Statistiek

STATISTIEK

STATISTIEK

WISKUNDE A

STATISTIEK

Een produktie ten behoeve van het project Hawex.

Ontwerpers:	Henk van der Kooy, Jan de Lange
Met medewerking van:	Christiane Hauchart Jan de Jong Martin Kindt Martin van Reeuwijk Anton Roodhardt
Vormgeving:	Ada Ritzer

Utrecht, juli 1989

Inhoudsopgave

1 Kijken en vergelijken

Van den Broek en Ruding streven Lubbers voorbij in populariteit

Van onze politieke redactie

DEN HAAG — CDA-premier Lubbers is in populariteit voorbijgestreefd door zijn partijgenoten Van den Broek en Ruding. Van de Nederlandse kiezers zegt 54 procent „veel vertrouwen" te hebben in minister Van den Broek (Buitenlandse Zaken). Minister Ruding van Financiën scoort bij 51 procent goed en Lubbers bij 49 procent. In vorige onderzoeken stond Lubbers steeds bovenaan.

Dit blijkt uit de laatste halfjaarlijkse peiling door het bureau Burke-Inter/View in opdracht van de Rijksvoorlichtingsdienst. In de top van de regeringspartijen wordt veel waarde toegekend aan deze cijfers, die een rol kunnen spelen bij toekomstige leiderskeuze.

de Volkskrant jan. 88

— De werkloosheid is gedaald met 3%.
— 49% heeft vertrouwen in Lubbers.
— Nieuwe motor is 7% zuiniger.
— 1 op de 4 leerlingen gebruikt wel eens drugs.
— Aspirine voorkomt hartinfarcten.

De media (radio, TV, tijdschriften, kranten) presenteren vaak informatie zoals hierboven. Vrijwel niemand vraagt zich af hoe je tot zulke uitspraken kunt komen. Toch heeft de manier waarop ze tot stand zijn gekomen invloed op de betrouwbaarheid.
In dit hoofdstukje bekijken we enige manieren om gegevens te verzamelen.

1. *Ik ben vooruitgegaan, dus het helpt!*
Vier leerlingen besluiten bijles te nemen in wiskunde. Ze staan gemiddeld:

ll. A: 4,4
ll. B: 4,1
ll. C: 5,3
ll. D: 3,2

Na vier weken bijles volgen twee proefwerken wiskunde. De vier leerlingen hebben de volgende resutaten:

ll. A:	6	5^-
ll. B:	$5\frac{1}{2}$	5
ll. C:	7	6^+
ll. D:	5	4

>a Vind je dat je op grond van deze gegevens kunt concluderen dat de bijlessen succesvol zijn geweest?

De ouders van de vier leerlingen en de bijlesleraar waren zeer tevreden. De vier leerlingen niet, en de lerares ook al niet. Wat was het geval?
Vóór dat de leerlingen bijles namen was het gemiddelde van de overige leerlingen ongeveer 6,1.
Het gemiddelde van de overige leerlingen bij de twee laatste proefwerken was: 7,9 en 7,7.

>b Bereken het gemiddelde van de laatste twee proefwerken zowel voor de leerlingen A, B, C en D als voor de overige leerlingen.
Hoeveel punten is ieder vooruitgegaan?
Verandert dit iets aan je antwoord op vraag >a?

>c Je zou ook kunnen kijken naar de relatieve vooruitgang (hoeveel procent is het cijfer vooruitgegaan). Bereken de relatieve vooruitgang van iedere leerling. Wat is je conclusie?

2. *Nieuw geneesmiddel tegen benauwdheid: wonderbaarlijk*
Een onderzoeker beweert een nieuw (en duur) middel tegen asthma uitgevonden te hebben. De bewijzen liegen er niet om: hij heeft het middel aan 43 patiënten toegediend en bij 36 was de benauwdheid na 3 weken volkomen verdwenen en bij de andere 7 minder geworden.

>a Als je bovenstaande alinea in de krant las en je had last van benauwdheid, zou jij dan dit middel willen gebruiken?

Enige jaren later begint een andere arts te twijfelen: zijn patiënten hebben er op den duur weinig baat bij.
Hij gaat het volgende doen:
37 patiënten krijgen het nieuwe geneesmiddel.
37 andere patiënten krijgen volkomen onschuldige neppilletjes, maar er wordt tegen ze gezegd dat het de 'echte' pillen zijn. Zo'n neppil wordt een 'placebo' genoemd.
Na 3 weken vergelijkt hij de resultaten van de twee groepen:

	genezen	verbetering	geen verandering
Groep 'echte' pil	31	5	1
Groep 'placebo'	28	9	0

> Wat is je opinie over het nieuwe geneesmiddel?

3. *Eindelijk: het haargroeimiddel*

Er is nu een haargroeimiddel dat werkt. Maar uw dokter weet of 't ook helpt.

Er is de laatste tijd veel publiciteit geweest over een nieuw haargroeimiddel dat in de medische wereld werd geïntroduceerd. Het betreft een vinding van Upjohn, die het heeft geregistreerd onder de naam Regaine®. The Upjohn Company is een wereldwijd opererende, op research gerichte producent van geneesmiddelen.

Regaine® is het enige klinisch bewezen effektieve middel voor de behandeling van mannelijke kaalhoofdigheid.

Het middel werkt niet bij iedereen. Klakkeloze aanprijzing is daarom niet verantwoord. De resultaten zijn afhankelijk van vele faktoren, zoals het aantal jaren dat en de mate waarin iemand kaal is, het soort haaruitval en de oorzaak hiervan, alsook leeftijd en algemene gezondheidstoestand.

De arts is de aangewezen deskundige om te beslissen wie voor een haargroeimiddel in aanmerking komt.

Daarom adviseert Upjohn-Nederland diegenen die overwegen het middel te gebruiken hun arts te raadplegen. Voor meer informatie kunt u bellen: 020-731233 of 020-731663.

Upjohn Maandkuur Regaine: f 101,25, verkrijgbaar bij de apotheek.

In 1986 werd een nieuw haargroeimiddel ontdekt. Het was eigenlijk een middel tegen hoge bloeddruk maar had als bijwerking een verhoogde haargroei. Was er dan eindelijk hoop voor al die kalende heren al of niet voorzien van een toupet? Maar eerst diende er zekerheid te komen: is het echt een effectief haargroeimiddel?

> Beschrijf hoe jij een experiment zou inrichten om er achter te komen of het een goed haargroeimiddel is.

4. *Vaccinatie levensgevaarlijk*

In oktober 1976 werd in de U.S.A. een griepvaccinatie gestart. Allereerst werden de ouderen en zwakken ingeënt. In de eerste week werden 24.000 mensen van 65 jaar en ouder behandeld. Drie ervan overleden kort daarna. Daarop stopten acht staten de vaccinatie.

> Welk commentaar zou je willen geven?

5. *Vitamines verlengen je leven*

De laatste jaren is er een enorme stijging in het gebruik van vitaminen. Nederland volgt wat dat betreft de ontwikkelingen in de U.S.A. en Duitsland.

Speciale 'Gezondheidswinkels' rijzen als paddestoelen uit de grond. Eén van die zaken, met een groot aantal filialen, wil voor reclamedoeleinden aantonen dat veel vitaminepillengebruik de gezondheid bevordert.

Er worden daarom enquêteformulieren neergelegd in alle filialen. De enquête wordt ingevuld door 3214 mensen. Van de ondervraagden zegt 91% baat te hebben bij extra vitaminengebruik.

De advertentie laat niet lang op zich wachten:

'Onderzoek toont aan: 91% van mensen heeft baat bij extra vitaminengebruik.'

> Deze zin is misleidend. Waarom?

6. *Ontevreden vrouwen*
Bij een onderzoek naar seksuele gewoontes werden 50.000 vrouwen aangeschreven met het verzoek een enquête in te vullen. Slechts 3750 voldeden aan dat verzoek.
De resultaten van het onderzoek werden vastgelegd in een rapport. Daarin werd steeds gesproken over bijvoorbeeld:
'drie van de vier vrouwen ontevreden over man'.

> Is zo'n uitspraak betrouwbaar?

7. *Nederland niet populair*
Een onderzoeker wil graag weten hoeveel Nederlanders er dit jaar de zomervakantie in eigen land willen doorbrengen.
Hij wil niet de fout maken enqueteformulieren *vrijwillig* te laten invullen omdat dit vertekeningen zou kunnen opleveren.
De onderzoeker kiest de volgende strategie: hij bezoekt 27 reisbureaus in de randstad (daar wonen veel mensen) en hij zal iedere zevende bezoeker van elk bureau interviewen. De uitkomst van het onderzoek:
'Record aantal Nederlanders naar buitenland'.
Het onderzoek lijkt zorgvuldig opgezet, toch zijn er nog wel verbeteringen mogelijk.

> Geef aan hoe je de opzet van dit onderzoek kunt verbeteren.

Samenvatting

Statistische onderzoeken zijn onderzoeken waarbij door het *verzamelen* van *getallen* geprobeerd wordt antwoorden op vragen te vinden.
Daarbij kan op verschillende manieren gewerkt worden.

Kijken en *beslissen*:
- De cijfers voor wiskunde gaan omhoog: bijles is goed.
- De benauwdheid verdwijnt: medicijn is goed.
- Er sterven mensen na vaccinatie: vaccinatie stoppen.

Deze methode wordt veel toegepast en moet met terughoudenheid worden gebruikt.

Kijken en *vergelijken*:
- De cijfers voor wiskunde gaan voor zowel de bijlesleerlingen als de anderen omhoog: bijles maakt niet uit.
- De benauwdheid verdwijnt ook bij neppillen: het effect is waarschijnlijk psychologisch.
- Er sterven misschien wel veel meer mensen als de vaccinatie niet plaats vindt: vaccinatie doorzetten.

Deze methode is veel beter omdat de resultaten van de 'behandelde' groep gezet worden naast een vergelijkbare niet-behandelde groep.

Bij onderzoeken (en dus ook bij enquêtes) moet je er voor zorgen dat de deelnemende mensen willekeurig worden uitgekozen, dus:
- *niet* mensen die overtuigd vitaminenkopers zijn vragen of ze er positief effect van ondervinden;
- *niet* mensen over hun thuisblijfplannen enquêteren op een reisbureau: de thuisblijvers komen vaak niet eens op het reisbureau;
- *niet* bij medische experimenten alleen met vrijwilligers of eigen patiënten werken;
- *niet* een onderzoek als volwaardig presenteren als slechts een klein deel van de ondervraagden heeft geantwoord.

2 De steekproef

Bijna de helft van de Nederlanders heeft op een bepaald moment in 1987 vertrouwen in Lubbers (zie kranteteksт op blz. 1). Wil dat zeggen dat ruim 6 miljoen Nederlanders aan het enquête bureau hebben meegedeeld vertrouwen in Lubbers te hebben? Nee dus. In feite had het bureau maar 1200 mensen opgebeld, waarvan er 588 vertrouwen in Lubbers hebben.
Onder welke voorwaarden kun je met zo'n kleine *steekproef* onder de totale bevolking volstaan en toch een redelijk betrouwbare uitspraak doen? Daarover gaat dit hoofdstuk.

1. *Steekproeven vergelijken*
   Drie manieren om een steekproef uit te voeren:

   1: In een winkelstraat in Amsterdam worden 1000 mensen gevraagd of ze vertrouwen in Lubbers hebben.
   2: Uit een adressenlijst van de abonnees van de grootste krant van Nederland worden 2000 namen willekeurig geselecteerd: iedere 350ste op de lijst van 700.000 krijgt een enquêteformulier thuis gestuurd met de vraag in welke politicus men vertrouwen heeft.
   3: Uit alle telefoonboeken van Nederland worden 1200 mensen geselecteerd. Uit ieder van de 50 regionale telefoongidsen 24 mensen.
   Die 24 mensen per boek worden als volgt uitgezocht. Het boek wordt 24 keer op een willekeurige bladzijde opengeslagen en met een speld wordt blind een naam geprikt.

   >a Vergelijk deze drie manieren.
   Bij welke manieren mag je verwachten dat de steekproef een goede afspiegeling geeft van de Nederlandse bevolking?
   >b Verzin zelf een voorbeeld van een goede steekproef onder 1000 Nederlanders om de populariteit van Lubbers te meten.

Een belangrijke eigenschap van een steekproef is dat hij *representatief* moet zijn. Dat betekent dat de (kleine) groep mensen die bij het onderzoek wordt betrokken een goede afspiegeling moet vormen van de totale groep waarop het onderzoek betrekking heeft.

2. > Zijn de steekproeven, genoemd bij de opgaven 5, 6 en 7 in hoofdstuk 1 representatief?

