

Twée standaarddeviaties voor een scheve verdeling?

Arno J. Krul

De standaarddeviatie van een lichaamsmaat wordt vaak gebruikt om percentielwaarden te schatten. Als zo'n maat scheef verdeeld is in plaats van normaal verdeeld, zoals bij veel omtrekmaten van het menselijk lichaam, kunnen de geschatte waarden flink afwijken van de werkelijkheid, in het bijzonder bij extreme percentielen. Bij toepassing, bijvoorbeeld in de kledingindustrie, kan dit tot forse fouten in de maatvoering leiden. Als mogelijke methode voor een goede percentielschatting werd onderzocht of de opdeling van een scheve verdeling bij de mediaan met uitbreiding van beide helften tot een hele verdeling tot betere schattingen leiden. Daarbij dienden gegevens over het lichaamsgewicht van Nederlanders als onderzoeksmateriaal. Aangetoond is dat deze methode aanzienlijke verbeteringen oplevert bij het schatten van percentielwaarden uit een scheve verdeling: de gemiddelde fout bij gebruik van één verdeling is 6,1% en bij twee verdelingen (met hetzelfde gemiddelde maar met verschillende standaarddeviaties) is die gedaald tot 1,1%. Voor een kledingfabrikant is dat bijvoorbeeld van belang om de productie beter op de praktijk te kunnen afstemmen.

Inleiding

Idealiter moet elk product dat de mens gebruikt zijn afgestemd op de lichaamsmaten van de doelgroep. Kennis van de afmetingen van het menselijk lichaam is daarom belangrijk. Denk bijvoorbeeld maar eens aan stoelen, waarbij de zittinghoogte aangepast moet zijn aan de kniehoogte van de doelgroep. Vaak worden bepaalde percentielwaarden aangehouden om ontwerpeisen te definiëren, zoals P5 en P95. Voor iedereen met een kniehoogte die daartussen valt, dus 90% van de populatie (90% van de hele doelgroep), moet de zittinghoogte dan geschikt zijn.

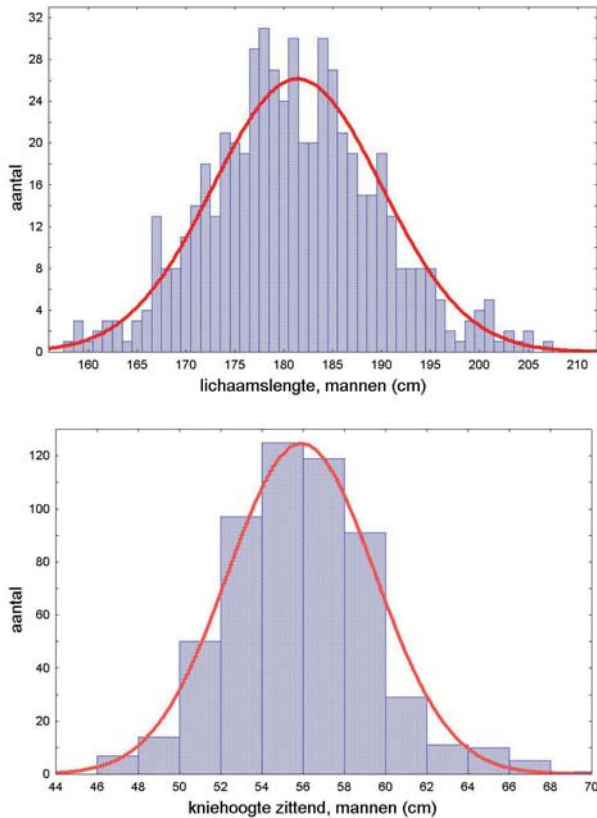
Als een bepaalde lichaamsafmeting een normale verdeling heeft, zoals bij lichaamslengte, kniehoogte en de meeste andere lengte-, breedte- en dieptematen het geval is, kan elk percentiel gemakkelijk berekend

