Statistiek voor Informatiekunde (I00099)

Bernd Souvignier

voorjaar 2005

Inhoud

Les 1 Beschrijvende statistiek 2
- 1.1 Representatie van gegevens 2
- 1.2 Klassen 4
- 1.3 Typische waarden 8
- 1.4 Spreiding 13
- 1.5 Momenten 16

Aanbevolen literatuur

- Larray Gonick, Woollcott Smith: The Cartoon Guide to Statistics. HarperResource, 1993, 240 p., ISBN: 0-06-273102-5
  nederlandse versie hiervan:
  Larray Gonick, Woollcott Smith: Het stripverhaal van de statistiek. Epsilon Uitgaven 32, 2004, 240 p., ISBN: 90-5041-037-5

- Murray R. Spiegel, Larry J. Stephens: (Schaum's Outline of Theory and Problems of) Statistics. McGraw-Hill Companies, 1999, 512 p., ISBN: 0-07-060281-6.

Les 1 Beschrijvende statistiek

In de statistiek gaat het erom, vanuit waargenomen gegevens een model te ontwikkelen dat de gegevens goed kan verklaren. Meestal houdt het model een kansverdeling in, daarom bestaat er een grote overlap tussen de methoden van de statistiek en van de kansrekening. Het verschil ligt erin dat men in de kansrekening een proces veronderstelt dat volgens een kansverdeling waarden met zekere kansen produceert, terwijl men in de statistiek van gegevens uitgaat die een zekere frequentieverdeling hebben en probeert conclusies over een hier achter liggende kansverdeling te trekken. In zekere zin bekijken dus kansrekening en statistiek dezelfde vraagstukken uit verschillende invalshoeken.

1.1 Representatie van gegevens

In de statistiek gaat het vooral om het onderzoeken van gegevens die op een of ander manier verzameld zijn, bijvoorbeeld door een of meerdere metingen of door een enquête. Om uitspraken over de gegevens te kunnen doen en structuren erin te kunnen herkennen, is het belangrijk om overzicht over de gegevens te krijgen.

Voorbeeld: We zullen in deze les vaker naar het volgende voorbeeld van gegevens kijken (resultaten bij een zekere toets):

$$54, 41, 59, 45, 34, 49, 58, 30, 61, 47, 43, 48, 80, 27, 56, 45.$$

Meestal is het niet zo handig, de gegevens gewoon op een rij te zetten, omdat de structuur dan verborgen blijft. Daarom worden verschillende manieren toegepast om gegevens grafisch te representeren.

We gaan er nu van uit dat we het over gegevens hebben, die numerieke waarden voor een eigenschap van zekere individuën zijn. Denk hierbij aan de uitslagen van studenten bij een tentamen, de lengte van kinderen op tienjarige leeftijd of iets dergelijks. Het is duidelijk dat het beschrijven van de type van de gegevens afhangt, deze kunnen discrete waarden, zo als aantallen hebben, maar ook continue waarden, waar in principe elke waarde mogelijk is. Natuurlijk zijn er ook gegevens die niet numeriek zijn, zo als eigenschappen, hobbies etc., maar deze kunnen we als gegevens met discrete waarden behandelen, door bijvoorbeeld de verschillende mogelijkheden te nummeren.

Maar eigenlijk bestaan er in de praktijk bijna nooit gegevens met echt continue waarden. Als je bijvoorbeeld naar de resultaten van een competitie in het verspringen kijkt, dan zijn die altijd op centimeters nauwkeurig aangegeven, terwijl we toch ook makkelijk millimeters zouden kunnen meten. Hetzelfde geldt voor tijden, die worden bijvoorbeeld bij het zwemmen in honderdste seconden aangegeven, ook al worden ze nauwkeuriger gemeten (namelijk minstens op duizendsten).

> Bij de olympische spelen van München 1972 hadden er over de 400m wisselslag bij het zwemmen de zweed Gunnar Larsson en de amerikaan Tim McKee een tijd van 4:31,98 minuten. Maar er werden ook duizendsten seconden gemeten en de preciezere tijden waren 4:31,981 voor

> Larsson en 4:31,983 voor McKee. Men heeft toen Larsson de gouden en McKee de zilveren medaille toegekend. Maar sindsdien is er besloten, om de metingen achter de honderdste seconden gewoon te negeren en bij een *dead race* twee gouden medailles uit te reiken.

Vaak worden waarden door *afronden* gediscretiseerd, alle waarden die in een zeker interval liggen worden hierbij door dezelfde waarde vervangen. We zouden ons daarom op gegevens met discrete waarden kunnen beperken, maar we zullen zien dat het vaak handig is, een verdeling door een continue functie te beschrijven.

Merk op: Bij het rekenen met afgeronde waarden neemt de nauwkeurigheid (in het algemeen) bij elke bewerking af. Het is daarom verstandig, zo lang mogelijk met hoge nauwkeurigheid te rekenen en pas het uiteindelijke resultaat af te ronden.

Bij het optellen worden de *absolute fouten* bij elkaar opgeteld, want

$$(x + \Delta x) + (y + \Delta y) = (x + y) + (\Delta x + \Delta y).$$

Bij het vermenigvuldigen worden de *relatieve fouten* bij elkaar opgeteld, want uit

$$(x + \Delta x) \cdot (y + \Delta y) = x \cdot y + \Delta x \cdot y + \Delta y \cdot x + \Delta x \cdot \Delta y$$

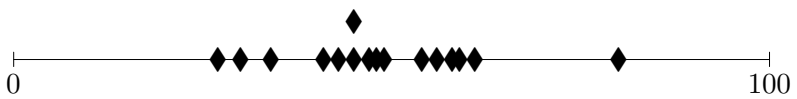
volgt voor $\Delta(x \cdot y) = (x + \Delta x) \cdot (y + \Delta y) - x \cdot y$:

$$\frac{\Delta(x \cdot y)}{x \cdot y} \approx \frac{\Delta x}{x} + \frac{\Delta y}{y}$$

waarbij we de term met twee Δ's hebben weggelaten. Als dus de zijden van een blok met een nauwkeurigheid van 5% gemeten kunnen worden en het volume van de blok als product van de zijden wordt berekend, heeft het volume slechts nog een nauwkeurigheid van 15%.

Stengel-en-blad diagram

Een eenvoudige mogelijkheid om waarden te representeren bestaat erin, de waarden op een lijn te markeren. Dit geeft soms al een overzicht waar de waarden liggen en waar bijvoorbeeld veel punten dicht bij elkaar liggen en hoe ver ze verspreid zijn. Voor ons voorbeeld ziet dit er zo uit:



Natuurlijk is er een probleem als we twee keer dezelfde waarde hebben, wat natuurlijk vooral bij discrete gegevens het geval is. We kunnen dit (zo als in het plaatje) bijvoorbeeld oplossen, door punten voor dezelfde waarde boven elkaar te zetten.

Een representatie die dit idee opneemt is het *stengel-en-blad* diagram, waarbij we alle waarden in een zeker interval naast elkaar schrijven. In het voorbeeld

nemen we het eerste cijfer van een waarde (de tienen) als waarde op de stengel, het laatste cijfer komt dan als blad erachter te staan. Vervolgens worden de bladen die achter een waarde op de stengel staan op volgorde gesorteerd. Voor ons voorbeeld ziet het stengel-en-blad diagram als volgt uit:

2	7						
3	0	4					
4	1	3	5	5	7	8	9
5	4	6	8	9			
6	1						
7							
8	0						

Deze manier om waarden samen te vatten is al een speciaal voorbeeld voor het vormen van *klassen* die we nu gaan behandelen.