3. Een op de vier leerlingen in het middelbaar onderwijs gebruikt wel eens drugs. Via een (anonieme!) enquête onder alle leerlingen van jouw school, wil je deze uitspraak controleren.

   > Krijg je op die manier een representatieve steekproef?

Een tweede belangrijke eigenschap van een steekproef is dat de personen echt willekeurig worden aangewezen.
Bij manier 3 van opgave 1 gebeurde dat met 'spelden prikken'.
Het laten 'spelden prikken' kan vervangen worden door meer professionele technieken. Daarbij spelen computers een grote rol. Zo zou je in theorie alle namen van de Nederlanders in een computer kunnen stoppen (of zitten ze er al in?) en de computer kan daar dan een steekproef uit trekken. Op echt willekeurige wijze.
Soms wordt computers gevraagd een lijst van willekeurige getallen van bijvoorbeeld vijf cijfers te produceren:

101	03918	86495	47372	21870	28522	99445	38783	83307
102	10041	35095	66357	64569	08993	20429	28569	63809
103	43537	58268	80237	17407	89680	04655	24678	61932
104	64301	47201	31905	60410	80101	33382	95255	10353
105	43857	42186	77011	93839	28380	49296	63311	49713
106	91823	39794	47046	78563	89328	39478	04123	19287
107	34017	87878	35674	39212	98246	29735	09924	27893
108	49105	00755	39242	50472	39581	44036	54518	46865
109	72479	02741	75732	99808	02382	77201	44932	88978
110	84281	45650	28016	77753	39495	41847	19634	82681
111	61589	35486	59500	20060	89769	54870	75586	07853
112	25318	01995	87789	41212	74907	90734	31946	24921
113	40113	37395	51406	98099	43023	70195	07013	72306
114	58420	43526	15539	24845	15582	16780	95286	69021
115	18075	45894	09875	42869	20618	07699	80671	54287
116	52754	73124	93276	71521	59618	44966	37502	15570
117	05255	53579	08239	99174	75548	95776	42314	13093
118	76032	35569	28738	38092	74669	00749	17832	64855
119	97050	31553	32350	51491	53659	89336	36912	05292
120	29030	43074	84602	95131	22769	44680	68492	33987
121	28124	29686	63745	12313	15745	11570	20953	17149
122	97469	41277	90524	36459	22178	63785	20466	67130
123	91754	40784	38916	12949	76104	20556	34001	59133
124	84599	29798	57707	57392	91757	76994	43827	69089
125	06490	42228	94940	10668	62072	58983	10263	08832
126	30666	02218	89355	76117	75167	69005	42479	79865
127	87228	15736	08506	29759	74257	85594	75154	48664
128	45133	49229	32502	99698	68202	44704	39191	73740
129	55713	98670	57794	64795	27102	83420	26630	95009
130	20390	38266	30138	61250	07527	02014	43972	49370
131	13400	68249	32459	41627	56194	93075	50520	96784
132	08900	87788	73717	19287	69954	45917	80026	55598
133	86757	47905	16890	99047	78249	73739	97076	00525
134	19862	54700	18777	22218	25414	13151	54954	80615
135	96282	11576	59837	27429	60015	40338	39435	94021
136	17463	26715	71680	04853	55725	87792	99907	67156
137	44880	55285	95472	57551	24602	98311	63293	58110
138	61911	78152	96341	31473	58398	61602	38143	93833
139	07769	22819	58373	88466	71341	32772	93643	92855
140	73063	63623	29388	89507	78553	62792	89343	27401
141	24187	60720	74055	36902	22047	09091	79368	35408
142	06875	53335	91274	87824	04137	77579	54266	38762
143	23393	37710	46457	03553	58275	11138	18521	59667
144	00980	73632	88008	10060	48563	31874	90785	78923
145	46611	39359	98036	25351	88031	72020	13837	03121
146	56644	79453	49072	30594	73185	81691	29225	70495
147	98350	36891	04873	71321	29929	37145	95906	41005
148	17444	61728	86112	76261	92519	61569	65672	95772
149	45785	21301	89563	23018	60423	50801	70564	45398
150	54369	08513	36838	19805	67827	74938	66946	01206

Zo'n lijst kan gebruikt worden om een goede steekproef samen te stellen (goede steekproef heet ook wel *a-selecte steekproef*). Hoe dat kan, bekijken we aan de hand van een voorbeeld.

Bij een autofabriek moeten de laatste 50 auto's van de produktielijn gecontroleerd worden. Men besluit een steekproef van zes te nemen. Daartoe worden de auto's genummerd van 01 tot 50.
We kiezen nu een willekeurige regel van de tabel met toevalsgetallen, bijvoorbeeld regel 121.

Deze luidt:

2 8 1 2 4 2 9 6 8 6 6 3 7 4 5 1 2 3 1 3 1 5 7 4 5

De eerste auto die we kiezen is nu: 28

2 8

De volgende: 12

2 8 1 2

enz:

2 8 1 2 4 2 9 6 8 6 6 3 7 4 5 1 2 3 1 3 1 5 7 4 5

x x x x x → te hoog

De auto's voor de steekproef hebben dus de nummers: 28, 12, 42, 23, 13 en 15.

4. > Gebruik regel 125 van de tabel met toevalsgetallen om een steekproef van zeven leerlingen uit je klas te nemen.

5. *Verantwoord medisch experiment*
Twintig patiënten hebben zich aangemeld voor een medisch experiment met een nieuw geneesmiddel.

1.	Brouwer	6.	Keyser	11.	Kuyt	16.	Van Akkeren
2.	Minneboo	7.	Koomen	12.	Van der Meer	17.	Doeven
3.	Dijkman	8.	Berkhey	13.	Mos	18.	Balvert
4.	De Jong	9.	Koetsier	14.	Pennings	19.	Van Doorn
5.	Jansma	10.	Van Dam	15.	Reys	20.	Van Lith

>a Beschrijf hoe je dat experiment op kunt zetten mèt controlegroep (binnen deze twintig) en gebruikmakend van de tabel met toevalsgetallen (regel 101).

De ziekte uit voorgaande opgave is dodelijk.

>b Wat vind je in zo'n geval van een 'goed' experiment met een controlegroep, als je weet dat de eerste ervaringen met het geneesmiddel zeer positief zijn?

6. *Onveilig gevoel*
Bij een 'goed' onderzoek in de Verenigde Staten bleek dat 45% van de mensen zich 's avonds op straat niet veilig voelde.
De steekproef was 1500 mensen groot. De V.S. hebben meer dan 200 miljoen inwoners.
Twee weken later wordt er aan 1500 andere mensen (weer een 'goede' steekproef) dezelfde vraag gesteld. De uitkomst is nu 47%.
De kranten schrijven:

"Bevolking voelt zich steeds onveiliger!"

>a Wat vind je van zo'n mededeling?

Nader onderzoek bij het enquêteringsbureau levert de volgende informatie op over de betrouwbaarheid. Een enquête-uitslag van 45% betekent dat het percentage voor de hele bevolking vrijwel zeker ligt tussen 42% en 48%.
Toch gebeurt het af en toe (gemiddeld in één op de twintig gevallen) dat het werkelijke percentage nog meer afwijkt van het steekproefpercentage.
Na de meting van 45% en die van 47% volgde twee weken later een derde van 50%.

>b Kun je nu *zeker* weten dat het percentage mensen dat zich 's avonds op straat niet veilig voelt, over deze periode van vier weken is toegenomen?

7. *Koppositie*

Om de vier jaar proberen Amerikaanse politici via geldverslindende campagnes de eigen partij te overtuigen van het feit, dat zij de aangewezen presidentskandidaat zijn.

De media proberen daarbij elk zwak moment uit te buiten om zo iemand op fouten te betrappen.

Voor de opvolging van president Ronald Reagan (1988) leek senator Hart aanvankelijk de belangrijkste kandidaat voor de Democratische partij. Totdat journalisten van de Miami Herald zijn zwakke plek vonden: een avontuurtje met het fotomodel Donna Rice.

Donna Rice, Hart(en)breekster

Hart, met echtgenote

Senator Hart raakt koppositie al weer kwijt

Vorige week meldde de Miami Herald, de krant die in mei vorig jaar Harts buitenechtelijke relatie met Donna Rice onthulde, dat hij tijdens zijn gooi naar de Democratische presidentskandidatuur in 1984 financiële contributies van een Californische zakenman heeft ontvangen die ver uitgaan boven wat de federale kieswet toestaat. Andere kranten volgden met het bericht, dat destijds nog twee geldschieters te diep in de portemonnee hadden getast ten behoeve van Hart.

Hart beloofde de zaak te zullen uitzoeken en moest vervolgens toegeven dat er althans in een geval sprake is geweest van ongeoorloofd hoge campagne-bijdragen. De kwestie had en heeft niet erg veel om het lijf, maar Hart — wiens standaard-antwoord op vragen over de affaire-Rice luidt dat hij in zijn privé-leven dan wel "gezondigd" mag hebben, maar dat zijn publieke staat vandienst onberispelijk is — kan er op zijn blazoen geen enkele smet meer bij hebben.

De voorzitter van de Democratische partij in Iowa — die eens in de vier jaar een gewichtig figuur is — voorspelde al onmiddellijk dat Hart de weerslag zou ondervinden van de nieuwe negatieve publiciteit, en dat is inderdaad gebeurd. In de jongste peiling onder de Democraten die vermoedelijk zullen deelnemen aan de indirecte voorverkiezingen, is hij teruggezakt naar de vierde plaats.

De nieuwe koploper is afgevaardigde Richard Gephardt. Hij geniet de voorkeur van negentien procent van de kiezers. Dukakis (achttien procent) en senator Paul Simon (zeventien procent) zitten hem evenwel zo dicht op de hielen, dat het gelet op de onnauwkeurigheidsmarge van de peiling zeer wel mogelijk is dat een van hen de werkelijke aanvoerder is. Hart volgt met dertien procent, de zwarte dominee Jesse Jackson staat op elf procent, ex-gouverneur Bruce Babbitt scoort tien procent, en senator Albert Gore, die nauwelijks campagne voert in Iowa en zich geheel toelegt op de voorverkiezingen in de zuidelijke staten, is hekkesluiter met een procent. Ruim een-tiende van de Democratische kiesgerechtigden heeft nog geen keus gemaakt.

de Volkskrant jan.88

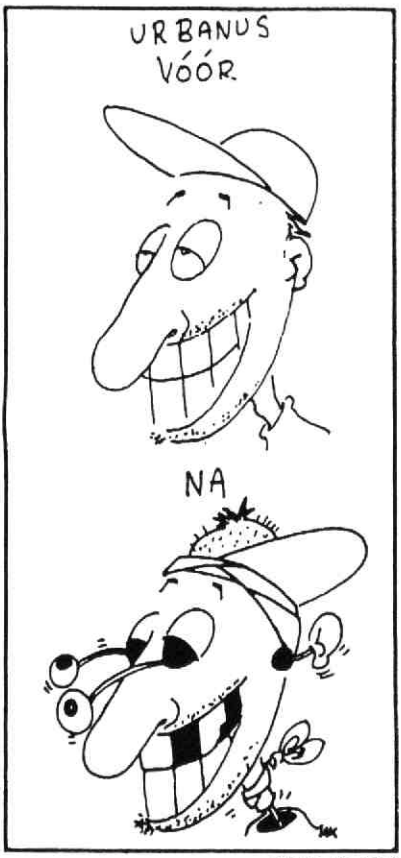
>a Als bij dit onderzoek ook geldt dat er een speling is van 3% naar boven en 3% naar beneden, wie komen er dan allemaal in aanmerking voor de koppositie?

>b Welke invloed kan de 'Ruim een-tiende van de Democratische kiesgerechtigden' uit de laatste zin van het artikel op de koppositie uitoefenen?

8. *Vuurwerk*
In december 1987 werd door de Stichting Consument en Veiligheid een campagne gehouden om te wijzen op het gevaar van vuurwerk. In spotjes op de t.v. en door middel van posters werd opgeroepen om toch vooral voorzichtig te zijn.

