie opgave 18), nu op volgorde gezet:

4.1, 4.5, 4.9, 5.2, 5.3, 5.5, 5.7, 6.3, 6.5, 6.6, 6.6,
6.7, 6.8, 7.3, 7.4, 7.4, 7.5, 7.8, 8.2, 8.4, 9.1, 9.6

Je hebt hier al een puntenband van getekend.Nu richten we ons op het tekenen van de boxplot bij deze cijfers. Daarvoor heb je de vijf boxplot-getallen nodig. Drie getallen heb je zo te pakken: kleinste 4.1; mediaan 6.65; grootste 9.6
Blijft over het bepalen van 1e en 3e kwartiel. Die kun je uit de gegevens vinden door te bedenken:

- *1e kwartiel = de mediaan van de eerste helft*
- *3e kwartiel = de mediaan van de tweede helft*

eerste helft 4.1, 4.5, 4.9, 5.2, 5.3, 5.5, 5.7, 6.3, 6.5, 6.6, 6.6,

---

tweede helft 6.7, 6.8, 7.3, 7.4, 7.4, 7.5, 7.8, 8.2, 8.4, 9.1, 9.6

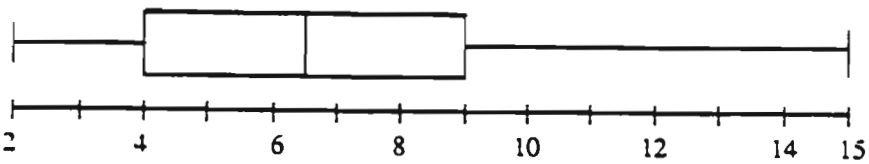
29.a Bepaal de kwartielen.

b Teken de boxplot.

c Probeer drie uitspraken te verzinnen met:
'50% van de leerlingen heeft een proefwerkcijfer .....'

30. Een rijtje gegevens:
15, 7, 11, 3, 9, 3, 10, 5, 2, 7, 3, 8, 6, 4, 6, 7, 5, 2
Wat is het eerste kwartiel, de mediaan, het derde kwartiel?

31. Kan de onderstaande boxplot bij de gegevens van opgave 30 horen?



32. Verzin een rijtje getallen waarvoor deze box-plot goed is.

Temperaturen vergelijken

In twee opeenvolgende jaren registreerde het KNMI de volgende temperaturen in de maand mei:

dagnr	temp '82	temp '83	dagnr	temp '82	temp '83
1	10.1	15.7	17	23.6	16.3
2	11.6	12.5	18	19.4	16.5
3	13.5	10.1	19	19.3	15.8
4	12.6	13.8	20	17.9	17.0
5	11.1	15.8	21	20.2	12.7
6	9.1	17.5	22	18.1	15.0
7	10.6	16.0	23	15.5	16.9
8	11.3	14.7	24	16.9	11.3
9	14.4	15.9	25	20.0	10.8
10	16.0	14.2	26	26.3	10.1
11	17.2	11.1	27	24.9	10.8
12	18.8	13.4	28	18.0	12.4
13	19.6	16.0	29	20.5	12.3
14	22.8	16.2	30	23.8	16.2
15	25.6	16.1	31	27.2	22.8
16	24.3	13.1			

33. Probeer zoveel mogelijk verschillende manieren te bedenken om de twee meimaanden met elkaar te vergelijken en maak daar een mooie bladzijde in je schrift van.

Zijn dit ook boxplots?