1.2 Klassen

Vaak is het handig om verschillende waarden samen te vatten die op een of ander manier op elkaar lijken. De zo samengevatte waarden noemt men dan een *klasse* van waarden. Als voorbeelden van klassen hebben we al intervallen gezien, waarbij alle waarden tussen zekere grenzen in een pot gegooid worden. Maar er zijn ook heel andere klassen mogelijk, bijvoorbeeld kunnen de woorden in een tekst op totaal verschillende manieren in klassen ingedeeld worden:

- aantal letters in het woord;
- aantal klinkers in het woord;
- syntactische klasse (werkwoord, naamwoord, artikel enz.);
- semantische klasse (wiskundig begrip, kleur, uitdrukking van beweging).

Als we eindig veel gegevens op klassen verdelen, krijgen we een frequentieverdeling voor de klassen, en als we naar de relatieve frequenties van de klassen kijken, voldoen deze aan de eisen van een kansverdeling.

> Merk op dat er een subtiel verschil is tussen een kansverdeling en de frequentieverdeling van klassen: Bij een kansverdeling veronderstellen we een proces die waarden met zekere kansen *produceert*, terwijl de frequentieverdeling gewoon een verzameling van gegevens *beschrijft*. Maar natuurlijk is het vaak nuttig een waargenomen frequentieverdeling met bekende kansverdelingen te vergelijken.

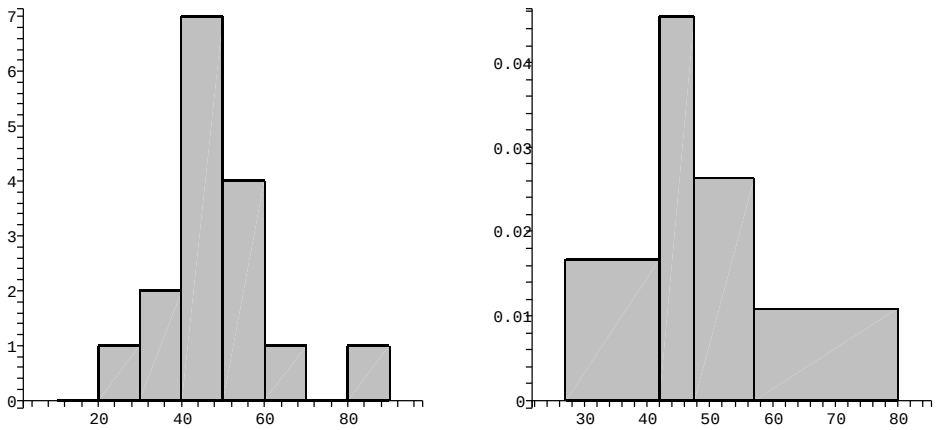
De indeling in klassen is een belangrijke voorwaarde voor de interpretatie van de gegevens. Te veel klassen geven vaak alleen maar versplinterde informatie omdat heel weinig gegevens in een klasse terecht komen, terwijl te weinig klassen geen structuur meer laten herkennen. Soms wordt als vuistregel gehanteerd, bij n gegevens het aantal klassen als $1 + {}^2\log(n)$ te kiezen, maar dit is ook niet meer dan een heuristische gok.

Soms kan zelfs de verschuiving van de klassen kritiek zijn, omdat er een duidelijk grootste klasse over twee ongeveer even grote maar veel kleinere klassen verdeeld wordt.

De frequentieverdelingen van klassen laten zich op verschillende manieren grafisch representeren. We zullen de meest belangrijke vormen kort bespreken.

Histogram

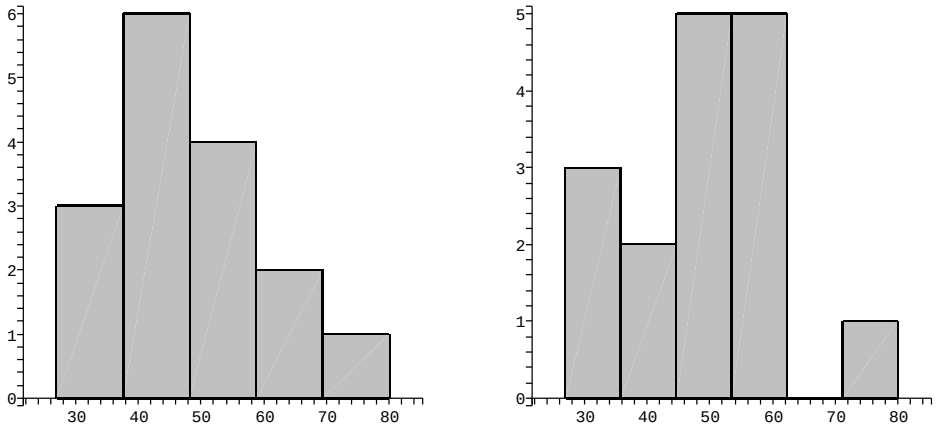
Bij een *histogram* worden de klassen door balken vertegenwoordigd, waarbij de *oppervlakte* van de balken de frequenties representeert. Als de balken ook dezelfde breedte hebben, zijn natuurlijk ook de hoogtes van de balken proportioneel met de frequenties. In Figuur 1 zijn twee histograms voor ons voorbeeld te zien: In het linkerplaatje zijn de klassen intervallen van breedte 10, in het rechterplaatje zijn de klassen automatisch zo gekozen dat elke klasse even veel (in dit geval 4) punten bevat, en de balken dezelfde oppervlakte hebben.



Figuur 1: Histograms met balken van dezelfde en verschillende breedtes.

Als we in ons voorbeeld het aantal klassen volgens de formule $1 +^2 \log(n)$ zouden kiezen, hadden we 5 klassen nodig. De histograms in Figuur 2 laten zien dat een opsplitsing in 5 of 6 klassen een duidelijk kwalitatief verschil in de histograms veroorzaakt: In het eerste geval is er een duidelijk grootste klasse, in het tweede geval zijn er twee grootste klassen en men kan zien dat er een uitschieter is, omdat er een gat tussen de klasse met de maximale waarde en de andere klassen valt.

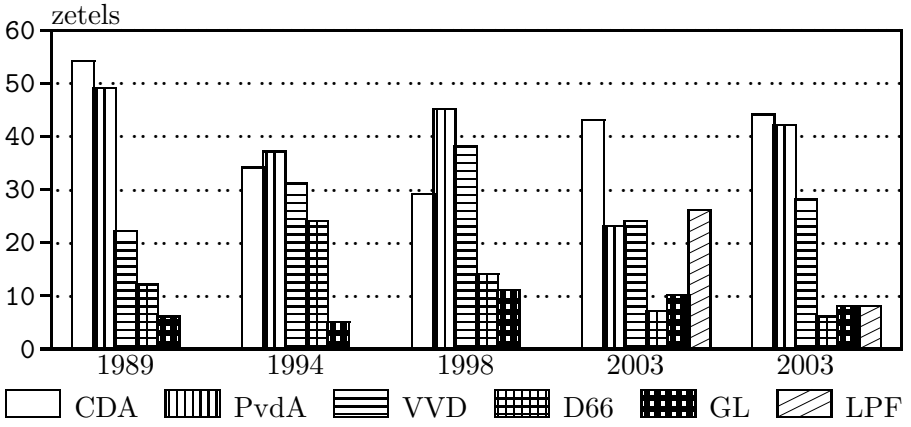
Er kunnen ook histograms van meerdere verzamelingen gegevens in een grafiek gecombineerd worden. Dit wordt vaak gebruikt om de ontwikkeling over de tijd te laten zien. De volgende tabel geeft het aantal zetels in de Tweede Kamer weer voor de verkiezingen sinds 1989 (beperkt tot partijen die in een van de verkiezingen minstens 10 zetels heeft gehaald).