De Volkskrant schreef in januari 1988:

Maar de jongste vuurwerkcampagne, die vier miljoen gulden zou hebben gekost als niet iedereen gratis had meegewerkt, heeft wel degelijk effect gehad. Het is gemeten door Aly Hendriks en Niels Rood van de Marktplan Adviesgroep in Amsterdam. Het doel van de campagne van de Stichting Consument en Veiligheid en SIRE was niet het vuurwerkgebruik met oudjaar en de dagen daaromheen te verminderen. Het ging er uitsluitend om de gebruikers bewust te maken van de gevaren van vuurwerk en dan in het bijzonder de gevaren voor *jezelf* (verknal je toekomst niet).
De Markplan Adviesgroep ondervroeg vijfhonderd scholieren voordat de campagne begon. Dezelfde groep werd na de campagne weer ondervraagd, maar daarnaast nog een andere groep van vijfhonderd jongeren. Dit laatste om te kunnen controleren of het feit dat de eerste groep twee keer werd ondervraagd, op zichzelf ook van invloed is geweest op de antwoorden.
De eerste vraag luidde: "Zou je een oorzaak kunnen noemen van ongelukken die speciaal in de winter of kerstvakantie kunnen gebeuren?"
Het meest genoemd (59 procent): uitglijden. Maar nummer twee in de spontaan genoemde gevaren was al meteen vuurwerk.
Voordat de campagne begon, noemde een kwart van de jongeren al vuurwerk als belangrijke ongelukkenmaker. Toen de campagne liep, werd het opnieuw gevraagd en toen werd door 37 procent van de jongeren vuurwerk als eerste genoemd. Ook bij de controlegroep lag het percentage in de buurt (34 procent).



Illustratie ZAK

Urbanus was één van de gratis medewerkers aan de campagne.

> Is de conclusie gerechtvaardigd (van de Volkskrant) dat de toename van 25% naar ongeveer 35% (uitsluitend) te danken is aan de campagne? Bedenk dat de tweede mening eind december werd gevraagd.

Samenvatting

Meestal is het onmogelijk om iedereen z'n mening te vragen omtrent een bepaalde vraag. Bijna altijd wordt volstaan met een *steekproef* van zo'n 500 - 2000 mensen.
Dat is alleen verantwoord als de steekproef echt een doorsnee van de hele groep vertegenwoordigt. Men spreekt dan van een *representatieve* steekproef.

Er zijn verschillende manieren om tot een *a-selecte* steekproef te komen.
Toevalsgetallen, gemaakt door de computer, worden daarbij vaak gebruikt.
Maar zelfs in ideale situaties blijft een steekproef een steekproef. Ook bij een goede steekproef moet rekening gehouden worden met een zekere speling; de onbetrouwbaarheidsmarge (zoals de speling van 3% naar boven en naar beneden bij opgave 6).
Wel geldt: hoe groter de steekproef, des te kleiner de speling.

3 Getallen in beeld: histogram

1. *Onveilig*

In het vorige hoofdstuk stond de volgende bewering:

"45% van de mensen voelt zich op straat onveilig."

Het bureau dat de enquêtering verrichtte tekende daarbij aan:

— De waarde 45% dient gelezen te worden als: ergens tussen de 42% en 48%.

— In één op de twintig gevallen kan de 'werkelijke' waarde zelfs buiten het 42-48 interval liggen.

Om deze uitspraak te controleren wordt een onderzoek uitgevoerd:

— Er worden 24 verschillende steekproeven van 1500 man genomen.

De resultaten zijn als volgt:

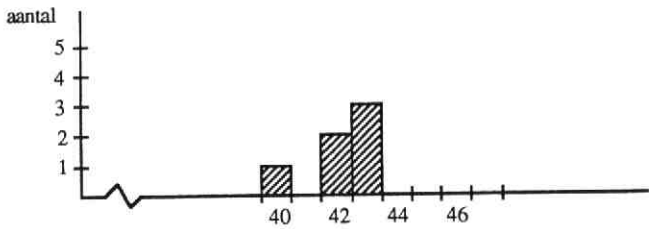
44	45	45	46	45	46
43	47	42	44	46	44
40	47	44	48	43	45
42	45	46	43	48	47

Deze resultaten kunnen in twee fasen eenvoudig verwerkt worden:

— Er wordt een turftabel gemaakt:

waarneming	40	41	42	43	...
aantal	/		//	///	...

— Vervolgens worden de waarnemingen (ook wel *data* genoemd) grafisch verwerkt.
Dit kan bijvoorbeeld met een *histogram*:



percentage mensen dat zich onveilig voelt

Horizontaal worden de waarnemingsgetallen uitgezet (in dit geval de percentages die bij de verschillende steekproeven zijn gevonden).
Verticaal staat aangegeven het aantal keren dat een waarnemingsgetal is gevonden.
Voorbeeld: bij de 24 steekproeven is drie keer een percentage van 43% gevonden.

>a Maak de turftabel en het histogram af.
Waarom zit er een knik in de horizontale as van het histogram?

>b Controleer aan de hand van het histogram de uitspraken van het enquêteringsbureau over de betrouwbaarheid van het steekproefresultaat.

Een half jaar na bovenstaand onderzoek werd weer een onderzoek uitgevoerd.
De uitkomst van het enquêteringsbureau was:

'48% van de mensen voelt zich op straat onveilig'

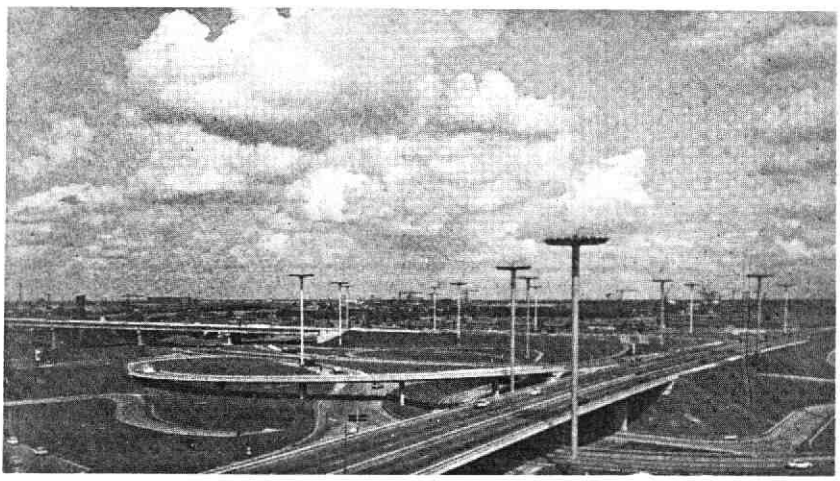
Ook nu wordt er controleonderzoek gedaan.
Er worden nu 18 verschillende steekproeven genomen met de volgende resultaten:

42	43	44	42	45	41
45	44	44	44	43	45
46	44	47	43	46	44

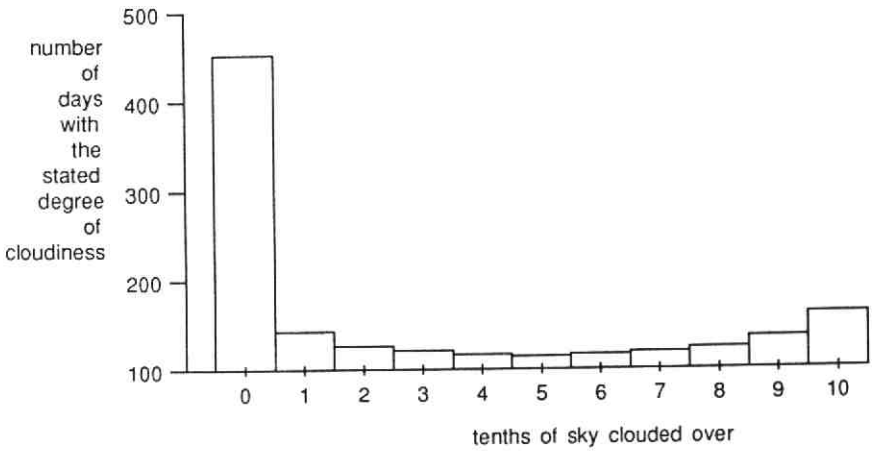
>c Verwerk deze resultaten in een histogram.

>d Vergelijk de twee histogrammen.
Wat is je conclusie?

Half bewolkt



2.



Voor alle maanden juli in de periode 1890-1955 werd op iedere dag de mate van bewolking gemeten. Was het de hele dag bewolkt, dan werd het getal 10 toegekend: $\frac{10}{10}$ bewolking. Was het de hele dag helder, dan sprak men van 0 (van $\frac{0}{10}$ bewolking).

>a Hoeveel dagen waren onbewolkt?
>b Hoeveel dagen half bewolkt ($\frac{5}{10}$)?

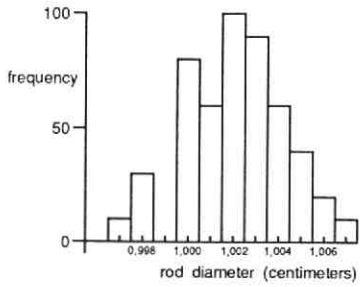
In juli 1987 werd de mate van bewolking ook gemeten.
De 'frequentietabel' ziet er zó uit:

vrijwel onbewolkt	4
licht bewolkt	3
half bewolkt	10
zwaar bewolkt	10
bewolkt	4

Om 1987 te kunnen vergelijken met de periode 1890-1955 gaan we het histogram aanpassen.

>c De elf verschillende staven moeten teruggebracht worden tot vijf staven. Dat kan door steeds twee staven samen te nemen. In één categorie moeten dan drie staven worden samengevoegd.
Bij welke categorie zou je dat doen? Waarom?
>d Vergelijk juli 1987 met de juli-maanden in de periode 1890-1955. Commentaar?
>e Is het mogelijk dat één of meerdere van de juli-maanden uit de periode 1890-1955 hetzelfde weerbeeld te zien gaven als juli 1987?

3. *De diameter van assen*



Bij een fabriek voor technische apparatuur worden asjes gemaakt die een diameter van precies 1 cm moeten hebben. Asjes die iets dikker zijn worden ook goedgekeurd. Asjes die *dunner* zijn worden afgekeurd.

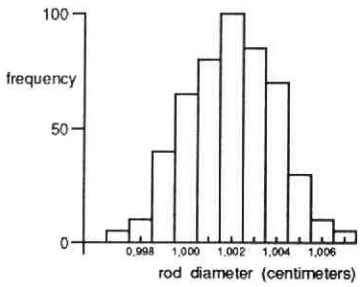
Er worden 500 assen gecontroleerd door de inspecteur van de fabriek. Daarna wordt een frequentietabel gemaakt die resulteert in bovenstaand histogram.

>a Hoeveel van de 500 assen worden afgekeurd?

De machine waarmee deze asjes worden gemaakt staat afgesteld op een diameter van 1.002 cm. Dat betekent nog niet dat ze allemaal precies die diameter hebben. Er zit een zekere speling in: de een wat te groot, een ander weer te klein.

De directeur van de fabriek had daarom eigenlijk een ander histogram verwacht.

Zo iets:



>b Welk opvallend verschil is er tussen het door de directeur verwachte histogram en het door de inspecteurs geleverde histogram?

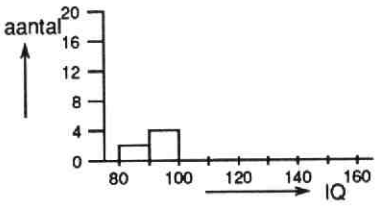
>c Kun je een schatting maken van het aantal assen dat afgekeurd *had moeten* worden?

4. *Steel-en-bladdiagram*
Bij een IQ-meting onder 50 Nederlanders werden de volgende resultaten gevonden:

106	103	117	103	116	96	102	130	83	107
127	118	114	117	108	115	119	147	109	142
99	141	125	88	104	132	124	114	127	136
111	96	108	112	153	136	106	108	111	101
105	110	101	120	101	106	98	118	106	100

>a Maak de volgende frequentietabel en het bijbehorende histogram af:

IQ-waarde	aantal
$80 \leq IQ < 90$	2
$90 \leq IQ < 100$	4
$100 \leq IQ < 110$	...
.....	...
.....	...



Tegenwoordig wordt ook vaak gebruik gemaakt van een *steel-en-bladdiagram.*
Voor de IQ-metingen ziet het diagram er als volgt uit:

8	38
9	6689
10	011123345666678889
11	0112445677889
12	04577
13	0266
14	127
15	3

13|0266 staat voor de getallen 130, 132, 136 en 136.
De getallen voor de streep worden de *stelen* genoemd. In dit voorbeeld zijn dat de verschillende tientallen (8 t/m 15) die bij de IQ-metingen voorkwamen.
Achter de streep staan de *bladeren* die bij zo'n steel horen.

>b Welke voordelen biedt een steel-en-bladdiagram boven een frequentietabel en histogram?

5. In een dubbel steel-en-bladdiagram staan de resultaten van een proefwerk in twee klassen (A en B) vermeld. De maximale score was 99.
De steel staat hier in het midden.