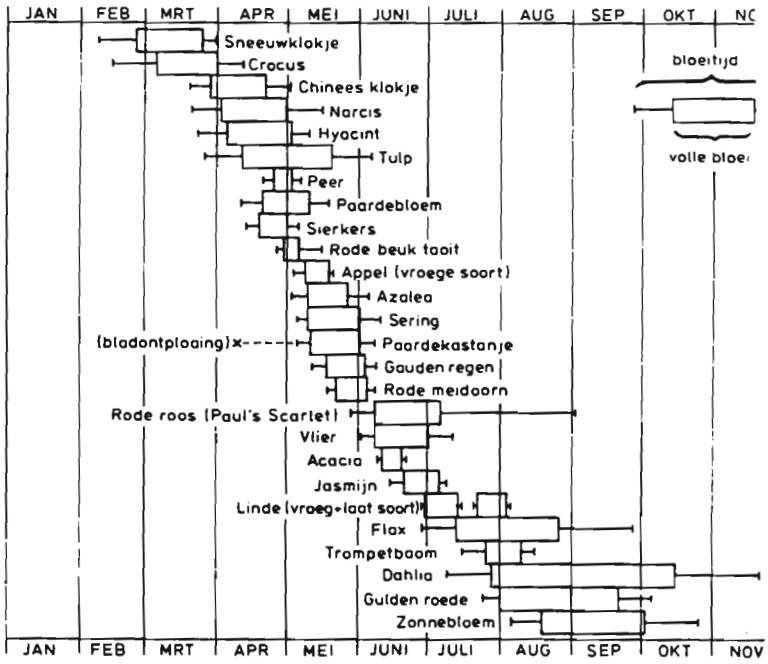
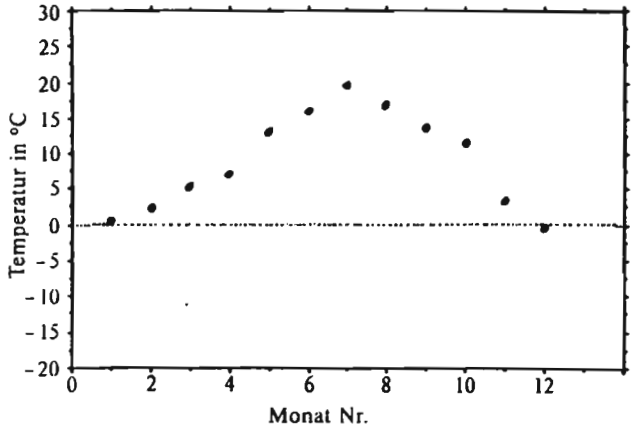


Fig. 67. Gemiddelde bloeiperioden van diverse plantesoorten voor De Bilt, zoals waargenomen door G. W. Th. M. de Bont over een periode van 25 jaar. Van jaar op jaar kunnen de bloeidata twee weken eerder of later vallen dan in de grafiek aangegeven. In het algemeen voltrekt de bloei zich wel ieder jaar in dezelfde volgorde.

Een heel jaar temperaturen

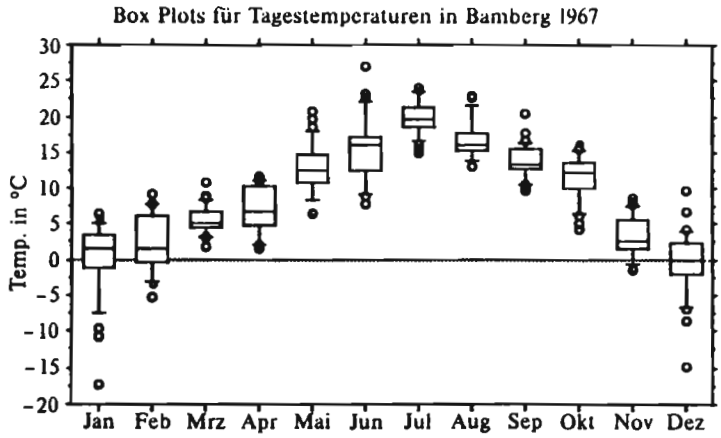
Als je een heel jaar lang elke dag de temperatuur meet, kun je daar een grafiek van tekenen. Dat kan op verschillende manieren. Hieronder zie je een grafiek van de gemiddelde maandtemperatuur in 1967 in het Duitse stadje Bamberg.