worden uit het gemiddelde en de standaarddeviatie van die maat. Daarbij wordt wel uitgegaan van aparte verdelingen voor mannen en vrouwen; als die twee worden samengevoegd is de gecombineerde verdeling niet noodzakelijk ook normaal. De nieuwe Dined-tabel (Daanen e.a., 2003) geeft deze uitgangswaarden voor mannen en vrouwen en beschrijft de methode voor het berekenen van percentielen. Volgens Molenbroek (1994) wijken berekende percentielen van deze in de ergonomie veel gebruikte maten nooit meer dan 2,5% van de werkelijke waarde in de populatie af. Figuur 1 toont een paar voorbeelden van zulke normaal verdeelde lichaamsafmetingen. In de figuren is telkens de best passende normale verdeling door een doorgetrokken lijn weergegeven; de kolomhoogte geeft het aantal personen in de betreffende categorie weer.

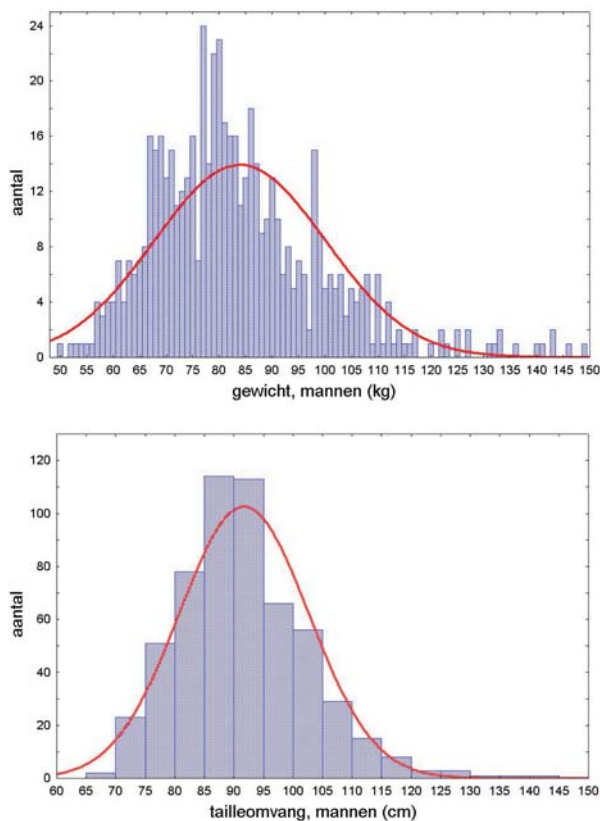
Voor maten met een niet-normale verdeling (zie de voorbeelden in figuur 2) levert deze procedure voor het berekenen van percentielen problemen op. Dat is lastig, want er zijn ook werkgebieden waar veel niet-normaal verdeelde afmetingen gebruikt worden en daarbij kunnen percentielwaarden dus niet goed berekend worden. Allerlei omtrekmaten, zoals borst-

Arno Krul is statistisch medewerker bij TNO Technische Menskunde.

Adres: Postbus 23, 3769 ZG Soesterberg;
Tel. 0346 – 356229. E-mail: krul@tm.tno.nl



Figuur 1: Enkele normaal verdeelde lichaamsafmetingen bij mannen: Lichaamslengte en kniehoogte



Figuur 2: Enkele niet-normaal verdeelde lichaamsafmetingen bij mannen: gewicht en tailleomvang

omvang, tailleomvang en heupomvang, zijn bijvoorbeeld voor de kledingbranche (confectiefabrikanten, KPU-bedrijf van Defensie) essentieel om een goed maatsysteem te kunnen hanteren, maar deze maten bezitten nu juist geen normale verdeling. Molenbroek (1994) meldt bij omtrekmaten en ook bij lichaamsgewicht verschillen tussen berekende percentielen en werkelijk gemeten waarden die tot 4% oplopen. Deze verdelingen zijn scheef naar rechts, met een duidelijke ondergrens en met een lange staart naar rechts.