Class A		Class B
2	1	23
9	2	
8	3	
9	4	
87	5	
87	6	
98	7	00122346
76532210	8	012448
91	9	0139

>a Hoeveel leerlingen hebben in elk van de twee klassen meegedaan aan het proefwerk?
>b Je ziet dat in klas A slechts één leerling een 38 had.
Hoeveel leerlingen in klas A hadden een score van 82?
En in klas B?
>c Hoeveel leerlingen in totaal hadden een onvoldoende (≤ 54).
>d Kun je aan deze tabel zien in welke klas het proefwerk gemiddeld het beste is gemaakt?

6. *Top 15 van 1984*

Vijftien pop-journalisten in de Verenigde Staten - daar komt tenslotte bijna alle pop vandaan - hebben ieder een lijst gemaakt van de beste tien l.p's van 1984.
Een eerste plaats leverde tien punten op, een tweede negen, enz.
De maximale score voor een l.p. was dus 150 punten.
De resulterende lijst ziet er als volgt uit:

	Album	Artist	Points
1.	*Born in the U.S.A.*	Bruce Springsteen	94
2.	*Purple Rain*	Prince	83
3.	*How Will the Wolf Survive?*	Los Lobos	55
4.	*Reckoning*	R.E.M.	46
5.	*Private Dancer*	Tina Turner	–
6.	*Let It Be*	Replacements	26
7.	*Learning to Crawl*	Pretenders	25
8.	*Double Nickels on the Dime*	Minutemen	24
9.	*The Magazine*	Rickie Lee Jones	24
10.	*The Unforgettable Fire*	U2	19
11.	*Lush Life*	Linda Rondstadt	–
12.	*Zen Arcade*	Hüsker Dü	15
13.	*Soul Mining*	The The	14
14.	*Meat Puppets II*	Meat Puppets	13
15.	*Sparkle in the Rain*	Simple Minds	12

>a Tina Turner haalde bij zes journalisten de top-tien met de scores: tiende, eerste, derde, vierde, derde, vijfde.
Hoeveel punten haalde ze ermee?

>b Linda Ronstadt haalde maar bij twee journalisten de lijst met een tweede en derde plaats.
Hoeveel punten is dat waard?

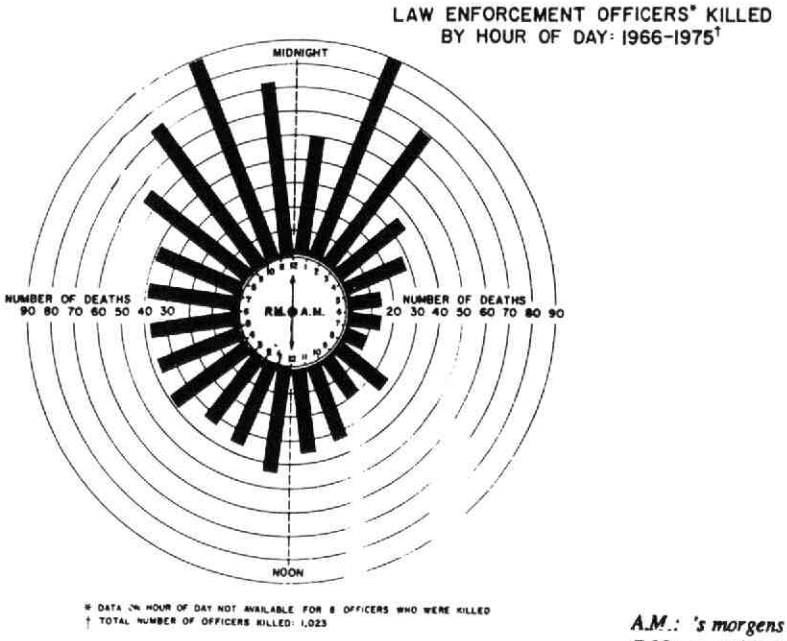
>c Hoeveel journalisten moeten Bruce Springsteen zeker genoemd hebben?

>d Teken een steel-en-bladdiagram; als steel gebruik je 9, 8, 7, ... (tientallen).
Geef commentaar op de verdeling.

>e Maak een lijst van je persoonlijke single-top 5 van dit moment. Verzamel de gegevens van de hele klas en maak een steel-en-bladdiagram voor de single-top 15 van de klas.

>f Lijken de twee steel-en-bladdiagrammen, wat vorm betreft, op elkaar? Hoe zijn eventuele verschillen te verklaren?

7. *Vermoorde agenten*

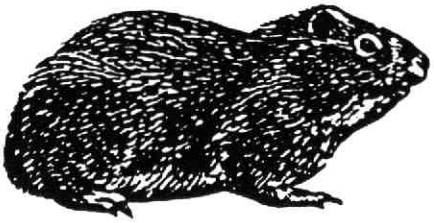


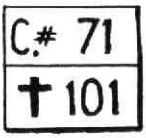
A.M.: 's morgens
P.M.: 's middags
Midnight: 12.00 uur
Noon: 12.00 uur overdag

Een variatie op het normale histogram is bovenstaande grafiek. Omdat de horizontale as de uren van de dag aangeeft, is het einde van de dag vastgemaakt aan het begin van de volgende.

>a Hoeveel agenten worden er gemiddeld 'per uur' overdag vermoord? (In de periode '66-'75.)
>b Hoeveel agenten worden er gemiddeld 'per uur' 's nachts vermoord?
>c Teken een 'normaal' histogram met langs de horizontale as de volgende indeling:
0- 3 uur 's nachts
3- 6 " "
6- 9 enz.

8. *Dierproeven*





In het kader van een medisch experiment worden 72 cavia's ingespoten met de tuberculose bacil. Er wordt gekeken na hoeveel dagen de cavia's overlijden.
Na 43 dagen gaat de eerste dood. De sterkste houdt het 598 dagen vol.
De volledige gegevens:

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								

Deze gegevens worden verwerkt in een histogram.
Het is duidelijk dat we daarbij de horizontale as niet precies in dagen gaan verdelen; we krijgen dan veel te veel staven en veel te korte staven. Zelfs een week als eenheid langs de horizontale as is te klein:
598 dagen is ruim 85 weken.
Kies als eenheid langs de horizontale as: 30 dagen.

>a Maak de volgende frequentietabel af:

aantal dagen	aantal dieren
$0 \leq$ aantal dagen < 30	0
$30 \leq$ aantal dagen < 60	7
$60 \leq$ aantal dagen < 90	...
...	...
...	...

>b Maak het histogram bij de frequentietabel.

>c Na 102 dagen is de helft van de cavia's al dood. Dat hoeft nog niet te betekenen dat de cavia's *gemiddeld* zo'n 102 dagen leven.
Zal het gemiddelde meer of minder dan 102 dagen zijn?

samenvatting

Een eenvoudige manier om getallen in beeld te brengen, is het histogram, of staafdiagram.
Eerst wordt er een turftabel gemaakt; in de 'nettere' vorm heet die een frequentietabel (frequentie = aantal).

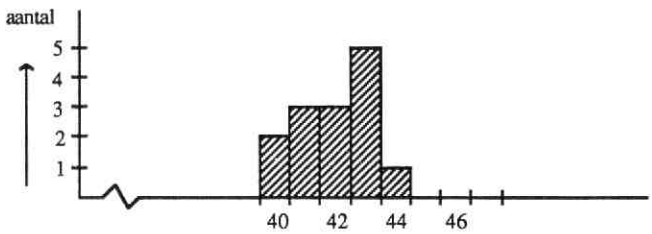
Turftabel:

Waarneming	*Aantal*
40	*//*
41	*///*
42	*///*
43	*~~////~~*
44	*/*

Frequentietabel:

Waarneming	*Aantal*
40	*2*
41	*3*
42	*3*
42	*5*
44	*1*

Vervolgens wordt een geschikte verdeling van horizontale en verticale as genomen, waarna de verschillende staven getekend kunnen worden.



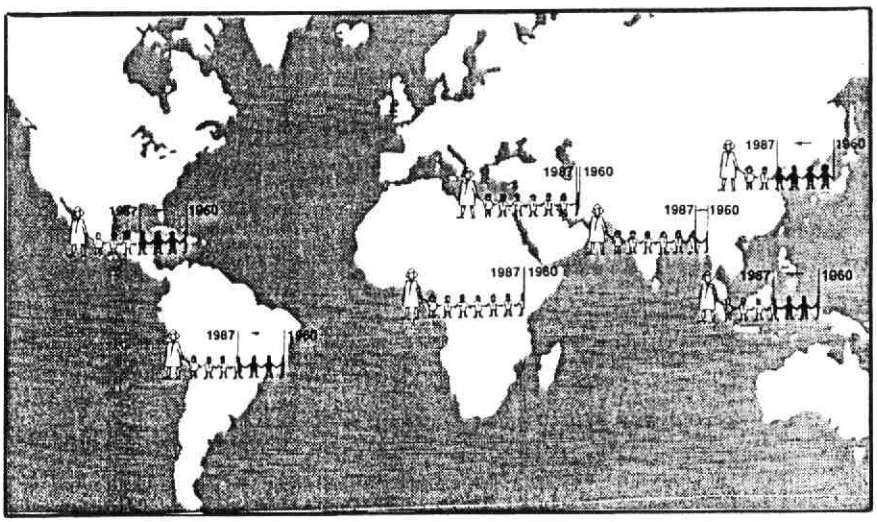
Liggen de waarnemingen wijd verspreid (zoals bij de cavia's en de IQ-metingen), dan wordt er meestal met een klasse-indeling gewerkt.
Het aantal klassen hangt af van de klasse-breedte.
In de praktijk wordt meestal gekozen voor een aantal klassen dat ligt tussen de 5 en de 20.

Het steel-en-bladdiagram is een bruikbare variant op de frequentie-tabel. Eigenlijk is het een frequentietabel en een histogram tegelijk, waarbij ook nog de afzonderlijke waarnemingen zichtbaar blijven.

4 Grafische verwerking

Getallen omzetten in plaatjes kan op veel manieren. Sommige manieren zijn 'grafiek-achtig' zoals het histogram. Vaak worden ook 'pictogrammen' gebruikt: plaatjes die op één of andere manier statistische gegevens weergeven. Als voorbeeld van deze categorie een plaatje waarin getracht wordt het effect van geboortebeperking in een aantal kinderrijke regio's weer te geven.

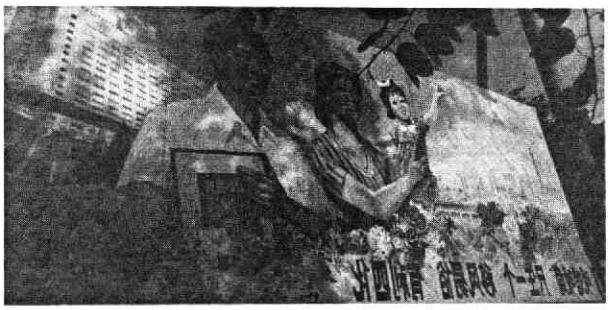
1. *Geboortebeperking*



De poppetjes geven het gemiddeld aantal kinderen per vrouw weer.

>a In welke regio is de geboortebeperking in de periode 1960-1987 het meest succesvol geweest?
En waar heeft het nauwelijks effect gehad?

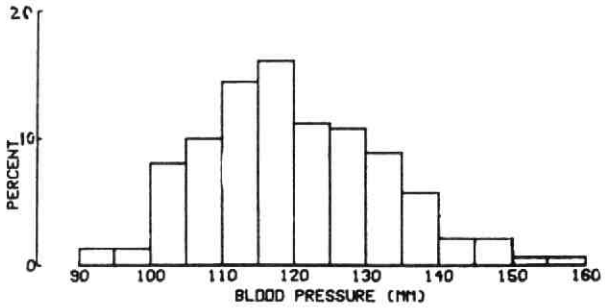
>b In welke regio was de geboortebeperking *relatief* gezien het meest succesvol?



Een poster in de Chinese stad Guanzhou. De tekst luidt: 'verander je levensstijl en draag bij aan De Vijf Moderniseringen. Doe aan geboortebeperking: neem maar één kind.'