Figuur 2: Histograms met 5 en 6 klassen.

Partij	1989	1994	1998	2002	2003
CDA	54	34	29	43	44
PvdA	49	37	45	23	42
VVD	22	31	38	24	28
D66	12	24	14	7	6
GroenLinks	6	5	11	10	8
LPF	0	0	0	26	8

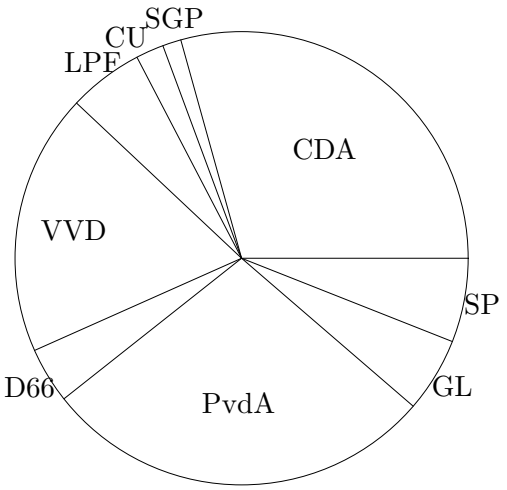
Als we voor elke partij een histogram voor het aantal zetels in de verschillende verkiezingen maken, ziet de combinatie van deze histograms uit als in Figuur 3 te zien. Natuurlijk kan men ook de verdelingen van zetels in een verkiezing als histogram zien, dan worden in deze grafiek gewoon verschillende histograms naast elkaar gezet.



Figuur 3: Verdeling van zetels in de Tweede Kamer.

Taart-diagram

Bij een *taart-diagram* (*pie chart*) wordt een cirkelschijf zo onderverdeeld dat de oppervlaktes van de sectoren de frequenties van de klassen representeren. Omdat de oppervlakte van een sector evenredig is met de hoek van de sector, geven ook de hoeken van de sectoren de frequenties weer. Voor de verkiezingen van 2003 is dit in Figuur 4 te zien.



Figuur 4: Taart-diagram voor de verdeling van zetels in de Tweede Kamer.

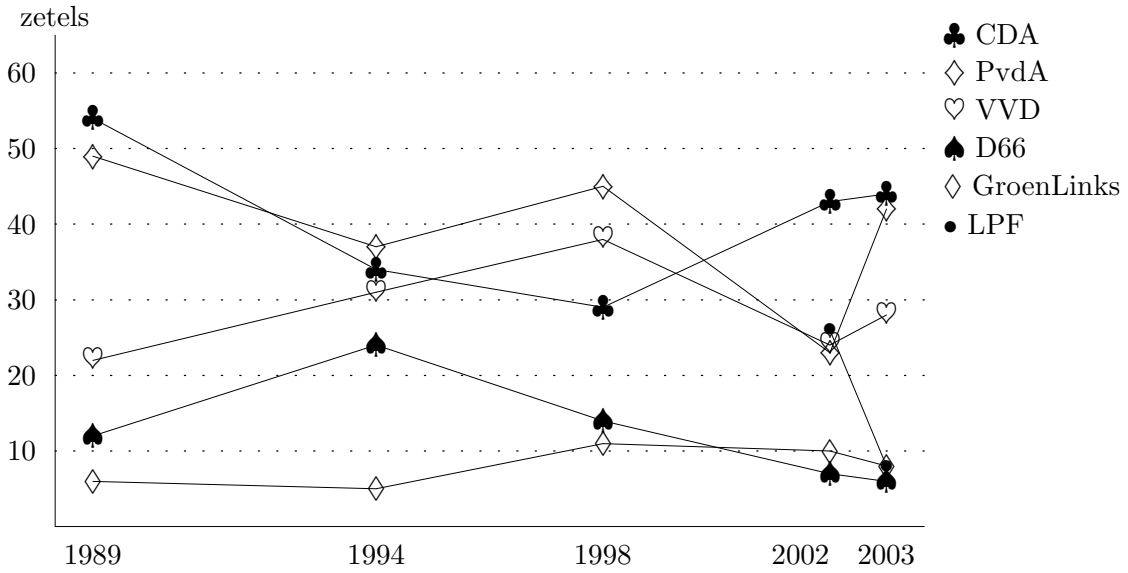
Frequentiepolygoon

In plaats van verschillende histograms in een grafiek te combineren, kan men ook de waarden van verschillende verdelingen over de tijd door *frequentiepolygonen* aangeven. Hierbij worden de waarden voor verschillende tijdstippen (bijvoorbeeld) door lijnstukken verbonden. Merk op dat de tussenwaarden meestal geen betekenis hebben. Ook al kun je op een lijnstuk tussen de verkiezingen van 1994 en 1998 een waarde voor het jaar 1996 aflezen, zegt dat niets over een mogelijke uitslag van verkiezingen in het jaar 1996. De ontwikkeling van het aantal zetels in de Tweede Kamer die in Figuur 3 door een combinatie van histograms beschreven werd, wordt in Figuur 5 door *frequentiepolygonen* gerepresenteerd.

Vervalsende representatie

Het kiezen van een vorm van representatie houdt altijd een manipulatie van de gegevens in. Dit hoeft niet per se negatief te zijn, want *een plaatje zegt meer dan duizend woorden*. Maar door een specifieke keuze van representatie kan er wel een zekere tendentie aan de gegevens gegeven worden. Dit leidt soms - bewust of onbewust - tot een vervalsing van de gegevens. Een paar bekende vervalsingen zijn:

- Schaling van de assen. Hierdoor wordt het stijgen of dalen stijler of vlakker en de veranderingen worden versterkt of verzwakt weergegeven.



Figuur 5: Frequentiepolygonen voor de verdeling van zetels in de Tweede Kamer.

- Afbreken van de y-as boven het nulpunt. Hierdoor lijken veranderingen veel extremer dan ze in werkelijkheid zijn.
- 'Slimme' keuze van klassen. Hierdoor kunnen effecten kunstmatig voortgebracht of onderdrukt worden.
- Representeren van de frequentie door een figuur waarvan de hoogte proportioneel met de frequentie is. Omdat de oppervlakte en niet de hoogte van de figuur waargenomen wordt, lijkt een twee keer zo hoge figuur vier keer zo groot.
- Suggereren van een *ontwikkeling* door representatie middels frequentiepolygonen.

1.3 Typische waarden

Om verschillende verzamelingen van gegevens te kunnen vergelijken, is het vaak handig om een *typische waarde* voor een verzameling aan te geven. Er zijn verschillende manieren, om zo'n typische waarde te definiëren, en er is geen *juiste* manier.

Het gemiddelde

Het *rekenkundig gemiddelde* (meestal kort *gemiddelde* genoemd) van waarden $x_1, x_2, \ldots, x_n$ is gedefinieerd door

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

De interpretatie hiervan is dat de gegevens bij elkaar opgeteld worden en vervolgens de som gelijkvormig over de individuën verdeeld wordt.

Een karakterisering van het gemiddelde is de eigenschap dat de verschillen tussen de gegevens en het gemiddelde bij elkaar opgeteld 0 geven, dus dat $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. Maar de belangrijkste eigenschap van het gemiddelde $\overline{x}$ is, dat het juist de waarde x is waarvoor de som van de kwadratische afstanden van de x_i minimaal wordt, dus waarvoor de functie

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

minimaal wordt. Dit wordt vaak zelfs als definitie van de gemiddelde gebruikt. Een minimum van $f(x)$ vinden we als nulpunt van de afgeleide $f'(x)$. Er geldt $f'(x) = \sum_{i=1}^{n}(2x - 2x_i)$ en dus $f'(x) = 0$ voor $n \cdot x = \sum_{i=1}^{n} x = \sum_{i=1}^{n} x_i$. Omdat de functie $f(x)$ een naar boven geopende parabool is, is dus $x = \overline{x}$ het eenduidige minimum van de functie.