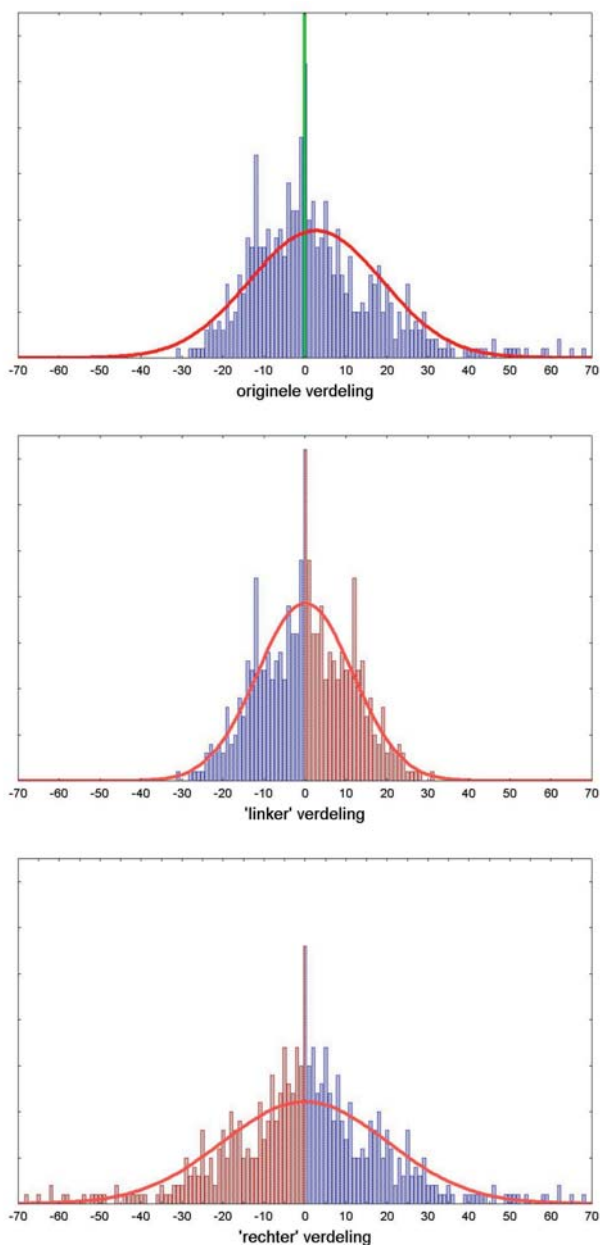
Molenbroek heeft gesuggereerd dat misschien een betere schatting van percentielwaarden bij scheve verdelingen kan worden bereikt door gebruik te maken van twee standaarddeviaties in plaats van één. Hij stelt voor om een scheve verdeling in tweeën te knippen bij de mediaan (P50, het 50e percentiel) en vervolgens van beide helften weer complete verdelingen te maken door aan elk hun spiegelbeeld toe te voegen. Figuur 3 toont een voorbeeld: de originele verdeling is scheef met een lange staart naar rechts, dat wil zeggen met veel meer hoge waarden dan bij een normale verdeling verwacht. Vervolgens bepaalt men van elk van die twee nieuwe verdelingen de standaarddeviatie. De percentielen worden dan op de gebruikelijke manier berekend, alleen hangt het van het percentiel af welke standaarddeviatie in de berekening wordt gebruikt: voor lage percentielen (tot P50) wordt de standaarddeviatie van de 'linker' verdeling gebruikt, voor de hoge percentielen die uit de 'rechter' verdeling.

Aan de hand van een bestaande dataset (Daanen & Robinette, 2001) wordt in dit artikel nagegaan of de door Molenbroek voorgestelde procedure helpt om tot betere percentielschattingen te komen bij een scheve verdeling.

Methode

Om de procedure te testen zijn gegevens uit het NedScan-databestand (Daanen & Robinette, 2001) gebruikt. In het NedScan-project is van 1255 mannen en vrouwen tussen 18 en 65 jaar een groot aantal lichaamsmaten bepaald. Als voorbeeld is in dit onderzoek lichaamsgewicht gebruikt om de uitkomsten van het werken met twee standaarddeviaties bij een scheve verdeling te bekijken. Omdat allerlei omtrekmaten een vergelijkbare verdeling vertonen, zullen die maten waarschijnlijk overeenkomstige resultaten opleveren.