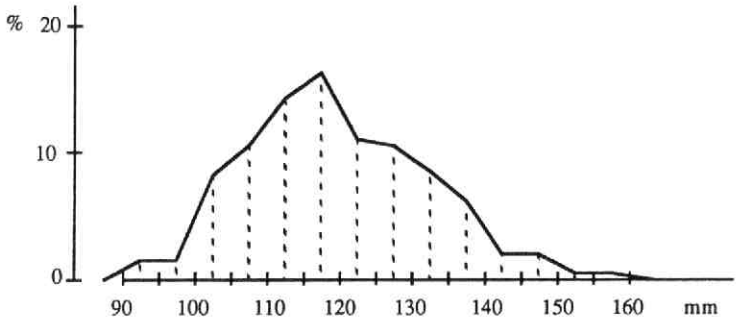
2. *De pil en bloeddruk*
Bij een uitgebreid onderzoek, waarbij 14.148 vrouwen waren betrokken, werd gezocht naar de invloed van 'de pil' op de bloeddruk.
Van alle vrouwen werd geregistreerd:
— de leeftijd
— het aantal kinderen
— gebruik van de pil (user) of niet (non-user)
— de bloeddruk

Het histogram waarin de bloeddruk van *alle* vrouwen is weergegeven:



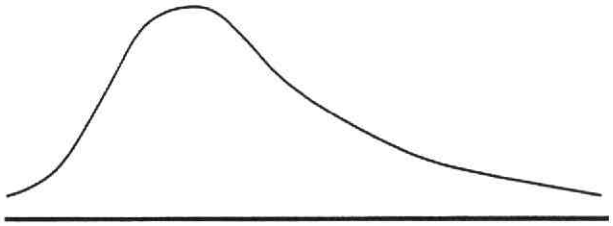
>a Schat het percentage vrouwen dat een bloeddruk heeft groter dan 130 mm.

In plaats van een histogram wordt ook wel eens een andere grafiek getekend: het *frequentiepolygoon* (frequentiecurve). Daarbij worden de *middens* van de toppen van de staven met elkaar verbonden. In bovenstaand geval levert dit:

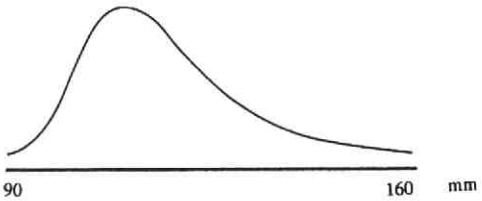


>b Probeer de volgende zin af te maken:
$\frac{2}{3}$ van de vrouwen heeft een bloeddruk die ligt tussen 105 en ... mm.

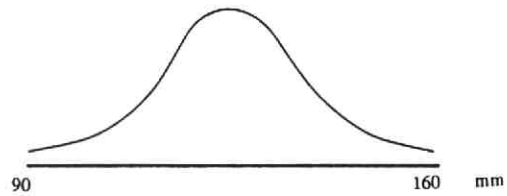
Soms wordt een frequentiecurve globaal getekend: het gaat dan alleen om de ruwe vorm.
Bij de vrouwen levert dit:



De onderzoekers hadden het vermoeden dat er eventueel een verband bestaat tussen de hoogte van de bloeddruk en het aantal kinderen.
De globale grafiek van de vrouwen met twee kinderen:



De globale grafiek van de vrouwen met vier kinderen:



>c Welke groep vrouwen heeft gemiddeld een hogere bloeddruk?
>d Kun je aan de hand van deze twee grafieken concluderen dat kinderen krijgen de bloeddruk verhoogt?

De laatste grafiek is symmetrisch en heeft de vorm van een klok. Deze 'klokvorm' zul je nog vaker tegenkomen. Met enige fantasie zou je de twee andere grafieken ook nog klokvormig kunnen noemen, maar dan wel enigszins 'scheef': een 'klok' met een staart naar rechts.

De volgende tabel toont de resultaten van de bloedmetingen onder de 14.148 vrouwen, ingedeeld in leeftijdsklassen. Bij iedere leeftijdsklasse is verder het onderscheid pilgebruiksters (users) en niet-pilgebruiksters (non-users) gemaakt.

Blood pressure (millimeters)	*Age 17–24*		*Age 25–34*		*Age 35–44*		*Age 45–58*	
	Non-users	*Users*	*Non-users*	*Users*	*Non-users*	*Users*	*Non-users*	*Users*
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
under 90	—	1	1	—	1	1	1	—
90–95	1	—	1	—	2	1	1	1
95–100	3	1	5	4	5	4	4	2
100–105	10	6	11	5	9	5	6	4
105–110	11	9	11	10	11	7	7	7
110–115	15	12	17	15	15	12	11	10
115–120	20	16	18	17	16	14	12	9
120–125	13	14	11	13	9	11	9	8
125–130	10	14	9	12	10	11	11	11
130–135	8	12	7	10	8	10	10	9
135–140	4	6	4	5	5	7	8	8
140–145	3	4	2	4	4	6	7	9
145–150	2	2	2	2	2	5	7	9
150–155	—	1	1	1	1	3	2	4
155–160	—	—	—	1	1	1	1	3
160 and over	—	—	—	—	1	2	2	5
Total percent	100	98	100	99	100	100	99	99
Total number	1.206	1.024	3.040	1.747	3.494	1.028	2.172	437

>e Teken het non-users histogram
— voor de leeftijd 17-24
— voor de leeftijd 45-58
met klasse-indeling langs de horizontale as:
90-100; 100-110; 110-120; 120-130; enz.
Bestaat er een verband tussen bloeddruk en leeftijd?

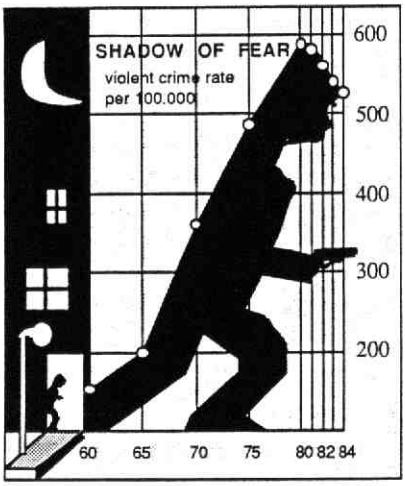
>f Doet het resultaat van >e je antwoord op >d nog veranderen?

Laten we nu kijken naar de invloed van de pil.

>g Teken in één figuur de frequentiepolygonen voor de non-users en voor de users in de leeftijdsklasse 25-34.
Gebruik dezelfde klasse-indeling als bij >e.
Mag je concluderen dat het gebruik van de pil bloeddrukverhogend is voor deze leeftijdsklasse?

>h Geldt deze conclusie ook voor de andere leeftijdsklassen?

3. *Criminaliteit*

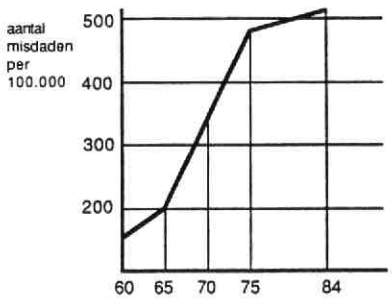


Deze grafiek geeft het aantal misdaden per 100.000 inwoners in de Verenigde Staten weer. Aanvankelijk om de vijf jaar, later om het jaar.

>a Hoeveel misdaden waren er volgens de grafiek in 1960?

>b Waarom wordt het aantal misdaden per 100.000 inwoners gegeven, en niet gewoon het totaal aantal misdaden?

Een fabrikant van beveiligingsapparatuur gebruikt in zijn advertentie de volgende grafiek:



misdaad
verdrievoudigd
STOP
die stijging !!!

>c Geef commentaar op de grafiek van de fabrikant.

De politie, die wilde aantonen dat de criminaliteit de laatste tijd toch was teruggedrongen, toonde een heel andere grafiek.

>d Maak een grafiek die laat zien dat de criminaliteit de laatste jaren alleen maar flink is teruggelopen.

Grafische verwerking van getallen (data) kan op vele manieren. Enkele ervan heb je gezien. Tevens is duidelijk dat conclusies trekken en het juist gebruiken van grafieken niet eenvoudig is.
Een tamelijk nieuwe manier van grafieken tekenen is de box-plot-grafiek (ongeveer in 1980 uitgevonden).
We gaan daartoe even terug naar de cavia-proeven.

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								

Zo'n tabel valt in een paar woorden samen te vatten:
— De middelste waarneming; MEDIAAN.
Dat zijn er hier twee: 102 en 103 (namelijk de 36ste en de 37ste). We nemen dan het midden: 102,5.
— Door bij elk van de twee helften weer de mediaan te nemen krijgen we de tabel in vier gelijke stukken verdeeld:

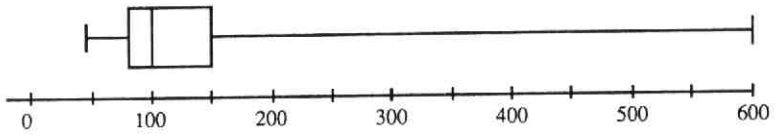
43	45	53	56	56	57	58	66	67	73
74	79	80	80	81	81	81	82	83	83
84	88	89	91	91	92	92	97	99	99
100	100	101	102	102	102	103	104	107	108
109	113	114	118	121	123	126	128	137	138
139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511
522	598								

De mediaan van de eerste helft waarnemingen heet het
eerste kwartiel: 82,5.
De mediaan van de tweede helft waarnemingen heet het
derde kwartiel: 151,5

De hele cavia-tabel kan nu als volgt in vijf getallen worden samengevat:

kleinste waarneming	: 43
1e kwartiel (Q_1)	: 82,5
Mediaan	: 102,5
3e kwartiel (Q_3)	: 151,5
grootste waarneming	: 598
of	: (43; 82,5; 102,5; 151,5; 598).

De *box-plot-grafiek* van deze tabel is:



De box-plot-grafiek bestaat uit vier stukken:

In ieder van die vier stukken staat 25% van de waarnemingsgetallen, beginnend met de 25% kleinste en eindigend met de 25% grootste waarnemingen.

4. Uit de box-plot-grafiek van de cavia's volgt dat de verdeling van de getallen niet klokvormig kan zijn.

   >a Hoe kun je dat zien?

   >b Maak op basis van de box-plot-grafiek een globaal frequentiepolygoon.

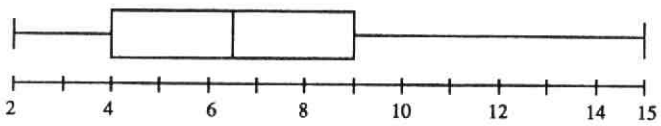
   >c Vergelijk het resultaat met het histogram dat je bij opgave 7 van het vorige hoofdstuk hebt gemaakt.

De mediaan en de kwartielen kunnen alleen bepaald worden als de waarnemingen eerst gerangschikt zijn van klein naar groot.
Voorbeeld:
Bepaal de vijf karakteristieke box-plot-getallen voor de volgende serie waarnemingen: 15, 7, 11, 3, 3, 9, 10, 5, 2, 7, 3, 8, 6, 6, 4, 7, 5, 11
Gerangschikt van klein naar groot wordt de rij getallen:

2	3	3	3	4	5	5	6	6	7	7	7	8	9	10	11	11	15
↑				↑				↑					↑				↑
kleinste				Q_1				mediaan					Q_3				grootste

Er is geen middelste getal, dus de mediaan is $6\frac{1}{2}$.

De bijbehorende box-plot-grafiek:



5. >a 'Vijftig procent van de waarnemingen ligt tussen 4 en 9'.
   Is dat waar?
   >b 'De 25% kleinste waarnemingen liggen meer gespreid dan de 25% grootste waarnemingen.
   Klopt dat?

6. Hieronder staan de inwonersaantallen (in duizendtallen) vermeld van de vijftig grootste Nederlandse gemeenten.
   Achter de plaatsnamen staat tussen haakjes het rangnummer. Amsterdam, de grootste, is nr. 1; Oss is met 50.000 inwoners de laatste; nr. 50.

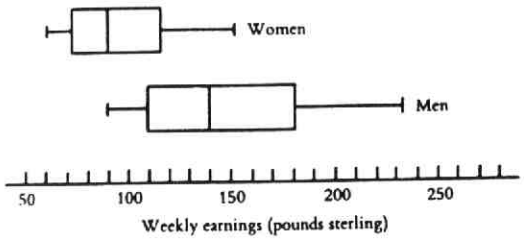
Alkmaar (26)	85	Gouda (39)	60	Nijmegen (9)	146
Almelo (36)	63	's-Gravenhage (3)	443	Oss (50)	50
Alphen (44)	55	Groningen (6)	168	Roosendaal (42)	57
Amersfoort (22)	88	Haarlem (8)	151	Rotterdam (2)	571
Amstelveen (32)	68	Haarlemmermeer (27)	85	Schiedam (31)	69
Amsterdam (1)	676	Heerlen (18)	94	Smallingerland (49)	51
Apeldoorn (11)	145	Den Helder (34)	64	Spijkenisse (41)	57
Arnhem (13)	128	Helmond (37)	62	Tilburg (7)	154
Breda (14)	119	Hengelo (29)	77	Utrecht (4)	230
Capelle (46)	54	's-Hertogenbosch (20)	89	Velsen (40)	58
Delft (25)	87	Hilversum (24)	87	Venlo (35)	63
Deventer (33)	65	Hoorn (48)	52	Vlaardingen (30)	76
Dordrecht (16)	107	Kerkrade (47)	53	Zaanstad (12)	128
Ede (21)	88	Leiden (17)	105	Zeist (38)	60
Eindhoven (5)	192	Lelystad (43)	57	Zoetermeer (28)	80
Emmen (19)	91	Maastricht (15)	114	Zwolle (23)	88
Enschede (10)	145	Nieuwegein (45)	55		

>a Bepaal de mediaan, de kwartielen Q_1 en Q_3 en de kleinste en grootste waarde en teken de box-plot-grafiek.