We kunnen het gemiddelde ook in samenhang met kansverdelingen interpreteren. Als we ons voorstellen dat de x_i waarden van een stochast X zijn, die met kans p_x het resultaat x oplevert, dan zullen we de waarde x in een verzameling van n waarden ongeveer $p_x \cdot n$ keer verwachten. Maar als we nu bij het gemiddelde $\overline{x}$ niet meer de som over de x_i maar over de waarden x met hun frequenties nemen, zien we dat $\overline{x}$ een benadering van de verwachtingswaarde $E[X] = \sum_x x \cdot p_x$ van de stochast X is.

Met een analoog argument zien we voor een stochast X met continue kansverdeling met dichtheidsfunctie $f(x)$ dat $\overline{x}$ ook hier een benadering van de verwachtingswaarde $E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$ is.

De mediaan

De *mediaan* $\tilde{x}$ van een verzameling gegevens is gedefinieerd als de waarde die in het midden van de geordende waarden ligt. Dit wil zeggen dat er even veel waarden kleiner dan $\tilde{x}$ zijn als er waarden groter zijn. Als we aannemen, dat de waarden opstijgend geordend zijn, dus $x_1 \leq x_2 \leq \ldots \leq x_n$, dan is voor oneven $n = 2m + 1$ de mediaan $\tilde{x}$ juist de middelste waarde x_m. Voor een even aantal $n = 2m$ neemt men gewoon het gemiddelde van de twee middelste waarden, dus $\tilde{x} = \frac{1}{2}(x_m + x_{m+1})$. Voor opstijgende waarden $x_1 \leq x_2 \leq \ldots \leq x_n$ hebben we dus:

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

We hebben gezien dat de som van de verschillen tussen de waarden en het gemiddelde $\overline{x}$ nul geeft en dat het gemiddelde de kwadratische afstanden minimaliseert. De mediaan heeft de eigenschap dat hij de gewone afstanden minimaliseert, dus dat $\tilde{x}$ de waarde is waarvoor

$$g(x) := \sum_{i=1}^{n} |x_i - x|$$

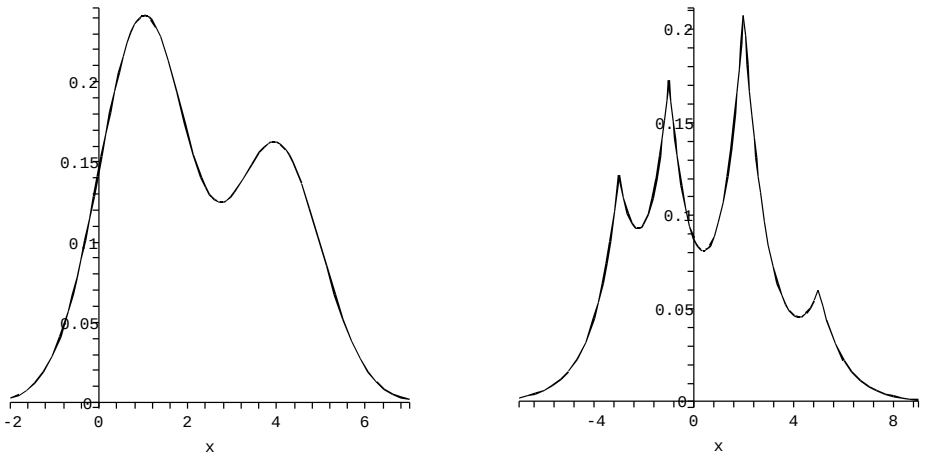
minimaal wordt. Dit ziet men (voor oneven n) als volgt in: Stel we hebben $x > \tilde{x}$, dan liggen er r waarden rechts van x en l waarden links van x en we hebben $l > r$. Als we nu x om Δx naar rechts schuiven, dan neemt $g(x)$ om $\Delta x(l - r)$ toe, als we x om Δx naar links schuiven, neemt $g(x)$ om $\Delta x(l - r)$ af. Dus is $g(x)$ niet minimaal als $l > r$ is. Met hetzelfde argument, toegepast of $x < \tilde{x}$, zien we dat $g(x)$ ook voor $l < r$ niet minimaal is. Dus moet $l = r$ gelden, en hieruit volgt $x = \tilde{x}$.

Voor even $n = 2m$ is $g(x)$ op het interval $[x_m, x_{m+1}]$ horizontaal met minimale waarde. Men neemt daarom het middelpunt van dit interval als mediaan.

De modus

Een verdere mogelijkheid om een typische waarde te definiëren is de *modus* $\hat{x}$ die de waarde met de hoogste frequentie is.

In veel gevallen geeft de modus een goede beschrijving die ook redelijk dicht bij het gemiddelde en de mediaan ligt, maar dit hangt sterk van de situatie af. Het kan bijvoorbeeld zijn, dat een verdeling twee duidelijke spitsen heeft, dan is de modus de hogere van de twee spitsen, maar gemiddelde en mediaan liggen waarschijnlijk tussen de spitsen. Een verdeling met twee spitsen heet *bimodaal*, een verdeling met nog meer spitsen *multimodaal*.



Figuur 6: Bimodale en multimodale verdelingen.

Het linkerplaatje in Figuur 6 laat een bimodale verdeling zien. De modus van deze verdeling is $\hat{x} = 1$, de mediaan is $\tilde{x} \approx 1.92$ en het gemiddelde is $\overline{x} = 2.2$.

In het rechterplaatje van Figuur 6 vinden we een multimodale verdeling met vier spitsen. In dit geval is de modus ook weer $\hat{x} = 1$, de mediaan is $\tilde{x} \approx 0.39$ en het gemiddelde is $\overline{x} = 0.4$.

Soms kan bij een multimodale verdeling de modus juist wel interessant zijn, maar vaak is het in dit geval nodig de verdeling als combinatie van een aantal unimodale verdelingen te beschrijven en door de typische waarden van deze verdelingen te karakteriseren.

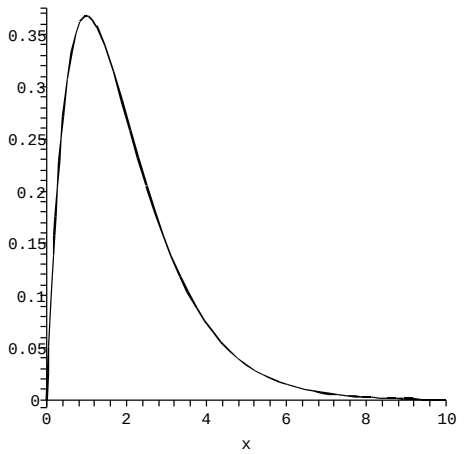
Samenhang tussen gemiddelde, mediaan en modus

Als een verzameling van gegevens een symmetrische unimodale verdeling heeft, vallen de waarden van het gemiddelde, de mediaan en de modus redelijk goed samen. Als de verdeling niet symmetrisch is en een langere staart naar rechts heeft, noemt men de verdeling *naar rechts scheef*. In dit geval is $\hat{x} < \tilde{x} < \overline{x}$. Omgekeerd heet een verdeling *naar links scheef* als hij een langere staart naar links heeft. In dit geval geldt $\overline{x} < \tilde{x} < \hat{x}$.

Een typische naar rechts scheve verdeling is

$$f(x) = \lambda^2 x e^{-\lambda x} \text{ met } \overline{x} = \frac{2}{\lambda}, \quad \tilde{x} \approx 1.678 \cdot \frac{1}{\lambda}, \quad \hat{x} = \frac{1}{\lambda}.$$

Deze verdeling is in Figuur 7 voor de parameter $\lambda = 1$ te zien. In het plaatje ligt dus de modus bij $\hat{x} = 1$, de mediaan bij $\tilde{x} \approx 1.678$ en het gemiddelde bij $\overline{x} = 2$.



Figuur 7: Naar rechts scheve verdeling $f(x) = xe^{-x}$.

Omdat de modus vaak niet eenvoudig te berekenen valt, wordt er voor unimodale verdelingen soms een heuristische formule voor de samenhang tussen modus, mediaan en gemiddelde toegepast, namelijk

$$\overline{x} - \hat{x} = 3(\overline{x} - \tilde{x}).$$

Voor de boven aangegeven verdeling $f(x) = \lambda^2 x e^{-\lambda x}$ zien we dat dit een uitstekende vuistregel is, maar let wel dat dit bij multimodale verdelingen meestal vreselijk mis gaat (zie de voorbeelden in Figuur 6).

Merk op: Het gemiddelde is veel gevoeliger voor *uitschieters* dan de mediaan. Op de modus heeft een uitschieter helemaal geen invloed. Als het erom gaat een robuuste schatting voor de typische waarde te hebben en er gevaar op uitschieters bestaat, is de mediaan soms een betere keuze dan het gemiddelde.

In ons voorbeeld van de tentamen resultaten kunnen we het gemiddelde en de mediaan makkelijk bepalen, we hebben $\overline{x} = 48.56$ en $\tilde{x} = 47.5$. Voor

de modus moeten we naar klassen kijken, als we bijvoorbeeld als klassen de intervallen van breedte 10 nemen, ligt de modus in het interval $[40, 50]$ en men neemt hiervoor de middelste waarde van het interval, dus $\hat{x} = 45$. Als we nu de uitslag van 80 punten als uitschieter beschouwen en weglaten, verandert dit het gemiddelde behoorlijk, we krijgen dan als nieuwe gemiddelde $\overline{x} = 46.47$, terwijl de mediaan veel minder verandert en nu $\tilde{x} = 47$ wordt. De modus blijft onveranderd.

We kunnen zelfs algemeen aangeven hoeveel het weglaten van een waarde het gemiddelde verandert. Stel we hebben bij n waarden en gemiddelde $\overline{x}$ en willen de waarde x weglaten. Het nieuwe gemiddelde wordt dan $\frac{n \cdot \overline{x} - x}{n-1}$ en voor het verschil van het oude en het nieuwe gemiddelde krijgen we:

$$\overline{x} - \frac{n \cdot \overline{x} - x}{n-1} = \frac{(n-1) \cdot \overline{x} - n \cdot \overline{x} + x}{n-1} = \frac{x - \overline{x}}{n-1}.$$

Het gemiddelde verandert dus om de afstand van de uitschieter van het gemiddelde, gedeeld door $n - 1$.

Andere gemiddelden

Soms is het rekenkundig gemiddelde niet geschikt om een gelijkmatige herverdeling te beschrijven. Dit is bijvoorbeeld het geval als de gegevens x_i een variabel beschrijven die niet opgeteld maar vermenigvuldigd wordt, zoals bij groeiprocessen. Stel een populatie groeit in n jaren met factoren $x_1, x_2, \ldots, x_n$, dan is de totale groei het product $\prod_{i=1}^{n} x_i$ van de x_i. Om nu een gemiddelde groei te berekenen, waarmee in n jaren dezelfde totale groei bereikt wordt, moeten we een waarde x_0 vinden zo dat $x_0^n = \prod_{i=1}^{n} x_i$. We moeten dus uit het product de n-de wortel trekken, dit geeft

$$x_0 = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

en x_0 heet het *meetkundig gemiddelde* van de x_i.

Een andere vorm van gemiddelde bestaat bij gegevens waarvoor eigenlijk x_i^{-1} opgeteld moet worden. Een beroemd voorbeeld hiervoor is het probleem van de piloot die op de heenweg wind tegen heeft maar de vertraging op de terugweg door de wind mee weer in te halen denkt. We noemen de afstand van de twee vliegvelden s, de tijd voor de heenweg t_1 en de tijd voor de terugweg t_2. Als de piloot zonder wind met een snelheid van v_0 vliegt, zou hij zonder wind de tijd $t = 2\frac{s}{v}$ nodig hebben. Bij wind met snelheid w is de snelheid op de heenweg $v_1 = v_0 - w$ en op de terugweg $v_2 = v_0 + w$. De tijden voor heen- en terugweg zijn $t_1 = \frac{s}{v_1}$ en $t_2 = \frac{s}{v_2}$. De vraag is nu, of $t_1 + t_2$ gelijk aan t is. Voor de gemiddelde snelheid $v = \frac{2s}{t_1+t_2}$ geldt:

$$v = \frac{2s}{t_1 + t_2} = \frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}} = \frac{2v_1v_2}{v_1 + v_2} \text{ en dus } \frac{1}{v} = \frac{\frac{1}{v_1} + \frac{1}{v_2}}{2}.$$

Men noemt $v = \frac{2v_1v_2}{v_1+v_2}$ het *harmonisch gemiddelde* van v_1 en v_2 en dit is gewoon het inverse van het rekenkundig gemiddelde van de inversen van v_1 en v_2. In het geval met $v_1 = v_0 - w$ en $v_2 = v_0 + w$ hebben we

$$v = \frac{2(v_0 - w)(v_0 + w)}{(v_0 - w) + (v_0 + w)} = \frac{2(v_0^2 - w^2)}{2v_0} = \frac{v_0^2 - w^2}{v_0} < v_0.$$

De vliegreis duurt dus inderdaad langer.

Tussen de verschillende gemiddelden bestaat altijd de volgende keten van ongelijkheden:

$$\text{minimum} \leq \text{harmonisch} \leq \text{meetkundig} \leq \text{rekenkundig} \leq \text{maximum}\ .$$

1.4 Spreiding

Het is duidelijk dat een verzameling gegevens met een gemiddelde waarde (of zelfs de verschillende soorten van gemiddelden) nog niet goed beschreven is, want de verdelingen kunnen er nog erg verschillend uit zien. Bijvoorbeeld kan het zijn dat bij een tentamen met een gemiddelde van 7 iedereen het gehaald heeft, omdat er even veel 6en als 8en en geen 9en en 10en waren. Maar het kan ook zijn, dat slechts 40% het gehaald hebben, omdat 40% een 10 en 60% en 5 gehaald hebben (dit is een typisch voorbeeld van een bimodale verdeling). Men wil daarom ook een uitspraak over de afwijking van de waarden van het gemiddelde hebben. Ook hiervoor zijn er verschillende mogelijkheden.

Standaardafwijking

We hebben al gezien dat het gemiddelde $\overline{x}$ de waarde is waarvoor de kwadratische afstanden van de gegevens minimaal is. De wortel uit dit minimum heet de *standaardafwijking* s, we hebben dus

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Voor veel (en belangrijke) verdelingen ligt een 'groot deel' van de waarden binnen een afstand van s van het gemiddelde. Voor de normaalverdeling zijn dit bijvoorbeeld 68% (en 95% liggen binnen een afstand van $2s$). Met behulp van het gemiddelde en de standaardafwijking laten zich gegevens normaliseren: De verschuiving $x_i' := x_i - \overline{x}$ geeft een verzameling gegevens met gemiddelde 0 en $z_i := \frac{x_i - \overline{x}}{s}$ geeft een verzameling gegevens met standaardafwijking 1. Men noemt de waarde

$$z := \frac{x_i - \overline{x}}{s}$$

ook de z-waarde van x_i. De z-waarde geeft de afwijking van een waarde van het gemiddelde in veelvouden van de standaardafwijking aan. Men zegt daarom ook soms dat een waarde *een afstand van 3 standaardafwijkingen* heeft, als de z-waarde 3 is.

Als we de standaardafwijking weer voor waarden bekijken die volgens een kansverdeling voor een stochast X geproduceerd zijn, zien we dat s^2 een benadering van de variantie $Var(X) = E[(X - E[X])^2]$ is. Voor een discrete kansverdeling is deze gegeven door $Var(X) = \sum_x (x - E[X])^2 \cdot p_x$, en voor een continue kansverdeling met dichtheidsfunctie $f(x)$ door $Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f(x)\, dx$.

In de kansrekening hebben we de wortel uit de variantie ook de *standaardafwijking* genoemd en toen met σ genoteerd. Het is inderdaad gebruikelijk, grootheden van kansverdelingen zo als verwachtingswaarde en standaardafwijking met *griekse letters* (μ, σ) te noteren, terwijl grootheden bij verdelingen van gegevens met *latijnse letters* genoteerd worden. Let wel dat niet iedere auteur dit soort conventies behartigt.

Kwartielen

Net als de mediaan voor de helft van de waarden worden ook *kwartielen* gedefinieerd waar een kwart van de waarden beneden of boven ligt. Het *onderste kwartiel* of *eerste kwartiel* is de waarde waar een kwart van de waarden onder en drie kwart boven liggen en is dus de mediaan van de onderste helft van de waarden. Net zo is het *bovenste kwartiel* of *derde kwartiel* de waarde waar drie kwart onder en een kwart boven ligt, dus de mediaan van de bovenste helft van de waarden. De mediaan zelfs heet soms ook het *tweede kwartiel*.

Algemeen noemt men de waarde waar p procent van de waarden onder en $100 - p$ procent boven liggen het *p-percentielpunt* en noteert dit met P_p. De mediaan is dus het 50-percentielpunt P_{50}, het onderste kwartiel het 25-percentielpunt P_{25} en het bovenste kwartiel het 75-percentielpunt P_{75}. Meestal zal een p-percentielpunt niet precies op een waarde vallen, en ook niet op het middelpunt tussen twee waarden. Bij n (geordende) waarden heeft het p-percentielpunt in de lijst de index $t = 1 + \frac{p}{100}(n-1)$. Als we t schrijven als $i + r$ met i een natuurlijk getal en $0 \leq r < 1$, dan berekenen we de waarde voor het p-percentielpunt als gewogen gemiddelde van x_i en x_{i+1} met gewichten $(1 - r)$ en r, dus als

$$P_p = (1 - r) \cdot x_i + r \cdot x_{i+1}.$$

Als we in ons voorbeeld van 16 waarden het 15-percentielpunt zouden willen vinden, hebben we $t = 1 + \frac{15}{100} \cdot 15 = 1 + \frac{225}{100} = 3 + \frac{1}{4}$. Het 15-percentielpunt ligt dus tussen x_3 en x_4, maar op een vierde van de afstand van x_3 naar x_4. We zouden dus in dit geval het 15-percentielpunt berekenen door $0.75 \cdot x_3 + 0.25 \cdot x_4$.

Percentielpunten worden ook gebruikt om parameters van systemen vast te leggen. Bijvoorbeeld geeft een spraakherkenningssysteem voor elke herkenning een *score* die aangeeft hoe goed de kwaliteit van de herkenning was. Dit geeft in het algemeen niet de kans op een correcte herkenning weer, maar slechts een heuristische waarde die met toenemende kwaliteit stijgt. Als men met het automatische systeem nu 90% van de aanvragen wil behandelen en de rest naar een menselijke operator doorstuurt, dan moet men op een testset van aanvragen

het 90-percentielpunt van de scores bepalen en dit als grens vastleggen waaronder aanvragen naar de operator doorgestuurd worden.

De afstand tussen de kwartielen geeft informatie over de spreiding van de waarden. Het interval tussen de kwartielen P_{25} en P_{75} heet het *interkwartielbereik*, hun verschil de *interkwartielafstand* IQR (voor *inter quartile range*). Vaak wordt ook de helft van de interkwartielafstand gebruikt, de *semi-interkwartielafstand* $\frac{1}{2}IQR := \frac{P_{75}-P_{25}}{2}$.

Er is geen zuivere definitie mogelijk wanneer een waarde die uit het algemene patroon valt een uitschieter is, over dit probleem zijn veel boeken geschreven. Een veel gehanteerde vuistregel is, waarden als uitschieters te beschouwen die meer dan $1.5 \cdot IQR$ buiten het interkwartielbereik liggen, dus:

$$x < P_{25} - 1.5 \cdot IQR \text{ of } x > P_{75} + 1.5 \cdot IQR \Rightarrow x \text{ is een uitschieter.}$$

Voor waarden die volgens dit criterium uitschieters zijn moet men *met de hand* beslissen of het gewoon extreme maar geldige waarden zijn of ongeldige waarden die uit het bestand verwijderd moeten worden (bijvoorbeeld omdat er bij een meeting iets is mis gegaan).

Voor verdelingen die niet erg scheef zijn, bestaat er een verband tussen de standaardafwijking s en die semi-interkwartielafstand $\frac{1}{2}IQR$, namelijk

$$\frac{1}{2}IQR \approx \frac{2}{3}s.$$

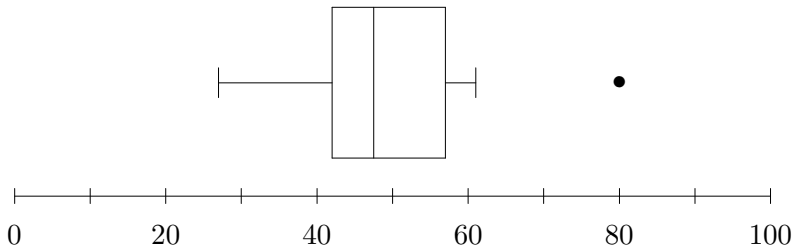
Dit is afgeleid van de normaalverdeling, waarvoor $\frac{1}{2}IQR \approx 0.6745$ geldt.

Natuurlijk leveren naast de kwartielen ook de minimale en de maximale waarde informatie over de spreiding van een verdeling. Dit soort informatie wordt vaak in een *doos-en-snorren* figuur (*box-and-whiskers plot* of kort *box-plot*) samengevat. Dit is een doos tussen de kwartielen met de mediaan gemarkeerd. Voor de einden van de snorren zijn er verschillende conventies:

- minimale en maximale waarden;
- minimale en maximale waarden die binnen een afstand van $1.5 \cdot IQR$ van de kwartielen liggen, de andere waarden worden als uitschieters beschouwd (en soms wel als punten weergegeven);
- 5-percentielpunt en 95-percentielpunt.

In ons voorbeeld van de tentamenresultaten hebben we $P_{50} = 47.5$, $P_{25} = 42$ en $P_{75} = 57$. Hieruit volgt $IQR = 15$. Omdat $42 - 1.5 \cdot 15 = 19.5$ kleiner is dan alle waarden, hebben volgens het genoemde criterium geen uitschieters naar beneden. Aan de andere kant is $57 + 1.5 \cdot 15 = 79.5$, dus is de waarde 80 net een uitschieter.

De doos-en-snorren figuur voor het voorbeeld ziet er dus als volgt uit:



De doos-en-snorren figuur wordt soms horizontaal (zo als hier) en soms verticaal getekend. De verticale versie heeft het voordeel dat de figuren voor verschillende verdelingen makkelijk naast elkaar geplaatst kunnen worden.