Tagestemperaturen in Bamberg 1967

34. a Welke maand was het (volgens deze grafiek) het warmst? Hoe warm was het toen?
    b En welke maand was het het koudst?
    c Kun je iets zeggen over de warmste dag van het jaar? En over de koudste dag?
    d Kun je iets zeggen over temperatuurverschillen binnen één maand?

De grafiek hieronder gaat ook over de temperatuur in Bamberg. Nu heeft men voor elke maand een boxplot getekend. De boxplot is anders getekend dan je gewend bent, want de strepen lopen niet door tot de uiteinden. In plaats daarvan heeft men de drie laagste en de drie hoogste waarnemingen apart met een rondje aangegeven.



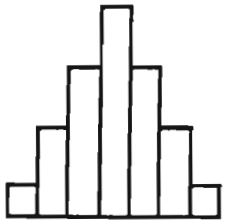
Tagestemperaturen in Bamberg 1967
Zusammenfassung nach 12 Monaten: Boxplots

35. Beantwoord de vragen van hiervoor nog een keer, maar nu aan de hand van deze grafiek.

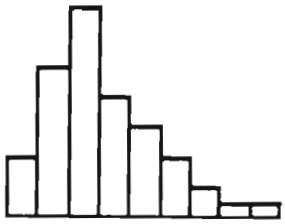
36. Vergelijk de twee grafieken eens met elkaar:
    - Welke grafiek is het makkelijkste te begrijpen?
    - Welke grafiek bevat de meeste informatie?

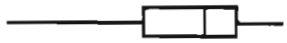
Histogram en box-plot vergeleken

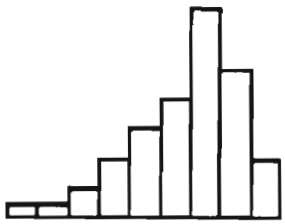
37. Hieronder staan vijf histogrammen en vijf box-plots. Ze horen twee-aan-twee bij elkaar. Welk histogram hoort bij welke box-plot?

a. 

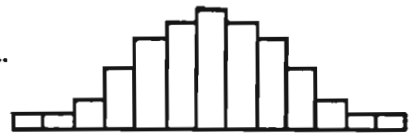
1. 

b. 

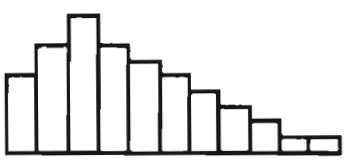
2. 

c. 

3. 

d. 

4. 

e. 

5. 

Symmetrisch of scheef

Bij het beantwoorden van de vorige vraag heb je vast en zeker gebruik gemaakt van *symmetrie* in de grafieken. Histogram symmetrisch? Dan is de boxplot die daar bij hoort ook symmetrisch!

38.a Bij een histogram kun je altijd een boxplot tekenen. Gaat het omgekeerde ook op?

b Iemand zegt: 'Hoge staven in het midden van een histogram betekent een grote doos in het midden van de bijbehorende boxplot'.
Welke denkfout wordt hier gemaakt?

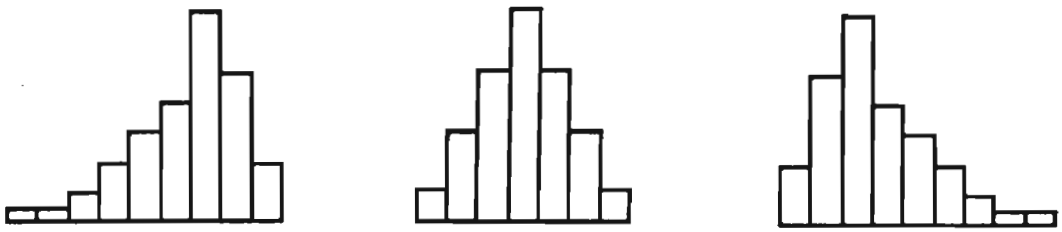
39. In het kader van een medisch experiment wordt een aantal cavia's ingespoten met een tuberculose bacil. Men kijkt na hoeveel dagen de cavia's overlijden. Na 43 dagen gaat de eerste dood en de sterkste houdt het 598 dagen vol.
De volledige gegevens zijn:

43	45	53	56	56	57	58	66	67
73	74	79	80	80	81	81	81	82
83	83	84	88	89	91	91	92	92
97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	121
123	126	128	137	138	139	144	145	147
156	162	174	178	179	184	191	198	211
214	243	249	329	380	403	511	522	598

a Maak bij deze gegevens een histogram. Kies eerst zelf een klassebreedte die je geschikt lijkt, rekening houdend met het aantal klassen je dan krijgt. Komen er ook lege klassen voor?

b Teken bij deze gegevens de boxplot.

Een histogram kan symmetrisch zijn of scheef:



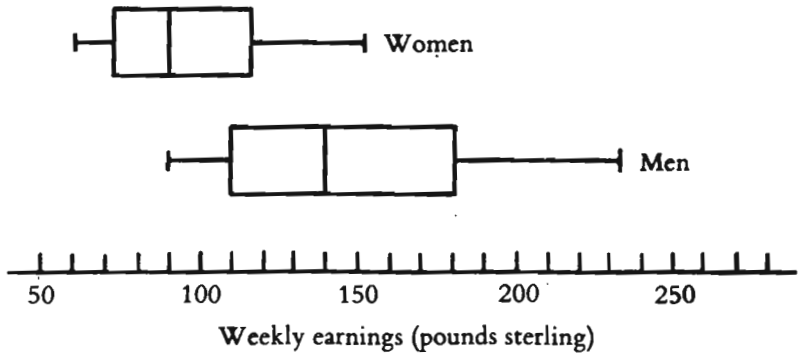
scheef naar links | symmetrisch | scheef naar rechts

40. Is de verdeling van de levensduur van de cavia's scheef naar links of scheef naar rechts?

41. Hoe kun je aan een boxplot zien of een verdeling scheef naar links of scheef naar rechts is?

Nogmaals salarissen

Hieronder staan nogmaals de boxplots van de salarissen in Engeland:

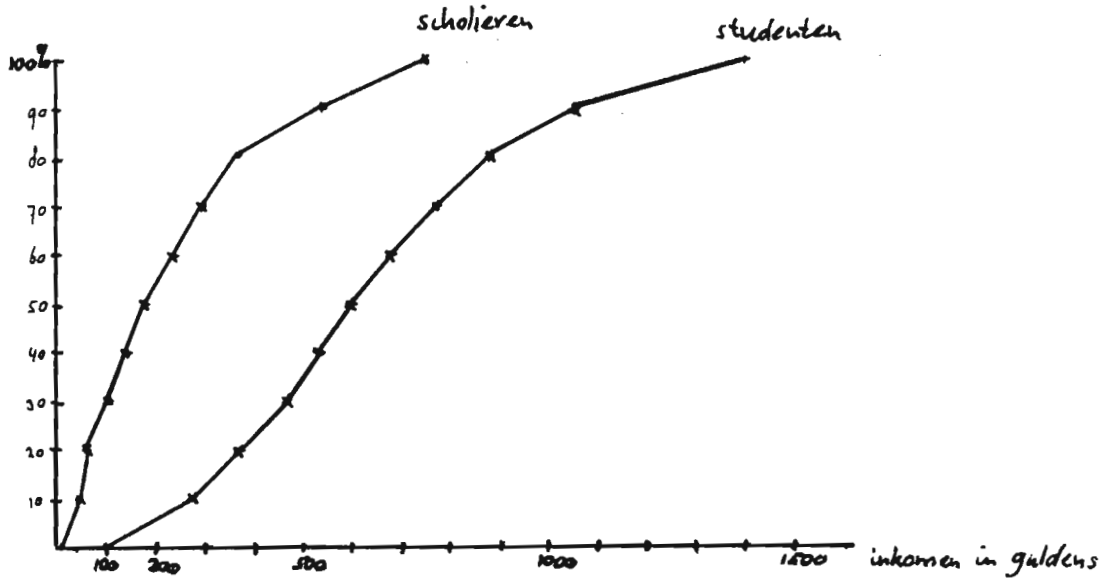


42. Wat kun je zeggen over de scheefheid van de inkomensverdelingen voor vrouwen en mannen?

43. Een lastige vraag:
Probeer eens of je een boxplot voor de mannen en vrouwen samen kunt maken!
Neem voor het gemak even aan dat de getekende boxplots betrekking hebben op 1000 vrouwen en 1000 mannen.
De volgende vragen helpen je een stukje op weg:
- wat is het verschil tussen het hoogste en het laagste salaris voor de hele groep?
- wat kun je zeggen over de mediaan van de hele groep?

Inkomsten van scholieren

Elk jaar wordt er een onderzoek gedaan onder de Nederlandse jeugd. Men kijkt dan onder andere naar de maandelijkse inkomsten van scholieren (13-17 jaar) en studenten (18-21 jaar). De grafiek hieronder brengt dat in beeld.

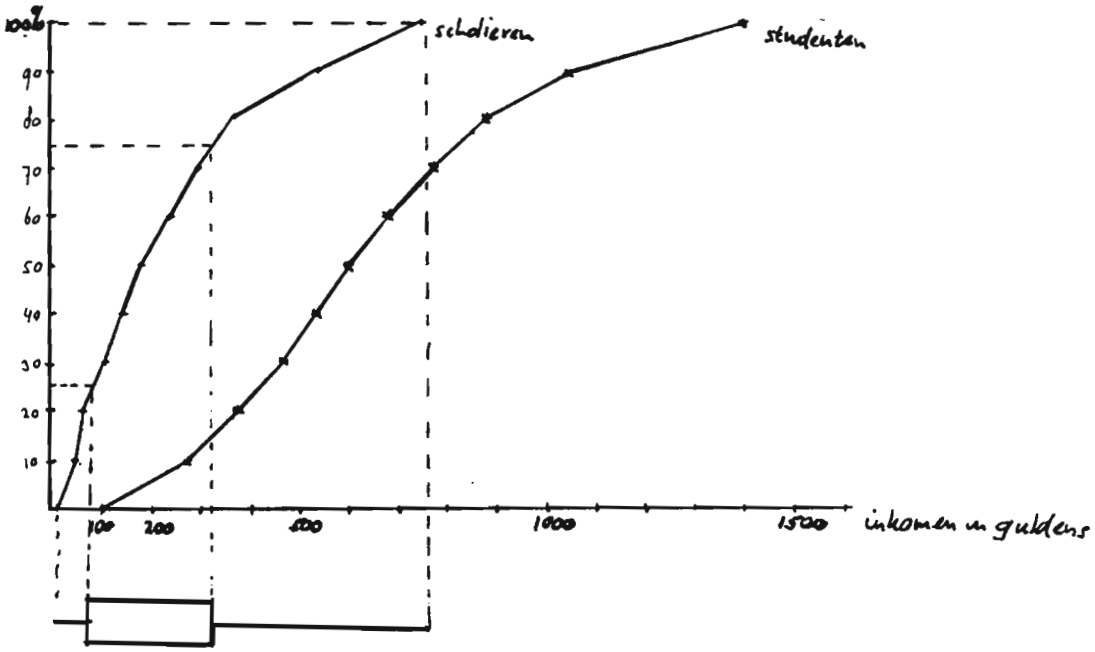


44. Een uitspraak bij deze grafiek is:
"30 procent van de scholieren heeft maandelijkse inkomsten van ten hoogste ...."

a Welk getal moet op de stippeltjes staan?

b Hoeveel procent van de studenten heeft maandelijkse inkomsten van méér dan 1000 gulden?

De inkomsten van de scholieren kunnen ook weergegeven worden in een boxplot. Die boxplot kun je afleiden uit de gegeven grafiek. Dat gaat zo:



45.a Kun je uitleggen hoe de boxplot getallen zijn gevonden?

b De mediaan ontbreekt nog, teken die zelf in.

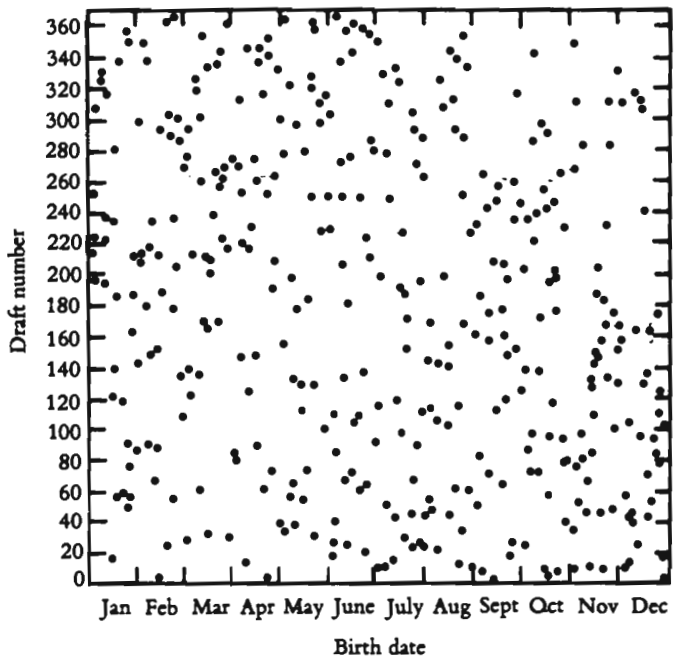
46. Maak op deze manier ook de boxplot voor de maandelijkse inkomsten van studenten.

Uit het scholierenonderzoek is ook gebleken dat het *gemiddelde* maandinkomen van scholieren 248 gulden bedraagt.

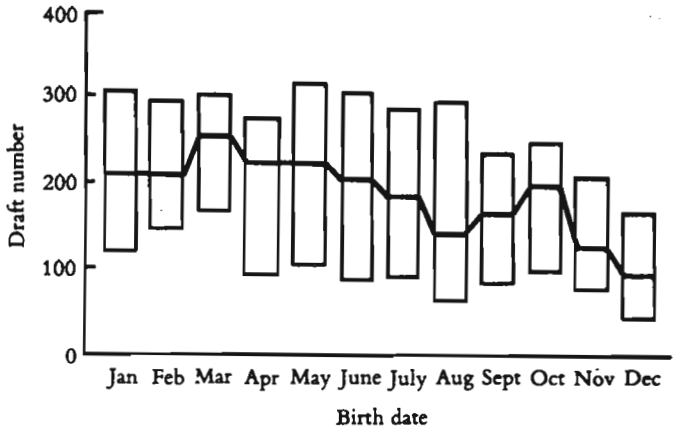
47. De zestienjarige Peter heeft maandelijks 200 gulden te besteden. Hij vindt dat zelf nogal weinig. Zijn moeder zegt dat dat heus wel meevalt.
- Bedenk een argument waarmee Peter moeder kan overtuigen.
- Bedenk een argument waarmee moeder Peter kan overtuigen.

Waar gebeurd verhaal

In 1970 werd in Amerika door middel van loting uitgemaakt wie er in militaire dienst moest. In die tijd betekende dat ook dat je als soldaat uitgezonden kon worden naar Vietnam, waar toen oorlog was. Men wilde volgens het toeval geboortedata uitkiezen. De mannen met de uitgekozen geboortedata en geboren tussen 1943 en 1952 zouden opgeroepen worden. De trekking van de geboortedata verliep als volgt. Alle dagen van het jaar werden stuk voor stuk op een briefje geschreven en de briefjes werden in een balletje gestopt. De balletjes gingen in een grote ton.
Het eerste balletje werd getrokken en alle soldaten met de geboortedatum van dit balletje kregen nummer 1. Daarna werd het volgende balletje getrokken, de soldaten met die datum kregen nummer 2. Enzovoort, tot alle 366 balletjes getrokken waren. De mannen zouden in volgorde van de nummers opgeroepen worden. Eerst de nummers 1, als er meer soldaten nodig waren de nummers 2, enzovoort. Met een laag nummer was je dus snel aan de beurt. De gelukkigen met een hoog nummer zouden waarschijnlijk helemaal niet opgeroepen worden.
In de figuur hieronder zijn de nummers uitgezet tegen de maanden.



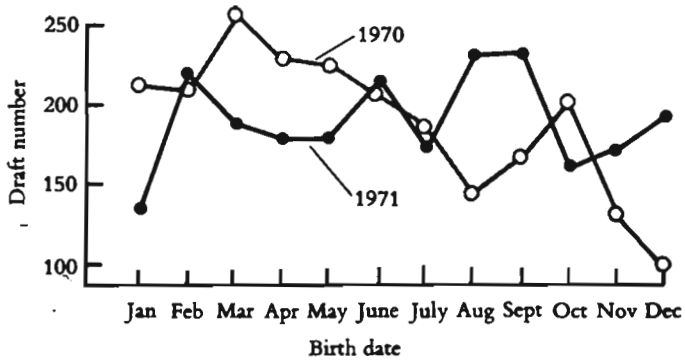
Het lijkt erop dat de nummers 1 t/m 5 terecht zijn gekomen in februari, april, september, oktober en december. De precieze volgorde is niet goed af te lezen. Ogenschijnlijk lijkt het alsof de stippen willekeurig over de maanden verdeeld zijn. Een Amerikaanse journalist heeft echter ontdekt dat de verdeling over de maanden niet louter toevallig is. December heeft teveel lage nummers gekregen.
Het bleek dat men de balletjes onvoldoende geschut had. Daardoor lagen de balletjes van december in verhouding teveel bovenop, waardoor ze een grotere kans hadden om getrokken te worden en dus lage nummers kregen. Pech voor de mannen die in december geboren waren!



Maand-medianen en -kwartielen voor de loterij van 1970

Het bovenstaande plaatje laat zien dat de data in december inderdaad lage nummers hadden. Per maand is de doos van de boxplot getekend (de strepen zijn dus weggelaten). De medianen zijn vervolgens met elkaar verbonden. Nu is duidelijk te zien dat december inderdaad veel lage nummers heeft.
Zo zie je hoe je door het kiezen van een geschikt plaatje méér informatie uit waarnemingen af kunt leiden!

Het jaar daarop, in 1971, heeft men de lotingsprocedure veranderd. Nu werden twee tonnen met balletjes gebruikt, één met de geboortedata en één met de getallen 1 t/m 366. De twee tonnen werden goed geschut. Daarna werd uit beide tonnen een balletje getrokken. De geboortedatum uit de ene ton kreeg het nummer van het balletje uit de andere ton. In de figuur hieronder staan de medianen per maand voor 1970 en 1971. De spreiding blijkt ook in 1971 vrij groot te zijn, maar het extreme resultaat van december 1970 komt niet meer voor. De loting was nu wel eerlijk.



Maand-medianen voor de loterijen van 1970 en 1971

archief FI 02.01.64

Histogram en boxplot

Leerlingentekst

Verhage, H.