>b Is het gemiddeld aantal inwoners van deze vijftig gemeenten groter of kleiner dan de mediaan?

>c Stel dat je iemand een indruk wilt geven van het aantal inwoners van de vijftig grootste gemeenten in ons land.
Welk aantal zal je dan noemen: de mediaan of het gemiddelde? Waarom?

7. De volgende box-plot-grafieken geven de inkomens van mannen en vrouwen in Engeland weer.



Twee uitspraken:

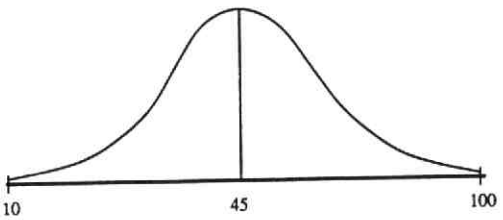
A. 'Bijna 50% van alle mannen verdient meer dan het topsalaris van de vrouwen'.

B. 'Alle mannen verdienen meer dan de 50% laagstbetaalde vrouwen'.

>a Zijn deze twee uitspraken uit de box-plot-grafieken af te leiden?

>b Bekijk voor zowel mannen als vrouwen de middelste 50% van de salarissen die worden verdiend.
Tussen welke grenzen liggen die salarissen?

8. > Teken een box-plot-grafiek bij de volgende klok-vormige verdeling.



9. *Wiskunde loont*

Wiskundigen verdienen ook geld. Als docent en onderzoeker op een Universiteit of School, bij de Regering, maar in toenemende mate ook bij het Bedrijfsleven. De volgende twee tabellen geven de salarissen aan van jonge wiskundigen bij de Universiteit (Tabel I) en bij het Bedrijfsleven (Tabel II) in de U.S.A. Je ziet de ontwikkeling van 1960 tot en met 1987 en voor de laatste jaren ook nog uitgesplitst naar mannen (M) en vrouwen (F). De salarissen zijn weergegeven door de karakteristieke box-plotgetallen:

(Min, Q_1, Median, Q_3, Max)

of

(laagste waarde, 1e kwartiel, mediaan, 3e kwartiel, hoogste waarde)

Tabel I

Year	Min	Q_1	Median	Q_3	Max
	TEACHING OR TEACHING AND RESEARCH (27 + 8)				
1960			NO DATA		
1965	78		104		121
1970	95		128		200
1975	87		145		204
1980	143		195		350
1982	100		250		500
1983	160		260		320
1984	134		260		450
1985	220	230	273	300	470
1986	220	265	320	360	480
1987	200	283	315	357	520
1984M	134		260		450
1984F	240		275		330
1985M	230	235	240	300	470
1985F	220	243	280	295	420
1986M	220	270	321	360	480
1986F	240	245	285	340	360
1987M	200	270	300	358	520
1987F	300	320	339	357	450

Tabel II

Year	Min	Q_1	Median	Q_3	Max
	BUSINESS AND INDUSTRY (30 + 12)				
1960	78		110		150
1965	100		136		180
1970	96		170		235
1975	114		187		240
1980	190		284		400
1982	198		354		550
1983	276		375		580
1984	180		378		660
1985	260	360	400	420	493
1986	324	373	425	477	750
1987	290	400	451	500	1500
1984M	180		383		660
1984F	200		342		416
1985M	260	360	400	425	493
1985F	295	330	370	409	430
1986M	324	390	453	492	750
1986F	350	357	375	400	440
1987M	290	400	465	517	1500
1987F	300	394	424	466	502

>a Teken een aantal box-plots van de salarissen van wiskundigen die duidelijk maken:
- het verschil tussen 1965 en 1987;
- het verschil tussen Universiteit en Bedrijfsleven (zowel in 1965 als in 1987);
- het verschil tussen mannen en vrouwen;

>b Zijn de volgende conclusies te verdedigen op basis van deze gegevens:
— Mannen verdienen altijd meer dan vrouwen.
— Het bedrijfsleven betaalt beter dan het onderwijs.

Samenvatting

Naast het histogram wordt ook het (frequentie)polygoon en de box-plot-grafiek vaak gebruikt. Vaak ook een 'globale' grafiek.
Dat voorzichtigheid geboden blijft, blijkt ook uit dit hoofdstuk: bloeddruk neemt toe met de leeftijd, niet met het aantal kinderen. Wel is het zo dat vrouwen met vier kinderen gemiddeld ouder zijn dan vrouwen met twee kinderen.
Oppassen ook bij grafieken die iets moeten bewijzen: door kleine manipulaties kan het resultaat drastisch veranderen - denk aan de criminaliteitsgrafiek.

De box-plot-grafiek is te karakteriseren met vijf getallen:

(laagste waarde, 1e kwartiel, mediaan, 3e kwartiel, grootste waarde)

Aan een box-plot grafiek kun je goed zien of de verdeling wel of niet klokvormig is.

Bij scheve (dus niet-klokvormige) verdelingen valt de mediaan *niet* samen met het gemiddelde.

De mediaan en de kwartielen kunnen alleen maar bepaald worden als de waarnemingsgetallen eerst gerangschikt zijn van klein naar groot.

5 Middelste en gemiddelde

De middelste waarneming (*mediaan*) verschilt in het algemeen van het *gemiddelde* van de waarnemingen. Het gemiddelde vind je door de som van alle waarnemingen te delen door het aantal waarnemingen.
Beide worden *centrummaten* genoemd.
Centrummaten worden vaak gebruikt om een grote serie waarnemingsgetallen door één enkel getal vast te leggen.

Voorbeelden:
Het rapportcijfer voor wiskunde is 6 (het gemiddelde; meestal tenminste).
De helft van de cavia's sterft binnen 102 dagen (de mediaan).

1. Het jaarcijfer voor wiskunde wordt berekend aan de hand van alle proefwerkcijfers van het afgelopen jaar.
   De negen proefwerkcijfers waren:
   $7,\ 6\frac{1}{2},\ 2,\ 7\frac{1}{2},\ 7,\ 7^{+},\ 7\frac{1}{2},\ 6\frac{1}{2},\ 7^{-}$

   >a Wat wordt het jaarcijfer als de docent het afgeronde gemiddelde als cijfer gebruikt?
   >b Welk cijfer krijg je, als hij de mediaan van de proefwerkcijfers neemt?
   >c Welke centrummaat geeft in dit geval het eerlijkste beeld van je prestaties voor wiskunde?
   >d Geef een serie proefwerkcijfers, waarbij het gemiddelde hoger uitvalt dan de mediaan.

2. *Hardlopen*
   Een hardloopster loopt driemaal per week hetzelfde parcours over duinen en strand. De laatste negen keer had ze de volgende tijden (over 11 km):
   56; 55; 68; 57; 58; 55; 54; 66; 57 minuten.
   De tijden variëren nogal door de invloed van het strand, dat soms heel rul is.

   >a Bereken de gemiddelde tijd.
   >b Maak een box-plot grafiek; geef daarin ook de gemiddelde tijd aan.

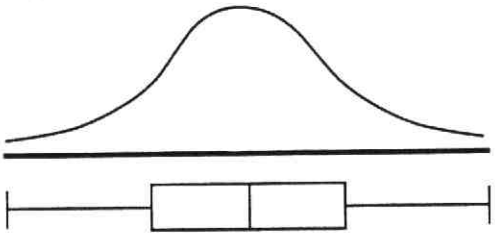
   Het gemiddelde ligt tussen de mediaan en het derde kwartiel (Q_3).
   In principe is het mogelijk dat het gemiddelde nog groter is dan Q_3.

   >c Verander één van de negen bovengenoemde tijden zó dat het gemiddelde rechts van Q_3 komt te liggen.

3. In Cook County, Verenigde Staten, is bijgehouden hoe groot de schadevergoedingen waren die aan mensen werden toegekend in verband met medische fouten, defecte apparaten, verwondingen, enz.
   De mediaan : 8000 dollar.
   Het gemiddelde : 69000 dollar.

   >a Hoe verklaar je dat grote verschil tussen de twee centrummaten?
   >b Teken een globale grafiek van de verdeling van de schadevergoedingen.

Als het globale histogram klokvormig is, dan zijn gemiddelde en mediaan ongeveer even groot:



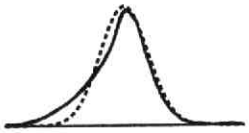
Zowel de globale grafiek als de box-plot-grafiek zijn dan symmetrisch.

4. In de volgende plaatjes worden globale grafieken bekeken, waarbij steeds iets is gewijzigd in de klokvorm. De oorspronkelijke klokvormige verdeling is elke keer met behulp van een stippellijn erbij getekend.

A

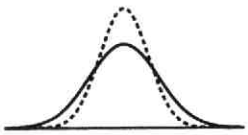
scheve verdeling met 'staart' naar rechts

B

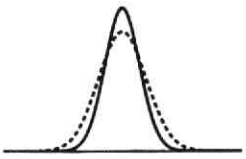


scheve verdeling met 'staart' naar links

C



D

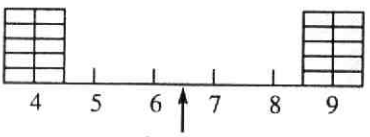


   >a Geef in elk van de gevallen aan hoe het gemiddelde ligt ten opzichte van de mediaan.
   >b Schets in elk van de vier gevallen een box-plot-grafiek.

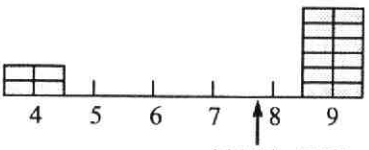
5. Het gemiddelde van 4 en 8 is 6.
Het gemiddelde van een aantal vieren en een aantal achten kan ook 6 zijn.

>a Welk verband bestaat er dan tussen het aantal vieren en het aantal achten?
>b Wat is het gemiddelde van 10 vieren en 30 achten?
>c Wat kun je zeggen over de aantallen vieren en achten, als het gemiddelde 5 is?

Bij het berekenen van een gemiddelde zijn niet alleen de waarnemingsgetallen zelf van belang, maar ook het aantal keren dat ieder van die getallen voorkomt (de *frequentie* van die getallen).
Een getal dat vaak voorkomt legt bij berekening van het gemiddelde meer 'gewicht' in de schaal dan een getal dat minder vaak voorkomt.
Een serie van 10 vieren en een serie van 10 negens zijn in balans ten opzichte van 6,5. Bij 4 vieren en 12 negens ligt het evenwicht bij 7,75.



gemiddelde: 6,5



gemiddelde: 7,75

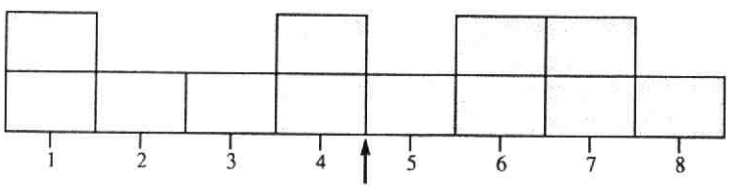
Op een andere manier gezegd:

Beide getallen leveren een bijdrage van 50% aan het gemiddelde.

Dus:
gemid. = 50% van 4 + 50% van 9
= 2 + 4,5
= 6,5

De bijdrage van de 4 aan het gemiddelde is 25%.
De 9 doet voor 75% mee.
Dus:
gemid. = 25% van 4 + 75% van 9
= 1 + 6,75
= 7,75

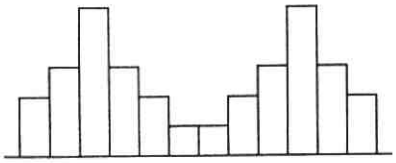
6.



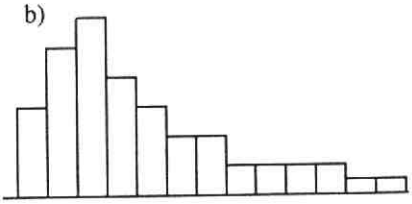
> Kun je, zonder berekening, beredeneren dat het gemiddelde hier precies in het midden ligt?