1.5 Momenten

We hebben al een paar keer iets over de scheefheid van een verdeling gezegd. Natuurlijk laat zich dit aan de hand van een grafiek meestal goed aflezen, maar het is handig hiervoor ook een kwantitatief begrip te hebben. Hiervoor zijn de *momenten* van een verdeling handig. Het k-de *moment* van een verzameling gegevens is

$$m'_k := \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

en het k-de *centrale moment* rond het gemiddelde is

$$m_k := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k.$$

De eerste en tweede momenten zijn oude bekenden, we hebben $\overline{x} = m'_1$, $m_1 = 0$ en $s = \sqrt{m_2}$.

Om momenten voor verschillende verdelingen goed te kunnen vergelijken, is het handig om ze te normaliseren. Dit gebeurt net als bij de z-waarde door delen door de standaardafwijking en men krijgt

$$a_k := \frac{m_k}{s^k} = \frac{m_k}{\sqrt{m_2}^k}.$$

Momenten worden op een analoge manier ook voor kansverdelingen gedefinieerd. Voor een stochast X met een discrete kansverdeling met kansen p_x zijn de k-de momenten μ'_k en de k-de centrale momenten μ_k gedefinieerd door

$$\mu'_k := \sum_x x^k \cdot p_x \text{ en } \mu_k := \sum_x (x - E[X])^k \cdot p_x.$$

Voor een stochast X met een continue kansverdeling met dichtheidsfunctie $f(x)$ geldt

$$\mu'_k := \int_{-\infty}^{\infty} x^k \cdot f(x)\, dx \text{ en } \mu_k := \int_{-\infty}^{\infty} (x - E[X])^k \cdot f(x)\, dx.$$

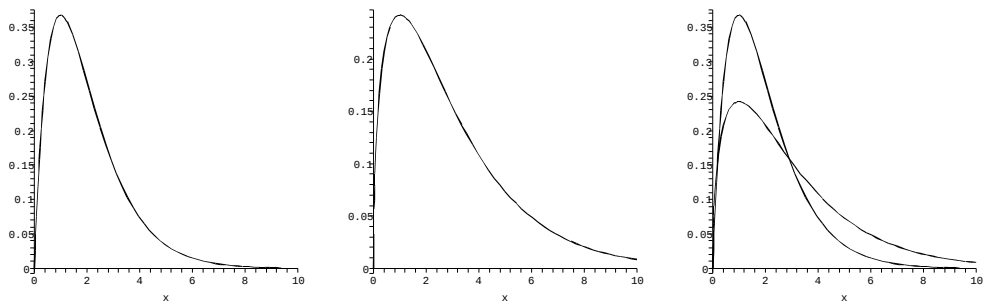
In het bijzonder is $\mu'_1 = E[X]$ en $\mu_2 = Var(X)$.

Merk op dat hogere momenten niet voor alle verdelingsfuncties van continue kansverdelingen hoeven te bestaan. Zo heeft bijvoorbeeld de integraal $\int_{-\infty}^{\infty} \frac{1}{1+x^2}\, dx$ de waarde π, maar de integralen $\int_{-\infty}^{\infty} x^2 \cdot \frac{1}{1+x^2}\, dx$ en $\int_{-\infty}^{\infty} x^4 \cdot \frac{1}{1+x^2}\, dx$ hebben geen eindige waarde.

Scheefheid

Omdat voor een scheve verdeling de waarden in de langere staart een hoger gewicht krijgen, is het derde centrale moment een maat voor de scheefheid (*skewness*) van de verdeling. Bij positieve waarden van m_3 of a_3 is de verdeling scheef naar rechts, bij negatieve waarden scheef naar links. Men noemt a_3 ook de *coëfficiënt van scheefheid*. Verdelingen die symmetrisch ten opzichte van hun gemiddelde zijn (zo als de normaalverdeling), hebben natuurlijk scheefheid 0.

In Figuur 8 zijn de grafieken van twee naar rechts scheve verdelingen te zien. De functie in het linkerplaatje is $f(x) := \lambda^2 x \cdot e^{-\lambda x}$ (voor $\lambda = 1$), de functie in het middelste plaatje is $g(x) := \frac{1}{\sqrt{2\pi}} \sqrt{x} \cdot e^{-\frac{x}{2}}$.



Figuur 8: Verdelingsfuncties van twee haar rechts scheve verdelingen.

De momenten voor $f(x)$ zijn $\overline{x} = m_1' = \frac{2}{\lambda}$, $s^2 = m_2 = \frac{2}{\lambda^2}$ en $m_3 = \frac{4}{\lambda^3}$. Hieruit volgt dat de coëfficiënt van scheefheid $a_3 = \frac{m_3}{\sqrt{m_2}^3} = \sqrt{2} \approx 1.414$ is. Merk op dat a_3 onafhankelijk van de parameter λ is.

De momenten voor $g(x)$ zijn $\overline{x} = m_1' = 3$, $s^2 = m_2 = 6$ en $m_3 = 24$. Hieruit volgt dat $g(x)$ de coëfficiënt van scheefheid $a_3 = \frac{m_3}{\sqrt{m_2}^3} = \frac{2}{3}\sqrt{6} \approx 1.633$ heeft. Zo als ook uit de plaatjes blijkt, heeft $g(x)$ een grotere scheefheid dan $f(x)$.

Een andere mogelijkheid om de scheefheid aan te geven gebruiken het verschil van gemiddelde en modus, bijvoorbeeld $\frac{\overline{x}-\hat{x}}{s}$. Als we hierin de heuristische benadering $\overline{x} - \hat{x} = (\overline{x} - \tilde{x})$ voor de modus toepassen, krijgen we $\frac{3(\overline{x}-\tilde{x})}{s}$ als uitdrukking voor de scheefheid.

Ook met behulp van de kwartielen of percentielen laat zich de scheefheid uitdrukken, bijvoorbeeld door

$$\frac{(P_{75} - \tilde{x}) - (\tilde{x} - P_{25})}{P_{75} - P_{25}} = \frac{P_{75} - 2\tilde{x} + P_{25}}{P_{75} - P_{25}} \quad \text{of}$$

$$\frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}.$$

Hierbij wordt gekeken hoe ver de p-percentielpunten P_{50-x} en P_{50+x} die bij een symmetrische verdeling even grote afstanden van de mediaan hebben van een symmetrische positie afwijken.

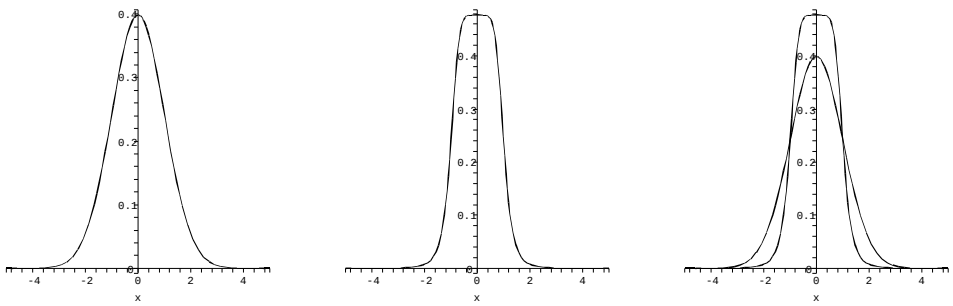
Scherptoppigheid

Het vierde moment zegt iets erover of een verdeling spits of plat is, dus over de *scherptoppigheid* of *gepiekdheid* (*kurtosis*) van de verdeling. Hiervoor vergelijkt men het genormaliseerde vierde moment a_4 met het vierde moment van de standaardnormaalverdeling dat de waarde 3 heeft en noemt a_4 ook de *coëfficiënt van scherptoppigheid*. Voor $a_4 > 3$ noemt men een verdeling *gepiekd* (*leptokurtic*, van het griekse *lepto-* = smal) omdat de verdeling dan een scherpere top heeft dan de normaalverdeling en de staarten dunner zijn. Voor $a_4 < 3$ noemt men de verdeling *afgeplat* (*platykurtic*, van *platy-* = plat) omdat ze een plattere top heeft dan de normaalverdeling. Een verdeling met $a_4 \approx 3$ heet *mesokurtic* (van *meso-* = gemiddeld).

Merk op: In de literatuur wordt vaak ook $a_4 - 3$ als coëfficiënt van scherptoppigheid gehanteerd, een positieve waarde hiervan staat dan voor een gepiekde verdeling, een negatieve waarde voor een afgeplatte verdeling.

Als eenvoudig voorbeeld bekijken we de symmetrische uniforme verdeling op het interval $[-c, c]$, deze heeft de dichtheidsfunctie $f(x) = \frac{1}{2c}$. Er geldt $m_2 = \int_{-c}^{c} x^2 \cdot \frac{1}{2c}\, dx = \frac{1}{2c} \cdot \frac{x^3}{3} \mid_{-c}^{c} = \frac{1}{3}c^2$ en $m_4 = \int_{-c}^{c} x^4 \cdot \frac{1}{2c}\, dx = \frac{1}{2c} \cdot \frac{x^5}{5} \mid_{-c}^{c} = \frac{1}{5}c^4$. Hieruit volgt $a_4 = \frac{m_4}{m_2^2} = \frac{9}{5} < 3$, dus is de uniforme verdeling afgeplat. Merk op dat de schalingsfactor c geen invloed op de scherptoppigheid van de verdeling heeft.

Een interessanter voorbeeld is de verdeling met dichtheidsfunctie $f(x) = \frac{3}{2\pi} \cdot \frac{1}{1+x^6}$ die in het middelste plaatje van Figuur 9 te zien is. Hier hebben we $m_2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \frac{1}{2}$ en $m_4 = \int_{-\infty}^{\infty} x^4 \cdot f(x)\, dx = 1$, dus is $a_4 = \frac{m_4}{m_2^2} = 4$ en $f(x)$ is een gepiekde verdeling. Dit wordt ook in het vergelijk met de normaalverdeling in Figuur 9 duidelijk.



Figuur 9: Verdelingsfuncties voor de normaalverdeling en een gepiekde verdeling.

Merk op dat de scherptoppigheid vooral bij (redelijk) symmetrische verdelingen een rol speelt. Bij scheve verdelingen heeft de scheefheid een groot invloed

op de coëfficiënt van scherptoppigheid en is het vergelijken met symmetrische verdelingen meestal niet bijzonder verklarend.

Belangrijke begrippen in deze les

- stengel-en-blad diagram
- klassen, frequentieverdeling
- histogram, taart-diagram
- gemiddelde, mediaan, modus
- uni-, bi-, multimodale verdelingen
- kwartielen, p-percentielpunten
- standaardafwijking, interkwartielafstand
- doos-en-snorren figuur
- momenten, scheefheid, scherptoppigheid

Opgaven

1. Gegeven is de rij waarnemingen

$$15.813, 15.705, 15.748, 15.801, 15.720, 15.743.$$

Bereken het gemiddelde en de standaardafwijking van deze gegevens

(i) zonder af te ronden;

(ii) met op twee decimalen achter de komma afgeronde waarden;

(ii) met op een decimaal achter de komma afgeronde waarden.

2. Dit is een standaardafwijkings-wedstrijd: Kies als gegevens 4 getallen uit de getallen $0, 1, \ldots, 10$, waarbij herhalingen toegestaan zijn.

(i) Vind getallen zo dat hun standaardafwijking minimaal is. Is het antwoord eenduidig?

(ii) Vind getallen zo dat hun standaardafwijking maximaal is. Is het antwoord eenduidig?

(iii) Behandel (i) en (ii) met 3 in plaats van 4 getallen.

3. Zij X het aantal ogen dat geworpen wordt met twee witte en één zwarte dobbelsteen, waarbij het aantal ogen van de zwarte dobbelsteen dubbel wordt geteld. In een experiment met 50 werpen zijn de volgende resultaten verkregen:

12	10	23	10	10	14	15	20	5	18
14	8	6	20	21	12	16	11	13	21
13	10	9	16	19	7	9	7	20	22
17	14	15	15	12	9	13	14	18	8
17	18	15	12	14	20	18	11	19	7

(i) Bereken de verwachtingswaarde $E[X]$ en de variantie $Var(X)$ van de stochast X (dit hangt niet van de verkregen resultaten af).

(ii) Bereken het gemiddelde $\overline{x}$ en de standaardafwijking s van de 50 waarnemingen.

(iii) Maak een histogram voor een zinvolle indeling van de waarnemingen in klassen.

4. De aantallen van stemmen voor de kandidaat presidenten in de VS in de verkiezingen sinds 1960 (dus sinds Kennedy) waren:

jaar	Republicans	Democrats	anderen
1960	34,108,157	34,226,731	0
1964	27,178,188	43,129,484	0
1968	31,785,480	31,275,166	9,906,473
1972	47,169,911	29,170,383	1,099,482
1976	39,147,973	40,830,763	756,631
1980	43,899,248	36,481,435	5,719,437
1984	54,455,075	37,577,185	0
1988	48,886,097	41,809,074	0
1992	39,104,545	44,909,889	19,742,267
1996	39,198,755	47,402,357	8,085,402
2000	50,456,002	50,999,897	2,882,955
2004	59,668,261	56,172,264	0

Met uitzondering van de verkiezingen in 2000 is steeds de kandidaat met de meeste stemmen president geworden.

(i) Maak frequentiepolygonen voor de relatieve aantallen stemmen voor de verschillende partijen.

(ii) Bepaal de verdeling van de stemaandelen die de gekozen president in de verschillende verkiezingen heeft behaalt. Maak een doos-en-snorren figuur voor deze verdeling. Zijn er uitschieters? Kun je die verklaren?

(iii) We beperken ons nu tot de stemmen voor de republikanen en de demokraten. In het jaar 2000 heeft dan bijvoorbeeld de kandidaat van de republikanen $50,456,002$ van $50,456,002+50,999,897 = 101,455,899$ stemmen, dus 49.73% van deze stemmen gehaald, en de kandidaat van de demokraten 50.27%. De *afstand* tussen republikanen en demokraten definiëren we als het verschil van deze aandelen, dus -0.54% voor het jaar 2000 (let op het teken).

Bepaal de verdeling van deze afstanden, hun gemiddelde, standaardafwijking, mediaan, kwartielen en interkwartielafstand.

Men zegt dat er een *aardverschuiving* heeft plaatsgevonden als de afstand bij een verkiezing sterk verschilt van de afstand bij de vorige verkiezing. Definieer een criterium, wanneer er sprake van een aardverschuiving is en geef aan bij welke verkiezingen een aardverschuiving heeft plaatsgevonden.

5. Zij $x_1, \ldots, x_n$ een verzameling gegevens waarbij de x_i alleen maar de waarden 0 of 1 kunnen hebben. Stel er zijn $p \cdot n$ gegevens met de waarde 0 en $(1-p) \cdot n$ gegevens met de waarde 1.

(i) Bereken het gemiddelde $\overline{x}$ en de centrale momenten m_k voor $k = 1, 2, 3, 4$.

(ii) Geef de scheefheid en scherptoppigheid van deze verzameling gegevens aan.

(iii) Laat zien dat de scheefheid 0 is dan en slechts dan als $p = 0.5$, dus als de verdeling over de twee mogelijke waarden symmetrisch is.