Het databestand is in vier groepen verdeeld: mannen en vrouwen, en van elk twee leeftijdsklassen (18-30



Figuur 3: Voorbeeld van het splitsen van een niet-normale verdeling bij de mediaan (de groene lijn) in twee verdiepingen

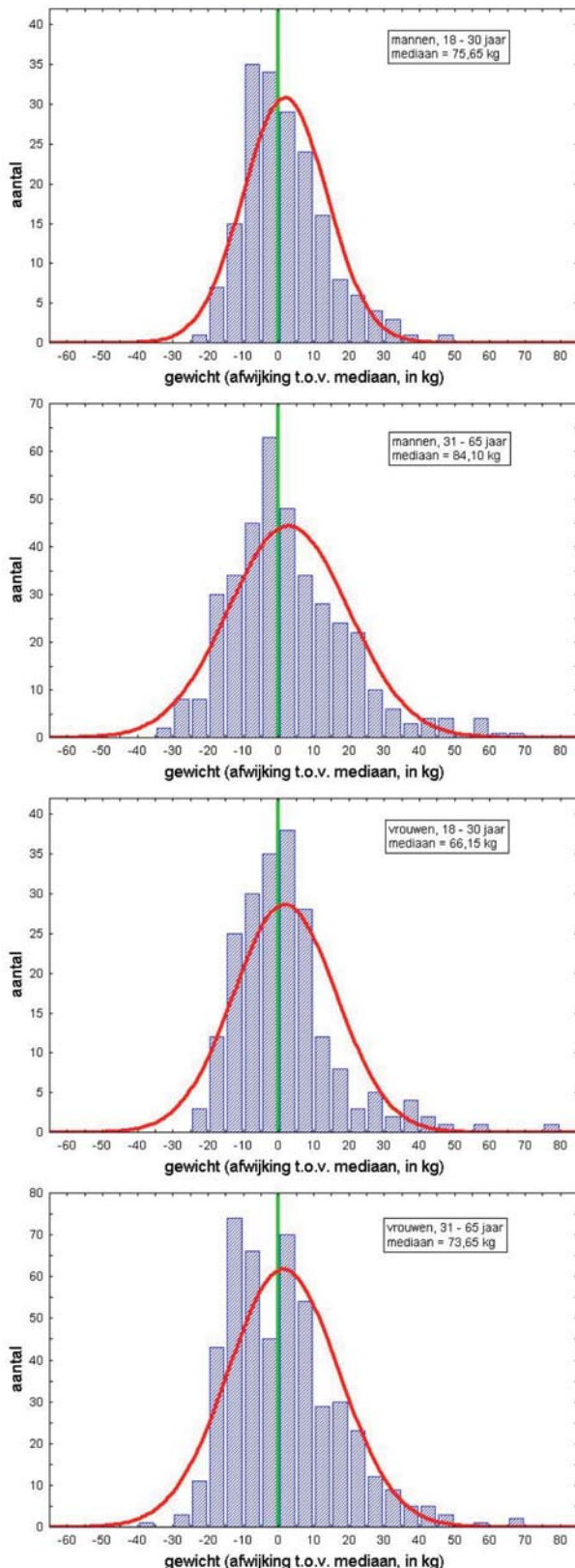
en 31-65 jaar). Voor elke groep is het gemiddelde gewicht en de standaarddeviatie bepaald. Ook is de scheefheid van de verdeling berekend. Deze maat, berekend als $\sum (X_i - \mu)^3 / N\sigma^3$, geeft aan of een verdeling afwijkt van een symmetrische verdeling. Is de verdeling symmetrisch, dan is de scheefheid 0, een links scheve verdeling levert een negatieve waarde voor de scheefheid op, een rechts scheve verdeling een positieve waarde. Is de absolute waarde van de berekende scheefheid groter dan twee maal de standaardfout ervan (geschat als $2\sqrt{6/N}$), dan wordt de verdeling als significant scheef beschouwd. De scheefheid is hier in alle 4 gevallen significant (dus afwijkend van een symmetrische verdeling). Verder is met de Kolmogorov-Smirnov-toets nagegaan of de gevonden verdeling significant afwijkt van de best-passende normale verdeling. Tabel 1 geeft de resultaten. De verdelingen zijn afgebeeld in figuur 4.

Elk van deze groepen wordt nu in tweeën gedeeld. Als scheiding wordt de mediaan gebruikt, omdat die niet, zoals het gemiddelde, gevoelig is voor uitbijters; het doet er niet toe hoe groot precies de hoogste waarden in de rechter 'staart' van de verdeling zijn. In figuur 4 is in de verdelingen de scheidslijn (de mediaan) ingetekend.

Van beide helften worden vervolgens complete verdelingen gemaakt door ze aan te vullen met hun spiegelbeeld. Voor de halve verdeling met waarden beneden de mediaan gaat dat als volgt: van elke waarneming wordt het verschil met de mediaan bepaald en vervolgens wordt een waarde met een even groot verschil met de mediaan, maar dan aan de andere kant van de mediaan, toegevoegd. Voor een jonge vrouw met een gewicht van 60 kg (in de groep met een mediaan van 66,10, zie tabel 1) wordt dus een extra waarde gecreëerd van $66,10 + (66,10 - 60) = 72,20$. Elke waarneming x levert zo een extra waarde medi-

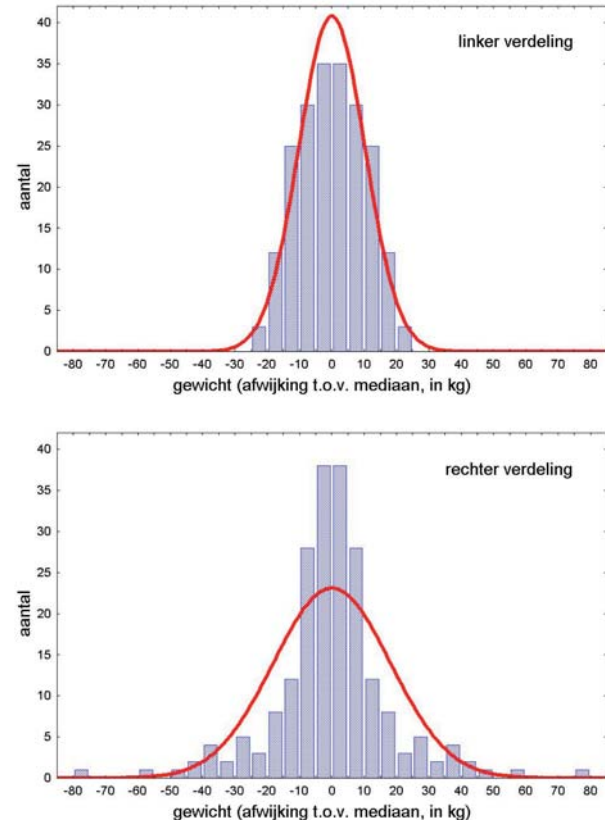
Tabel 1: Kenmerken van lichaamsgewicht voor 4 groepen, verdeeld in sexe en leeftijd, uit de NedScan-data

geslacht	leeftijd (jaar)	gemiddelde	standaard-deviatie	mediaan	scheefheid (0 = niet scheef)	Kolmogorov-Smirnov toets (afwijking van normaliteit)
man	18 - 30	77,45	11,94	75,65	0,85	niet significant
	31 - 65	87,08	17,03	84,1	0,9	$p < 0,05$
vrouw	18 - 30	68,06	14,3	66,15	1,62	$p < 0,01$
	31 - 65	75,08	15,62	73,65	0,89	$p < 0,05$



Figuur 4: Verdeling van lichaamsgewicht bij mannen en vrouwen (2 leeftijdscategorieën) Data uit NedScan (zie tabel 1)

aan $+ (\text{mediaan} - x)$ op, waardoor een symmetrische verdeling ontstaat. Uiteraard werkt het net zo voor



Figuur 5: Verdelingen voor gewichtsverdeling bij vrouwen (18-30), verkregen door spiegeling t.o.v. de mediaan

de halve verdeling met waarden boven de mediaan, alleen liggen de extra waarden dan juist beneden de mediaan: elke waarde x levert dan een waarde $\text{mediaan} - (x - \text{mediaan})$ op. Deze verdelingen worden aangeduid als respectievelijk de linker en de rechter verdeling. Figuur 5 toont als voorbeeld de nieuw ontstane verdelingen voor vrouwen tussen 18 en 30 jaar (ontstaan uit de verdeling van figuur 4c). Duidelijk is te zien dat door de lange rechter staart in de originele verdeling de spreiding bij de linker verdeling veel kleiner is dan bij de rechter verdeling.

Voor elk van de vier groepen zijn er nu twee verdelingen, waarvan weer standaarddeviaties berekend kunnen worden. Daarmee kunnen schattingen voor P5 en P95 worden berekend (uit gemiddelde $\pm 1,645 \times$ standaarddeviatie), voor P5 met de standaarddeviatie van de linker en voor P95 met de standaarddeviatie van de rechter verdeling. Daarnaast worden ook uit de originele verdelingen P5 en P95 berekend en worden uit die verdelingen de echte P5 en P95 bepaald. Nu is het mogelijk om te zien of de uitgevoerde procedure tot verbetering van de percentielschattingen heeft geleid. Als de procedure effect heeft, is de verwachting dat de percentielen van de originele verdelingen (de echte P5 en P95) beter

geschat worden door de P5 en P95 die uit de nieuwe verdelingen berekend zijn dan door de P5 en P95 die uit de originele verdelingen met behulp van de standaarddeviatie zijn berekend.

Voor alle statistische analyses en voor het maken van de figuren is gebruik gemaakt van het programma Statistica (StatSoft, 2004).

Resultaten

Tabel 2 geeft voor elk van de vier groepen dezelfde gegevens als tabel 1, maar nu niet over de originele verdelingen maar over de gesplitste, dus voor elke groep twee verdelingen. De gemiddelden zijn nu natuurlijk gelijk aan de medianen van de originele verdelingen en de scheefheid wordt automatisch 0 (ze zijn immers symmetrisch ten opzichte van het gemiddelde). Uit de laatste kolom van tabel 2 blijkt dat deze verdelingen beter normaal verdeeld zijn dan de originele. Alleen bij vrouwen van 18 - 30 jaar veroorzaakt de rechter staart van de verdeling nog een significante afwijking van normaliteit (zie ook figuur 5), terwijl bij vrouwen boven de 30 jaar de linker verdeling enigszins niet-normaal is

Zowel uit de originele verdelingen als uit de geconstrueerde verdelingen kunnen nu percentielen geschat worden. Verder zijn uit de originele verdelingen ook de echte P5 en P95 bepaald. Tabel 3 geeft de uitkomsten van de vergelijking van percentielschattingen en echte percentielwaarden voor de 4 groepen.

De grijze rijen in tabel 3 geven de uitkomsten voor de originele verdelingen voor elk van de vier groepen. Zowel de echte P5 en P95 als de op basis van de originele verdeling geschatte waarden daarvan zijn hier

vermeld. De fouten lopen uiteen van 3,5 tot 5,8 kg voor P5 en van 2,3 tot 5,5 kg voor P95 of in percentages van de echte waarden: bij P5 van 5,7% tot 11,4% en bij P95 van 2,0% tot 5,7%.

Na elke grijze regel volgen in tabel 3 telkens twee regels met de resultaten als de verdeling gesplitst wordt bij de mediaan. De eerste regel geeft de standaarddeviatie van de linker verdeling en de daaruit geschatte P5-waarde en de tweede regel de standaarddeviatie van de rechter verdeling en de daaruit geschatte P95-waarde.

Als nu opnieuw schattingen voor P5 en P95 berekend worden, maar dan met gemiddelden en standaarddeviaties van respectievelijk de linker en de rechter verdeling, blijken de afwijkingen ten opzichte van de echte waarden veel kleiner te zijn. Voor P5 is de afwijking nu 0,2 tot 2,0 kg (0,4 tot 3,6%) en voor P95 0,0 tot 1,5 kg (0,0 tot 1,5%).

De gemiddelde fout bij gebruik van de originele verdeling is 4,2 kg (6,1%). Met twee verdelingen neemt deze gemiddelde fout af tot 0,8 kg (1,1%).

Discussie

Onderzocht is of het gebruik van twee waarden voor de standaarddeviatie van een verdeling (één voor het deel met waarden tot de mediaan en één voor de waarden die groter zijn dan de mediaan) helpt om bij scheve verdelingen betere percentielschattingen te maken. Bij de onderzochte gevallen treedt inderdaad een duidelijke verbetering van die schattingen op.

De gepresenteerde methode is hier getest met gegevens over het lichaamsgewicht van Nederlanders. De

Tabel 2: Kenmerken van lichaamsgewicht voor 4 groepen, verdeeld in sexe en leeftijd, en elk opgedeeld in 2 verdelingen (kleiner en groter dan de mediaan)

geslacht	leeftijd (jaar)	verdeling (t.o.v. mediaan)	gemiddelde = mediaan	standaarddeviatie	Kolmogorov-Smirnov toets (afwijking van normaliteit)
man	18 - 30	linker	75,65	8,97	niet significant
		rechter		14,54	niet significant
	31 - 65	linker	84,10	12,38	niet significant
		rechter		21,09	niet significant
vrouw	18 - 30	linker	66,15	9,74	niet significant
		rechter		17,88	p < 0,01
	31 - 65	linker	73,65	12,43	p < 0,05
		rechter		18,38	niet significant

gebruikte steekproef is afkomstig uit NedScan, een goed gedocumenteerd bestand, representatief voor de inwoners van Nederland (Daanen & Robinette, 2001). Lichaamsgewicht is een maat met een duidelijk scheve verdeling: een minimum, onder andere bepaald door skelet en lichaamsbouw, een als gezond beschouwd gebied (BMI tussen 20 en 25), waarin zich een groot deel van de bevolking bevindt, en een groep met overgewicht, variërend van een BMI net boven de 25 tot heel ver daarboven, dus een verdeling met een lange rechter staart. Aangenomen mag daarom worden dat de hier gevonden uitkomsten ook gelden voor andere maten met een soortgelijke scheve verdeling, bijvoorbeeld omtrekmaten van het menselijk lichaam.

Voor normaal verdeelde gegevens (dus met een symmetrische verdeling) zal het niet uitmaken of men de methode wel of niet toepast, omdat dan de standaarddeviaties van beide helften van de gesplitste verdeling ongeveer gelijk zullen zijn aan die van de originele verdeling.

Hoe belangrijk is nu de correctie bij scheve verdelingen? Bij de uitkomsten voor lichaamsgewicht gaat het bij de procentuele fout (ten opzichte van de werkelijke percentielwaarden) om een vermindering van maximaal 11,4% naar 3,6% voor P5 en van maximaal 5,7% naar 1,5% voor P95. Dit zijn natuurlijk forse verbeteringen. Men moet echter wel overwegen dat de grootste fout bij gebruik van de originele verdeling 5,8 kg bedraagt (P5 bij vrouwen van 31-65 jaar) en dat met gebruik van de hier gepresenteerde methode nog een fout van 2,0 kg overblijft. Toch kan

zo'n verbetering veel uitmaken. Voor een kledingfabrikant bijvoorbeeld, die als doelgroep mensen in een relatief extreme groep kiest (mensen t/m P10 of vanaf P90 bijvoorbeeld), betekent verbetering van de schatting van de verdeling van omtrekmaten dat hij veel preciezer zijn productie kan afstemmen op de werkelijke afmetingen van de potentiële klanten. De gewenste precisie van de percentielschattingen zal dus bepalend moeten zijn voor het toepassen van de methode.

Conclusie

Bij een scheve verdeling is de standaarddeviatie geen goede maat voor de spreiding. Daardoor zijn (uit gemiddelde en standaarddeviatie van die verdeling berekende) schattingen voor percentielwaarden onbetrouwbaar. Om toch met normale verdelingen te kunnen rekenen, kan men een kunstgreep toepassen: opdeling van de (scheve) verdeling bij de mediaan, completeren van beide halve verdelingen door middel van spiegeling rond de mediaan en bepaling van de standaarddeviatie voor elk van die nieuwe verdelingen apart. De percentielwaarden die met behulp van de standaarddeviaties van deze nieuwe verdelingen geschat worden, benaderen de echte percentielwaarden in de meeste gevallen heel goed, duidelijk beter in ieder geval dan de schattingen op basis van de originele verdeling. Of men de methode wil gebruiken, zal afhangen van het belang dat men hecht aan de nauwkeurigheid van percentielschattingen.

Lees verder op de volgende pagina

Tabel 3: Echte en geschatte waarden van het 5e en 95e percentiel in lichaamsgewicht (P5 en P95), verdeeld in sexe en leeftijd

geslacht	leeftijd (jaar)	verdeling	gemiddelde	standaarddeviatie	P5		P95	
					echt	geschat	echt	geschat
man	18-30	origineel	77,45	11,94	61,3	57,8	99,6	97,1
		gesplitst	75,65	links: 8,97 rechts: 14,54		60,9		99,6
	31-65	origineel	87,08	17,03	64,4	59,1	117,4	115,1
		gesplitst	84,1	links: 12,38 rechts: 21,09		63,7		118,8
	18-30	origineel	68,06	14,3	50,2	44,5	97,1	91,6
		gesplitst	66,15	links: 9,86 rechts: 17,88		50		95,6
vrouw	31-65	origineel	75,08	15,62	55,2	49,4	103,7	100,8
		gesplitst	73,65	links: 12,43 rechts: 18,38		53,2		103,9

Referenties

Daanen, H.A.M., Krul, A.J. en Molenbroek, J.F.M. 2003. DINED-2003. Kerckebosch, Zeist.

Daanen, H.A.M. en Robinette, K.M. 2001. CAESAR: The Dutch dataset. TNO rapport TM-01-C026, Soesterberg.

Molenbroek, J.F.M. 1994. Op maat gemaakt. Delftse Universitaire Pers, Delft.

StatSoft, Inc. 2004. Statistica (data analysis software system), version 6. www.statsoft.com.

DANKWOORD

De auteur bedankt Johan Molenbroek en Hein Daanen voor hun inbreng bij de totstandkoming van dit artikel.

Summary

The standard deviation of a body dimension is often used to estimate percentile values. When the measure is skew and not normally distributed, as for instance in many circumference measures of the human body, it is possible to obtain large deviations between estimated and actual values, especially for extreme percentiles. This may lead to substantial errors in applications (e.g. in the clothing industry). It is important to make better percentile estimates. Body weight was taken as an example for this study. The skewed distribution of body weight was divided in two parts (mirrored around the median, each distribution with its own standard deviation). The use of the standard deviations of these two distributions compared to that of the original distribution, diminished the mean error in the estimation of the percentiles from 6.1% to 1.1%. 



Observe this!

Noldus Information Technology bv
Wageningen, The Netherlands
Phone: +31-317-497677
E-mail: info@noldus.nl

Noldus Information Technology GmbH
Freiburg, Germany
Phone: +49-761-4701600
E-mail: info@noldus.de

Noldus Information Technology Inc.
Leesburg, VA, U.S.A.
Phone: +1-703-771-0440
Toll-free: 1-800-355-9541
E-mail: info@noldus.com

www.noldus.com

As a researcher, you have to spend your resources as efficiently as possible. Saving time during the collection, analysis and presentation of observational data can mean a big help. And that's exactly what The Observer can do for you. Whether you are observing live or from video, The Observer offers you an easy and accurate tool for the study of activities, postures, movements or human-system interactions.

Noldus

Information Technology

Innovative tools for behavioral research

Er volgt nog een kwart advertentie. Opdrachtgever onbekend