7. Bepaal in elk van de vier onderstaande gevallen de plaats van het gemiddelde en van de mediaan.

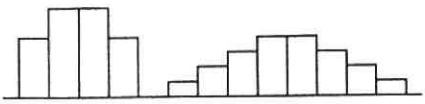
a)



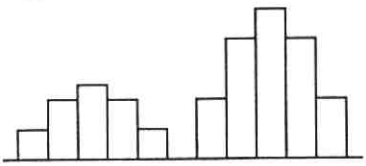
b)



c)



d)



8. Gegevens over gezinsgrootte (CBS-jaarboek, 1983):

Aantal kinderen per gezin	Aantal gezinnen met dit kindertal (in duizendtallen)
0	1176
1	810
2	1016
3	417
4	149
5	59
6	23
7 of meer	16

Het gemiddeld aantal kinderen per gezin is alleen uit te rekenen als '7 of meer' nauwkeuriger wordt vastgelegd.

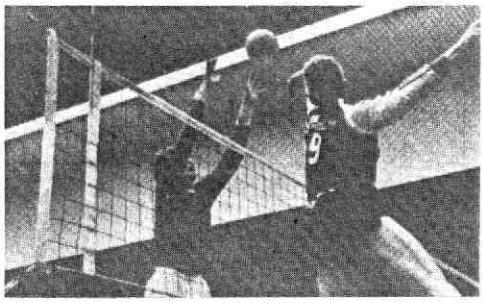
>a Bereken het gemiddelde als '7 of meer' vervangen wordt door '7'.

>b Hoe verandert het gemiddelde, als '7 of meer' gelezen wordt als '8'? En als er '10' gelezen wordt in plaats van '7 of meer'?

Kennelijk maakt het in dit geval niet zo veel uit wat er voor '7 of meer' gelezen wordt.

>c Hoe verklaar je dat?

Bij een mini-onderzoek onder leerlingen van 3 havo en 3 mavo werd gevraagd naar het bedrag dat per weekend besteed wordt aan ontspanning (disco, film, sport, ...).



De leerlingen die daaraan niets uitgeven, zijn bij dit onderzoek buiten beschouwing gelaten.
Het resultaat:

Uitgaven in guldens	3 mavo		3 havo	
	jongens	meisjes	jongens	meisjes
< 10	4	3	6	7
10 - 20	8	8	0	1
20 - 30	2	1	2	0
≥ 30	0	0	0	0

Het bedrag dat gemiddeld besteed wordt, is nu niet precies uit te rekenen, omdat er alleen maar globale gegevens zijn.
Zo zijn er acht meisjes uit 3 mavo die elk een bedrag uitgeven dat ligt tussen *f* 10,- en *f* 20,-.
In de praktijk wordt vaak met dit soort globale gegevens gewerkt. Om toch een redelijk goed idee te krijgen van het gemiddelde, wordt in die gevallen gerekend met het *klasse-midden*. Voor de klasse 10 - 20 is dat *f* 15,-.
Voor alle personen die bij die klasse genoemd worden, nemen we aan dat ieder van hen *f* 15,- uitgeeft.

9. >a Wanneer is deze aanname redelijk?
   >b Door deze aanname kan er een verschil ontstaan tussen het *berekende* gemiddelde en het *werkelijke* gemiddelde.
   Hoe groot is dat verschil maximaal?

10. >a Bereken met behulp van de klasse-middens, het bedrag dat de meisjes uit 3 mavo gemiddeld uitgeven.
    >b Geven de 3 havo-leerlingen gemiddeld meer uit dan de 3 mavo-leerlingen?

11. Regelmatig worden via steekproeven de lichaamslengtes van de Nederlanders van 18 jaar en ouder gemeten.
De gegevens van 1986 voor de mannen waren:

	18-29 j.	30-39 j.	40 49 j.	50-59 j.	60-69 j.	≥ 70 j.	totaal
Mannen	%						
-167 cm	3.0	4,2	7,2	9,4	10,9	17,2	6,8
168-172 cm	9,8	15,6	18,7	25,1	26,7	31,2	18,3
173-177 cm	16,6	20,6	26,0	25,3	27,1	21,9	21,9
178-182 cm	26,8	27,7	24,2	23,7	22,0	20,2	25,1
183-187 cm	25,1	21,3	16,2	11,2	9,9	7,6	17,8
188-192 cm	11,5	7,2	5,4	4,3	2,6	1,3	6,6
≥ 193 cm	7,2	3,5	2,3	0,9	0,8	0,6	3,4
streekproefaantal abs (= 100%)	2563	2461	1673	1378	1020	782	9877
Gemiddelde lengte (cm)	181,3	179,0	177,3	175,8	175,0	173,4	178,0

Bekijk de eerste kolom.
168 - 172 cm betekent: alle lengtes van 168 cm tot aan 173 cm.
Dus elke klasse heeft een *breedte* van 5 cm.

>a Welk klasse-midden hoort bij de klasse 168 - 172 cm?

De eerste en de laatste klasse zijn *open*:
- 167 cm betekent: alle lengtes kleiner dan 168 cm.
≥ 193 cm betekent: alle lengtes vanaf 193 cm.
Om het gemiddelde uit te kunnen rekenen moeten we ook voor deze open klassen iets afspreken:
Neem voor de klasse - 167 cm als klasse-midden $165\frac{1}{2}$
en voor de klasse ≥ 193 cm als klasse-midden $195\frac{1}{2}$.

>b Bereken de gemiddelde lengte van mannen van 18 - 29 jaar.

Het gemiddelde dat vermeld staat (181,3 cm) is berekend aan de hand van de 2563 lengtes die in de steekproef voorkwamen.

>c Het bij >b berekende gemiddelde komt niet precies uit op 181,3 cm.
Noem een paar oorzaken voor het verschil in de uitkomsten.

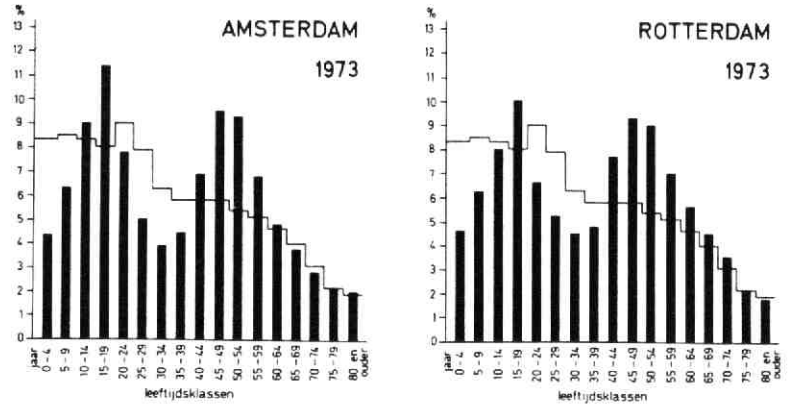
De gemiddelde lengte van ouderen is kleiner dan van jongeren. Dat is direct af te lezen in de rij 'gemiddelde lengte'.

>d Vergelijk de percentages bij de kolommen 18 - 29 jaar en 50 - 59 jaar.
Is het mogelijk om, zonder berekening, daaruit te concluderen dat de gemiddelde lengte voor de leeftijdsgroep 18 - 29 jaar groter is dan de gemiddelde lengte voor de leeftijdsgroep 50 - 59 jaar?

>e De gemiddelde lengte van alle mannen is op twee verschillende manieren uit de tabel te berekenen.
Hoe?

Gemiddelde en mediaan zijn de meest gebruikte centrummaten.
Soms is er behoefte aan een derde centrummaat: de *modus*.
Dat is het waarnemingsgetal (of klasse van waarnemingen) dat het meest voorkomt.
Wanneer, bijvoorbeeld, een gemeente het woningbeleid wil afstemmen op de gezinsgrootte, dan geeft het gemiddelde geen goede informatie.
Het gemiddeld aantal kinderen per gezin is 1,4 (zie opgave 8). Voor het bouwen van nieuwe woningen is het belangrijk om te weten dat gezinnen met 0 kinderen of met 2 kinderen het meest voorkomen.

12.



Leeftijdsopbouw in 1973 van de bevolking van de in de jaren vijftig gerealiseerde nieuwbouwwijken in Amsterdam en Rotterdam vergeleken met die in westelijk Nederland (1972).

De staven geven de leeftijdsopbouw in de nieuwbouwwijken weer. Voor westelijk Nederland is de opbouw getekend met een doorgetrokken lijn.
De opbouw in Amsterdam is vrijwel gelijk aan die in Rotterdam. Beide verschillen sterk van de opbouw in heel westelijk Nederland.

>a Hoe kun je dat verklaren?
>b Bepaal voor de grote steden en voor westelijk Nederland de mediaan en de modus.
>c Vergelijk de gemiddelde leeftijd in de nieuwbouwwijken met die van westelijk Nederland.

Samenvatting

Voor het vastleggen van een serie waarnemingen in één getal, worden de volgende centrummaten gebruikt:

het gemiddelde: alle waarnemingen optellen en vervolgens de uitkomst delen door het *aantal* waarnemingen.

de mediaan: het middelste waarnemingsgetal, als de getallen gerangschikt zijn van klein naar groot.

de modus: de waarneming die het meest voorkomt.

Bij het gebruik van klassen van waarnemingen wordt het gemiddelde berekend met behulp van de *klasse-middens*.

Bij frequentietabellen wordt ook vaak gebruik gemaakt van *relatieve frequenties* (of: procentuele frequenties).

Voorbeeld:

waarneming	absolute frequentie
3	5
4	12
5	6
6	2

→

waarneming	relatieve frequentie (in %)
3	20
4	48
5	24
6	8

gem. $= \frac{5*3 + 12*4 + 6*5 + 12*6}{25}$

$= 4{,}2$

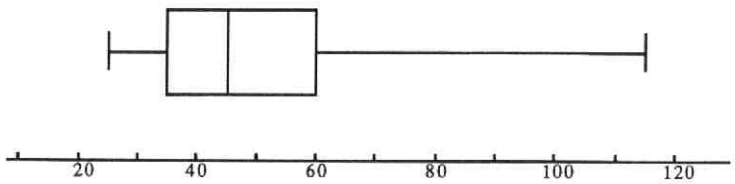
gem. $= 0{,}20*4 + 0{,}48*4 + 0{,}24*5 + 0{,}08*6$

$= 4{,}2$

6 Spreidingsmaten

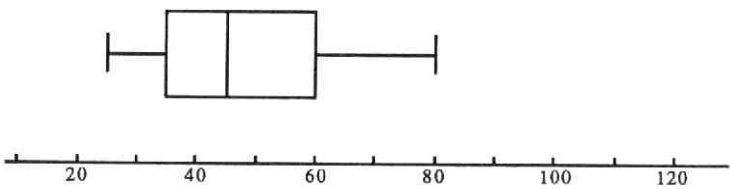
Met behulp van een centrummaat wordt een serie waarnemingen als het ware samengevat in één getal. Aan dat getal is niet te zien hoe de waarnemingen verspreid liggen. Daarom wordt bij een centrummaat meestal ook de *spreiding* vermeld.
Een voor de hand liggende maat daarvoor is de *absolute spreiding*: de afstand tussen het grootste en het kleinste waarnemingsgetal.

1. Een box-plot-grafiek van een serie van 150 waarnemingen.



>a Hoe groot is de absolute spreiding?
>b Hoe groot is de afstand tussen eerste kwartiel (Q_1) en derde kwartiel (Q_3)?

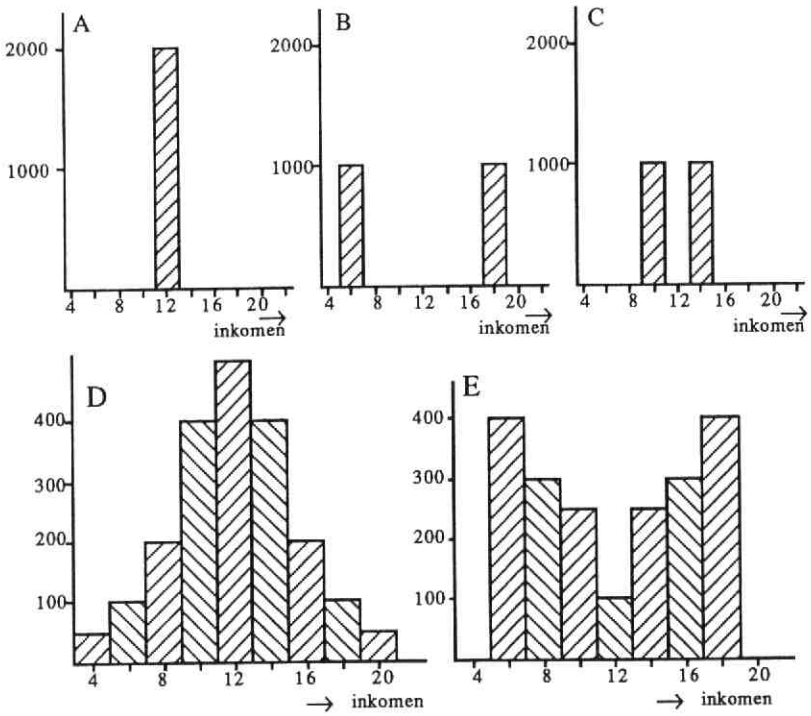
Uit die 150 waarnemingen wordt de grootste weggelaten.
De box-plot-grafiek van de resterende 149 waarnemingen is:



>c Welk effect heeft weglaten van de grootste waarneming hier op de absolute spreiding?
En op de afstand van Q_1 tot Q_3?
>d Hoe verklaar je dat?

Bij de mediaan wordt de afstand van Q_1 tot Q_3 (de zogenaamde *interkwartiele afstand*) als spreidingsmaat voor de waarnemingen gebruikt.
Deze spreidingsmaat geeft aan hoe wijd de middelste 50% van de waarnemingsgetallen verspreid liggen.
In tegenstelling tot de absolute spreiding wordt deze maat bijna niet beïnvloed door eventuele uitschieters (zie opgave 1).

2. Vijf histogrammen met verdelingen van de inkomens van telkens 2000 artsen.



In alle gevallen geldt dat het gemiddelde inkomen 12 is.

>a Hoe zie je dat, zonder berekening?

De absolute spreiding van de inkomens is heel verschillend.

>b Bij welk histogram is de absolute spreiding het grootst?
En bij welk het kleinst?

Bij de histogrammen *B* en *E* is de absolute spreiding gelijk.
Toch zijn de verdelingen heel verschillend. Bij *B* liggen alle inkomens ver van het gemiddelde vandaan. Bij *E* is dat veel minder het geval.
De absolute spreiding geeft niet weer *hoeveel* waarnemingen ver van het gemiddelde vandaan liggen of juist er dichtbij.

Een spreidingsmaat die deze eigenschap wel vertoont is de *gemiddelde absolute afwijking* (g.a.a.).
Deze wordt berekend door de afwijkingen ten opzichte van het gemiddelde van *alle* waarnemingen op te tellen en vervolgens te delen door het aantal waarnemingen.
Een voorbeeld:

Serie A: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Serie B: 2, 5, 5, 6, 6, 7, 8, 8, 9, 9, 12

Beide series hebben 7 als gemiddelde.

Voor serie A zijn de afwijkingen ten opzichte van het gemiddelde achtereenvolgens:

5, 4, 3, 2, 1, 0, 1, 2, 3, 4, 5.

De g.a.a. van serie A is dus:

$$\frac{5+4+3+2+1+0+1+2+3+4+5}{11} = \frac{30}{11} \approx 2{,}7.$$

3. Voor de afwijkingen zou ook genomen kunnen worden:
   -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5
   Een minteken betekent dat het waarnemingsgetal links van het gemiddelde ligt.

   > Waarom is dat niet zinvol?

4. >a Bereken de g.a.a. van serie B.
   >b De g.a.a. van serie B is kleiner dan de g.a.a. van A.
   Is dat terecht?

5. Bij de vijf histogrammen van opgave 2 zijn de gemiddelde absolute afwijkingen allemaal verschillend.

   >a Zet (zonder berekening) de histogrammen in de volgorde van kleinste g.a.a. tot grootste g.a.a.
   >b Bereken de g.a.a. bij de histogrammen D en E.
   Vergelijk de twee uitkomsten.

De g.a.a. geeft een goed beeld van de spreiding van de waarnemingen ten opzichte van het gemiddelde.
Toch wordt in de praktijk meestal met een andere maat gewerkt:
de *standaardafwijking* (afgekort tot *S.D.* = Standard Deviation).
De berekening van de *S.D.* is een tamelijk vreemd verhaal.

Voorbeeld:
Bereken het gemiddelde en de *S.D.* van de serie 2, 3, 3, 4, 5, 7, 7, 8, 10, 11

1. Bereken het gemiddelde: $\frac{2+3+3+4+5+7+7+8+10+11}{10} = 6$.

2. Bepaal de afwijking van alle waarnemingen ten opzichte van het gemiddelde:
   4 3 3 2 1 1 1 2 4 5

3. Kwadrateer alle afwijkingen:
   16 9 9 4 1 1 1 4 16 25

4. Tel ze op en deel door het *aantal* waarnemingen. (zo krijg je de 'gemiddelde gekwadrateerde afwijking'):
   $\frac{16+9+9+4+1+1+1+4+16+25}{10} = \frac{86}{10} = 8{,}6$

5. Neem de wortel:
   $\sqrt{8{,}6} \approx 2{,}9$

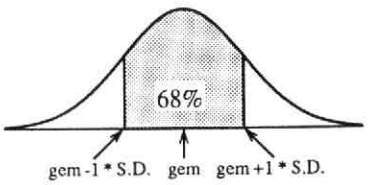
6. > Bereken de *S.D.* van de series A en B (zie blz. 44).

7. >a Bepaal gemiddelde en *S.D.* van
   — 1, 2, 3, 4, 5
   — 11, 12, 13, 14, 15
   >b Geef, zonder berekening, gemiddelde en *S.D.* van
   — 126, 127, 128, 129, 130.

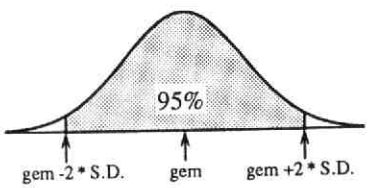
8. >a Welke van de volgende twee series waarnemingen heeft de grootste *S.D.*?
   A: 9, 12, 10, 10, 9, 10
   B: 7, 8, 11, 13, 10
   >b Controleer je antwoord door van beide series het gemiddelde en de standaardafwijking te berekenen.

Het lijkt vreemd om een spreidingsmaat in te voeren die zo moeilijk te berekenen is, terwijl er een goed alternatief (de g.a.a.) voorhanden is.
De standaardafwijking speelt een belangrijke rol wanneer de verdeling van de waarnemingen min of meer klokvormig is. In dat geval heeft de *S.D.* een paar prettige eigenschappen.
Bij een klokvormige verdeling geldt:

— ongeveer 68% van alle waarnemingen wijkt minder dan 1 maal de *S.D.* van het gemiddelde af.



— ongeveer 95% van de waarnemingen ligt minder dan 2 maal de *S.D.* van het gemiddelde verwijderd.



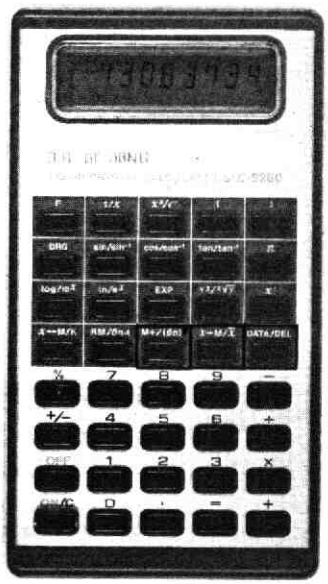
Bij het onderzoek 'pilgebruik-bloeddruk' (hoofdstuk 4) kwamen onder andere de volgende twee groepen vrouwen voor:

bloeddruk (in mm)	users (25-34)	non-users (17-24)
90-100	4	4
100-110	15	21
110-120	32	35
120-130	25	23
130-140	15	12
140-150	6	5
150-160	2	-
totaal (%)	99	100
totaal aantal	1747	1206

De gemiddelde bloeddruk van de pilgebruiksters in de leeftijd van 25-34 jaar is 121 mm. De standaardafwijking voor deze groep bedraagt 13 mm.
Beide getallen zijn afgerond op hele millimeters.

De berekening ervan is een hele klus.
Gelukkig kan het ook met de rekenmachine.
Daarop zit een toets (meestal 'S.D.' of 'STAT') waarmee de rekenmachine ingesteld wordt op statistische berekeningen.
Alle waarnemingsgetallen moeten nu worden ingevoerd.
Dat gaat als volgt:

toets het eerste waarnemingsgetal in en vervolgens de knop [x] (soms ook: [DATA]), daarna de volgende waarnemingsgetallen op dezelfde manier.
Komt een waarnemingsgetal meerdere keren voor, dan kan dat zo worden ingevoerd:
waarnemingsgetal [*] aantal [x]



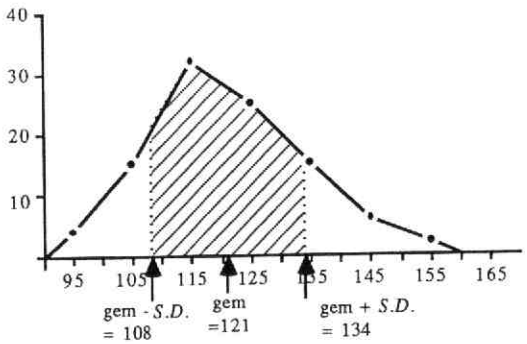
Bij de groep 25-34 jaar nemen we als waarnemingsgetallen de klassemiddens.
De invoer verloopt dan als volgt:

95	[*]	4	[x]
105	[*]	15	[x]
		...	
		...	
		...	
155	[*]	2	[x]

Alle getallen zijn nu ingevoerd.
Het gemiddelde krijg je te zien met de toets [$\bar{x}$] en de standaardafwijking met de toets [σn] .

9. >a Controleer met je rekenmachine het gemiddelde (121 mm) en de *S.D.* (13 mm).

Het frequentiepolygoon voor de leeftijdsgroep 25-34.



Omdat de verdeling min of meer klokvormig is, zal ongeveer 68% van de groep vrouwen een bloeddruk hebben die ligt tussen 108 en 134 mm.

>b Controleer in de tabel of dat klopt.

Verder moet ongeveer 95% van de vrouwen een bloeddruk hebben die ligt tussen 95 (= 121 – 2 * 13) en 147 (= 121 + 2 * 13).

>c Controleer dit ook met de tabel.

10. >a Bereken gemiddelde en *S.D.* voor de non-users in de leeftijdsgroep 17-24. Gebruik je rekenmachine.
    >b Teken voor deze groep vrouwen het frequentiepolygoon en kleur daarin het gedeelte dat ligt tussen de grenzen 'gemiddelde – 1 * *S.D.*' en 'gemiddelde + 1 * *S.D.*'.

11. Het resultaat van een steekproef (zie hoofdstuk 3) was:

'45% van de mensen voelt zich op straat onveilig'.

Het bureau dat de enquêtering verrichtte, tekende daarbij aan:
— de waarde 45% dient gelezen te worden als: ergens tussen de 42% en 48%.
— In ongeveer één op de twintig gevallen kan de 'werkelijke' waarde zelfs tussen het 42-48 interval liggen.

> Met welke standaardafwijking houdt dit bureau kennelijk rekening?

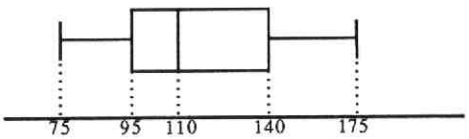
De klokvormige verdelingen blijken in de praktijk vaak op te treden. Daarom zullen we er in een apart boekje uitgebreid aandacht aan besteden.

Samenvatting

Spreidingsmaten geven, gecombineerd met centrummaten een goed beeld van de verdeling van een grote serie waarnemingen.

De *absolute spreiding* geeft de afstand van grootste tot kleinste waarneming. Omdat hij zo gevoelig is voor uitschieters wordt hij weinig gebruikt.

De *interkwartiele afstand* geeft aan hoever de middelste 50% van de waarnemingen uiteen liggen.
Gekoppeld aan de mediaan is dit een veel gebruikte spreidingsmaat.



mediaan: 110
interkwartiele afstand (Q_3 - Q_1): 45

De *gemiddelde absolute afwijking* (g.a.a.) wordt gebruikt bij het gemiddelde als centrummaat.
De g.a.a. geeft aan hoeveel alle waarnemingen gemiddeld van het gemiddelde afwijken.

Wanneer de verdeling bij benadering klokvormig is, wordt de *standaardafwijking* (*S.D.*) als spreidingsmaat gebruikt.
Daarbij gelden de volgende vuistregels: