

Hoofdstuk 10 A. Een verticale boxplot bouwen

Bonus hoofdstuk bij het boek *Datavisualisatie met Excel* door Wim de Groot (c)

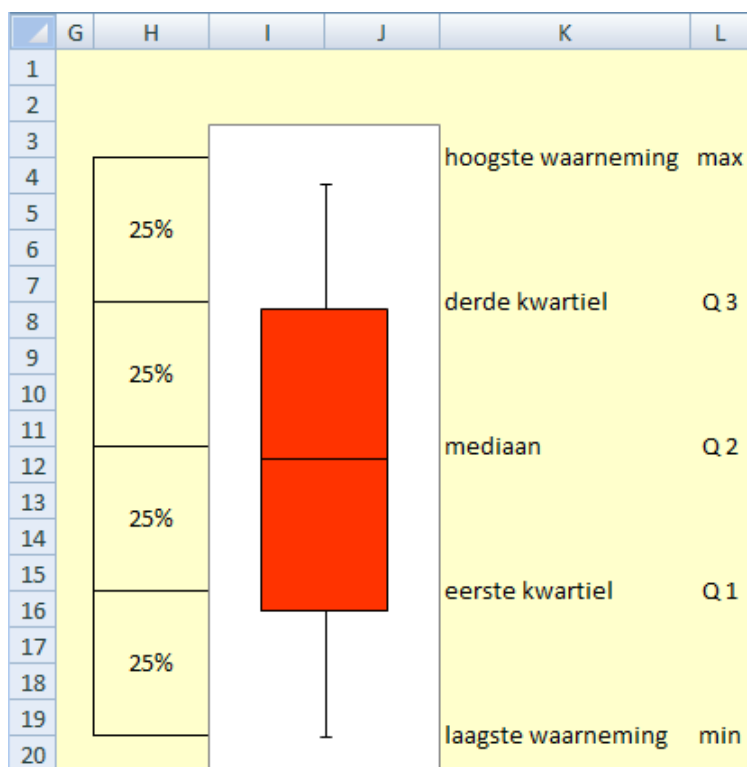
*In mijn boek **Datavisualisatie met Excel** leest u in hoofdstuk 10 'Een box en whisker opstellen', hoe u met Excel 2016 een boxplot maakt, zowel voor één groep als voor drie groepen gegevens. U kunt deze ook in oudere versies van Excel maken. Dat leg ik hier uit. U leest hoe u een verticale boxplot maakt, compleet met markeringen voor de uitschieters.*

Een boxplot bouwen

Voor de verticale boxplot gebruiken we de zogeheten Gestapelde kolom. Dat houdt in, dat we een aantal waarden als het ware op elkaar stapelen. We moeten de afstanden tussen een aantal gegevens berekenen en die vervolgens als blokjes op elkaar zetten.

Vanaf de onderkant van de doos (Q1) loopt er een lijn naar de kleinste waarde en van de bovenkant van de box (Q3) loopt er een lijn naar het maximum. Deze lijnen worden *whiskers* genoemd (snorharen, als van een kat). Eventuele uitschieters in de verzameling worden weergegeven als cirkels boven en onder deze lijnen.

Bij een boxplot staat in het midden een rechthoek, de doos (een *box* in het Engels). Deze doos geeft de beide middelste kwartielen weer en loopt van Q1 tot en met Q3. De afstand tussen deze beide kwartielen wordt berekend met $Q3 - Q1$; binnen deze grenzen liggen 50 procent van alle waarden.



Afbeelding 1. Een boxplot verdeelt de gegevens in vier gebieden. Zo ziet de boxplot eruit als de waarden precies gelijkmatig verdeeld zijn.

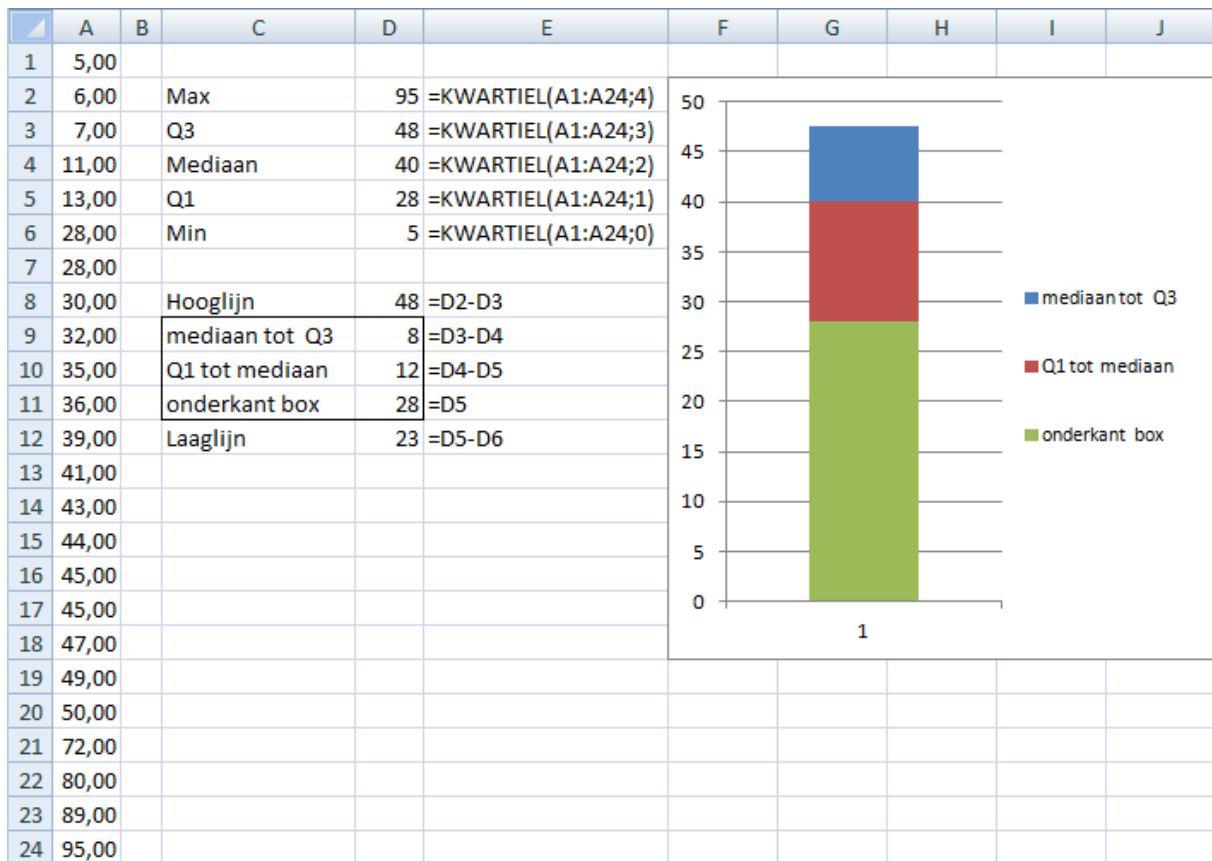
Waarden voor de boxplot berekenen

Om een boxplot te kunnen maken, hebben we nodig:

- de hoogste waarde (maximum),

- het derde kwartiel (Q3); deze vormt de bovenkant van de rechthoek,
- de mediaan, deze wordt gemarkeerd door een lijn in de rechthoek,
- het eerste kwartiel (Q1); deze vormt de onderkant van de rechthoek,
- de laagste waarde (minimum).

In dit voorbeeld gaan we uit van de bedragen die 24 klanten in een supermarkt hebben uitgegeven. Typ in de cellen A1 tot en met A24 willekeurige hele bedragen tussen 5 en 95. Neem de vijf aanduidingen en formules uit de afbeelding over in uw werkblad.



Afbeelding 2. Voor een verticale boxplot neemt u de formules over en begint u met de grafiek Gestapelde kolom.

- Ik bereken deze vijf waarden voor het gemak steeds met de functie KWARTIEL. De mediaan is gelijk aan het tweede kwartiel, het minimum gelijk is aan het nulde kwartiel en het maximum aan het vierde kwartiel.
- Als de boxplot erg nauwkeurig moet zijn en u werkt in Excel 2010 of 2013, dan kunt u de functie KWARTIEL.EXC gebruiken. Deze methode gaat anders met de mediaan om: bij een even aantal waarnemingen aan de linkerkant neemt deze de mediaan niet mee (vandaar: 'exclusief'), zodat er een oneven aantal getallen is; en daarvan neemt deze functie de middelste positie.

De verticale boxplot tekenen

Voor de verticale boxplot gebruiken we de zogeheten Gestapelde kolom. Dat houdt in, dat we een aantal waarden als een blokkentoren op elkaar stapelen. Daarvoor kunnen we de waarden

uit rij 2 tot en met 6 nog niet zonder meer gebruiken, we moeten de afstanden daartussen berekenen en die vervolgens op elkaar zetten. Neem hiervoor de formules in rij 8 tot en met 12 over uit de afbeelding.

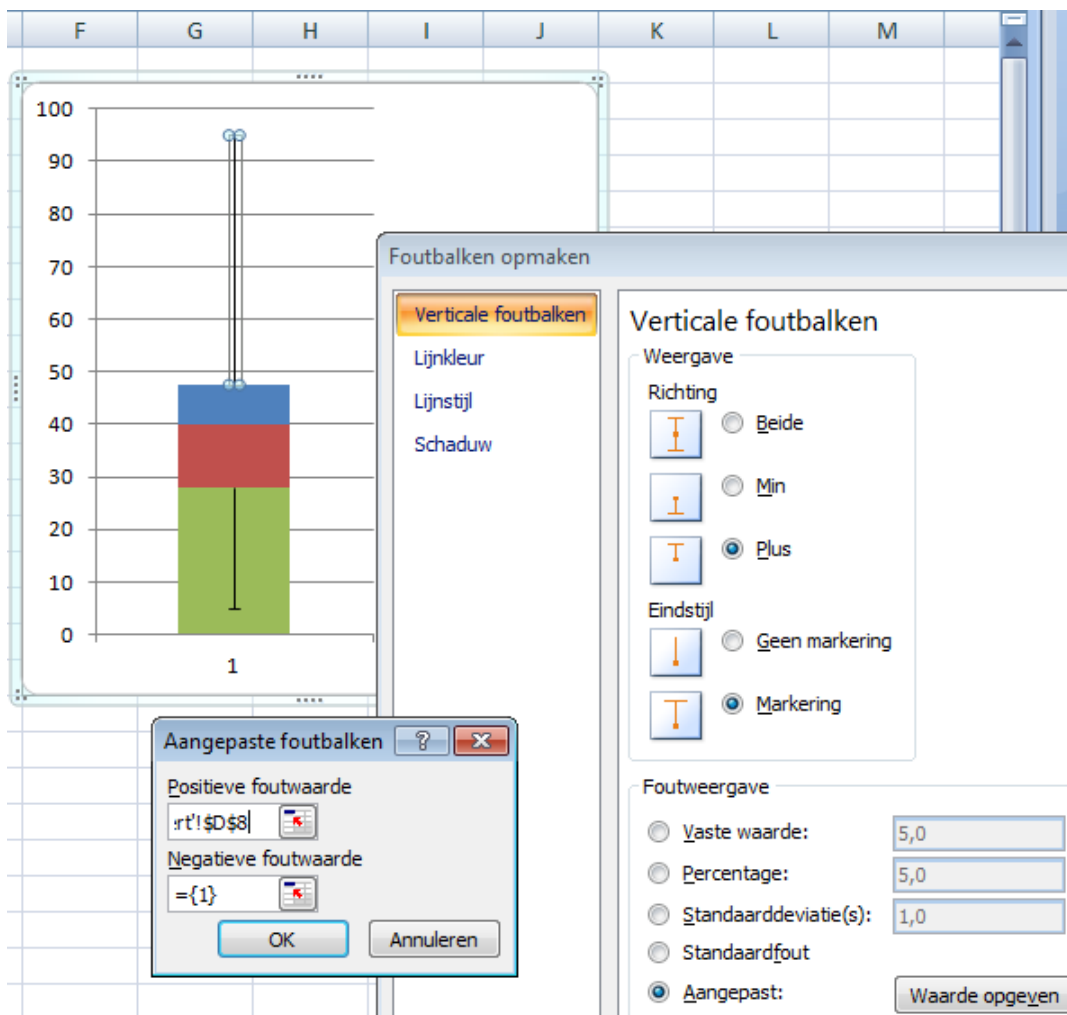
Selecteer de cellen C9 tot en met D11 ('mediaan tot Q3', 'Q1 tot mediaan' en 'onderkant box' met hun waarden). Klik op **Invoegen**, klik op **Kolom** en kies onder **2D- kolom** de **Gestapelde kolom** (let op: kies niet de 100% gestapelde kolom); er verschijnt een grafiek met drie kolommen naast elkaar. Klik op de grafiek, klik op de tab **Ontwerpen** en kies **Rijen/kolommen omdraaien**; nu hebt u één kolom die uit drie delen bestaat. Alleen de volgorde klopt nog niet.

Klik daarom op de grafiek en klik in de tab **Ontwerpen** op **Gegevens selecteren** (of kies dat na een rechtermuisklik); in het dialoogvenster dat verschijnt, ziet u de drie reeksen terug. Verander met behulp van de pijltjes midden in dit venster de volgorde zo, dat u hierin van boven naar beneden 'onderkant box', 'Q1 tot mediaan' en 'mediaan tot Q3' ziet. In dit dialoogvenster staan ze precies andersom dan in de grafiek zelf (zie de legenda), maar zo komt het goed.

- Maak deze grafiek maar meteen wat smaller door de rechterrاند naar binnen te slepen;
- haal de horizontale as onder aan de grafiek weg;
- verwijder de legenda.

De whiskers aanbrengen

Het middelste en het bovenste vlak vormen straks de box. Hier vandaan trekken we de lijnen; de ene *whisker* moet van de onderkant van het middelste blok tot het minimum lopen. Klik hiervoor op het onderste deel van de kolom (de reeks 'onderkant box'), klik op de tab **Indeling**, klik op **Foutbalken** en op **Meer opties voor foutbalken**. In het venster dat opengaat, klikt u op de optie **Min**, op **Aangepast** en op **Waarde opgeven**. Er gaat een klein venster open. Klik in het vak onder **Negatieve foutwaarde**, haal $=\{1\}$ weg en klik op cel D12 (de waarde van de 'Laaglijn'). Dit zorgt voor een lijn in de grafiek van de onderkant van het middelste vlak tot de minimumwaarde. De lengte van deze Laaglijn is het verschil tussen Q1 en de minimumwaarde (D5 minus D6), in ons voorbeeld is dat $28 - 5$ is 23. Vanaf Q1 (28) gaat deze lijn 23 omlaag en zo staat het dwarsstreepje precies op de minimumwaarde 5. Klik voor een lijn vanaf de box naar boven op het bovenste vlak (de reeks 'mediaan tot Q3'), klik weer op **Foutbalken** en op **Meer opties voor foutbalken**. Klik in het venster dat opengaat, op de optie **Plus**, op **Aangepast** en op **Waarde opgeven**. Maak in het kleine venster het vak onder Positieve foutwaarde leeg en klik op cel D8 (de waarde van de 'Hooglijn'); zo krijgt u een lijn van de bovenkant van het vlak tot het maximum.



Afbeelding 3. Met foutbalken brengt u lijnen aan van de box naar de uiteinden.

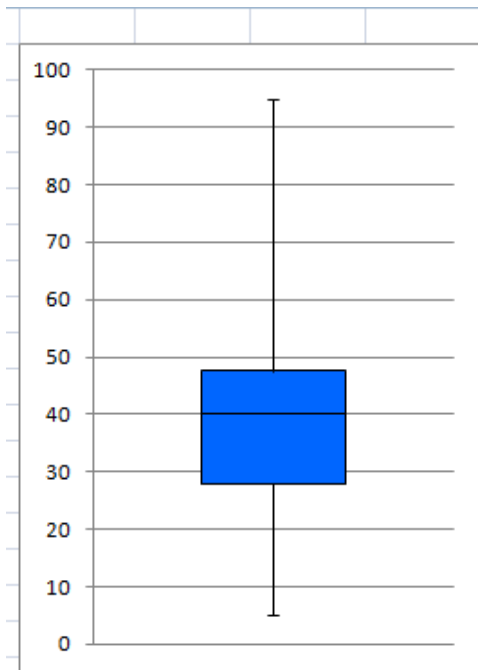
De vlakken aanpassen

Het onderste vlak van de kolom draagt de box, maar dit vlak hoeven we niet te zien. Om dat onzichtbaar te maken klikt u op het onderste vlak, klikt u in de tab **Opmaak** op **Opvulling van vorm** en op **Geen opvulling**.

De beide vlakken die overblijven, vormen samen de box. Excel heeft deze automatisch twee verschillende kleuren gegeven, maar het is gebruikelijk om de beide delen van de box dezelfde kleur te geven. Klik hiervoor op het ene vlak, klik in de tab **Opmaak** op **Opvullen van vorm** en kies uw kleur. Herhaal dit voor het andere vlak (of klik daarop en druk op Ctrl+Y).

Om de mediaan te kunnen zien, hoeven we alleen maar de rand om de beide vlakken van de box aan te brengen. Klik hiervoor op het ene vlak, klik in de tab **Opmaak** op **Omtrek van vorm** en op **Automatisch**. Herhaal dit voor het andere vlak.

- Wilt u op de linkeras de cijfers van 0 tot en met 100 in stappen van 10 zien, klik dan met de rechtermuisknop op die as en kies **As opmaken**. In het venster dat opent, kiest u onder **Opties voor as** bij **Minimum**, **Maximum** en **Primaire eenheid** de optie **Vast** en vult u respectievelijk 0, 100 en 10 in;
- hoeft u de horizontale rasterlijnen niet te zien, klik dan op een van die lijnen en druk op de Delete-toets;
- verwijder de legenda.



Afbeelding 4. En uw verticale boxplot is klaar. Zoals u ziet, ligt de helft van de uitgaven in de supermarkt tussen 28 en 48 euro. Aan de plaats van de mediaan in de box ziet u meteen of de verzameling scheef verdeeld is.

Uitschieters bepalen

Wat u misschien wel ziet aan de getallen zelf, maar niet aan de boxplot, is dat er eigenlijk vier uitschieters zijn (ook wel uitbijters genoemd, in het Engels *outliers*). We kunnen de boxplot uitbreiden om de uitschieters ook weer te geven. De uitschieters worden opgespoord met de regel: $1,5 \times$ interkwartielafstand (IKA of in het Engels IQR, *Inter Quartile Range*). Eerst wordt de interkwartielafstand genomen, dat is het verschil tussen Q3 en Q1, dus de afstand tussen de beide kwartielen aan weerszijden van de mediaan. Deze afstand wordt met $1,5$ vermenigvuldigd en vervolgens van Q1 afgetrokken om de ondergrens te bepalen en bij Q3 opgeteld om de bovengrens te bepalen. Alle waarden die buiten dit bereik vallen, zijn dan uitschieters.

Als de uitschieters worden weergegeven, loopt de onderste lijn niet meer van Q1 naar de minimumwaarde, maar van Q1 naar de nieuwe ondergrens (die ligt op $1,5 \times$ de interkwartielafstand onder de box). En de bovenste lijn loopt nu niet meer van Q3 naar het maximum, maar van Q3 naar de nieuwe bovengrens (die ligt $1,5 \times$ IKA boven de box).

Interkwartielafstand berekenen

We moeten de interkwartielafstand berekenen en de nieuwe grenzen. We nemen als voorbeeld weer de bestedingen in de supermarkt, die in A1 tot en met A24 staan. De formules uit de volgende afbeelding verschaffen de informatie voor de boxplot.

- Hebt u de verticale boxplot al gemaakt met de aanwijzingen uit de paragraaf *De verticale boxplot tekenen*, dan moet u alleen de formules op rij 2 en 6 veranderen en de formules vanaf rij 14 toevoegen.

	A	B	C	D	E	F
1	5,00				Formules in kolom D	<i>Betekenis:</i>
2	6,00	Max		77	=MIN(D16;MAX(A1:A24))	
3	7,00	Q3		48	=KWARTIEL(A1:A24;3)	←
4	11,00	Mediaan		40	=KWARTIEL(A1:A24;2)	<i>interkwartielafstand</i>
5	13,00	Q1		28	=KWARTIEL(A1:A24;1)	←
6	28,00	Min		5	=MAX(D17;MIN(A1:A24))	
7	28,00					
8	30,00	Hooglijn		29	=D2-D3	<i>Max-Q3</i>
9	32,00	mediaan tot Q3		8	=D3-D4	<i>Q3-Mediaan</i>
10	35,00	Q1 tot mediaan		12	=D4-D5	<i>Mediaan-Q1</i>
11	36,00	onderkant box		28	=D5	<i>Q1</i>
12	39,00	Laaglijn		23	=D5-D6	<i>Q1-Min</i>
13	41,00					
14	43,00	1,5 * IKA		29	=(D3-D5)*1,5	<i>1,5 * interkwartielafstand</i>
15	44,00					
16	45,00	bovengrens uitschieters		77	=D3+D14	<i>Q3 + 1,5 * IKA, uitschieters boven Max</i>
17	45,00	ondergrens uitschieters		-1	=D5-D14	<i>Q1 - 1,5 * IKA, uitschieters onder Min</i>
18	47,00					

Afbeelding 5. Om de uitschieters weer te geven in de grafiek, verleggen we eerst de grenzen van de boxplot. Neemt u de formules in kolom D over.

De formule in D14 berekent de interkwartielafstand, dat is het verschil tussen Q3 en Q1, en die wordt met 1,5 vermenigvuldigd, met de formule:

$$= (D3 - D5) * 1,5$$

In D16 wordt deze afstand bij Q3 opgeteld om de bovengrens te bepalen met:

$$= D3 + D14$$

De waarden daarboven zijn de uitschieters omhoog.

In D17 wordt deze afstand van Q1 afgetrokken om de ondergrens te bepalen met:

$$= D5 - D14$$

De waarden daaronder zijn de negatieve uitschieters.

De formule in D2 is veranderd. Deze regelt nu hoever de hooglijn loopt. Die gaat niet meer gewoon tot de maximumwaarde, maar tot het punt waarboven de waarden als uitschieters gelden. Maar als er geen uitschieter is hoger dan de maximumwaarde, gaat de lijn tot dat maximum. Deze formule betekent: als D16 kleiner is dan het maximum, neem dan D16; neem anders het maximum. Dat had gekund met:

$$= \text{ALS} (D16 < \text{MAX} (A1 : A24) ; D16 ; \text{MAX} (A1 : A24))$$

maar die heb ik ingekort tot:

$$= \text{MIN} (D16 ; \text{MAX} (A1 : A24))$$

Ook de formule in D6 is aangepast en geeft niet zonder meer de minimumwaarde. Deze is het spiegelbeeld van D2 en regelt hoever de laaglijn loopt. Die gaat niet meer tot de laagste waarde, maar de waarde waaronder de uitschieters omlaag beginnen. Maar als er geen uitschieter is lager dan de minimumwaarde, gaat de lijn tot dat minimum. Hier staat voluit: als D21 groter is dan het minimum, neem dan D21; neem anders het minimum. De lange versie hiervoor is:

$$= \text{ALS} (D21 > \text{MIN} (A1 : A24) ; D21 ; \text{MIN} (A1 : A24))$$

Die heb ik ingekort tot

$$= \text{MAX} (D21 ; \text{MIN} (A1 : A24))$$

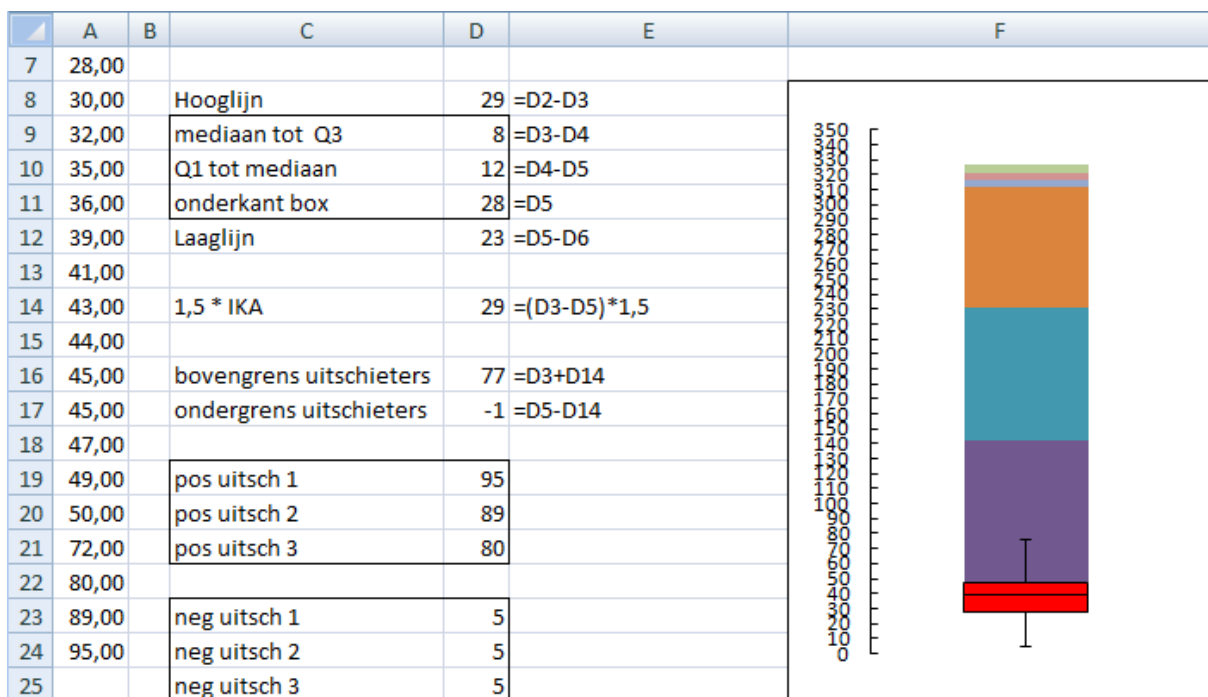
Hebt u de boxplot gemaakt met de aanwijzingen in de vorige paragrafen, dan hoeft u de celverwijzingen van de grafiek niet te veranderen. De waarde van de whiskers (de foutbalken) haalt de grafiek uit dezelfde cellen (namelijk D8 en D12), alleen wordt hun lengte nu anders berekend.

Met de formules uit de tabel zal het dwarsstreepje van de bovenlijn iets verschoven zijn: dat stond eerst bij het maximum van 95, maar staat nu op het punt waarboven de uitschieters beginnen, bij 77 in dit voorbeeld. Het streepje van de onderste lijn in dit voorbeeld is niet verschoven. Want pas onder -1 zou er sprake zijn van uitschieters, maar er zijn geen lagere waarden van -1, de laagste waarde is 5, dus dat was en blijft de ondergrens.

Uitschieters aan de boxplot toevoegen

Om de uitschieters in de grafiek weer te geven, berekenen we die in een aparte tabel. We zullen ze straks op een speciale manier in de grafiek plaatsen. Zouden we die nu meteen automatisch berekenen, dan zijn ze straks lastig terug te vinden om te bewerken. Daarom doen we het andersom: we typen eerst zelf een aantal waarden en pas als die goed in de grafiek staan, plaatsen we in die cellen de formules die de uitschieters automatisch berekenen. Typ daarom in D19, D20 en D21 drie verschillende getallen, die groter zijn dan het getal dat bij 'bovengrens uitschieters' staat. Typ ook in D23, D24 en D25 drie verschillende getallen, die kleiner zijn dan Q1 (maar neem nul niet).

Klik met de rechtermuisknop op de grafiek en klik op **Gegevens selecteren**; het venster **Gegevensbron selecteren** gaat open. Klik op **Toevoegen**, laat in het venstertje dat open gaat, het vak **Reeksnaam** leeg, maak het vak **Reekswaarden** leeg, klik op cel D19 en klik één keer op **OK**. Klik weer op **Toevoegen**, maak het vak **Reekswaarden** leeg, klik op cel D20 en klik één keer op **OK**. Herhaal dit voor alle zes cellen. U hebt nu de reeksen 6 tot en met 11 aan de grafiek toegevoegd. Er is voor iedere reeks een vlak op de toren gestapeld; daar gaan we meteen iets aan doen.



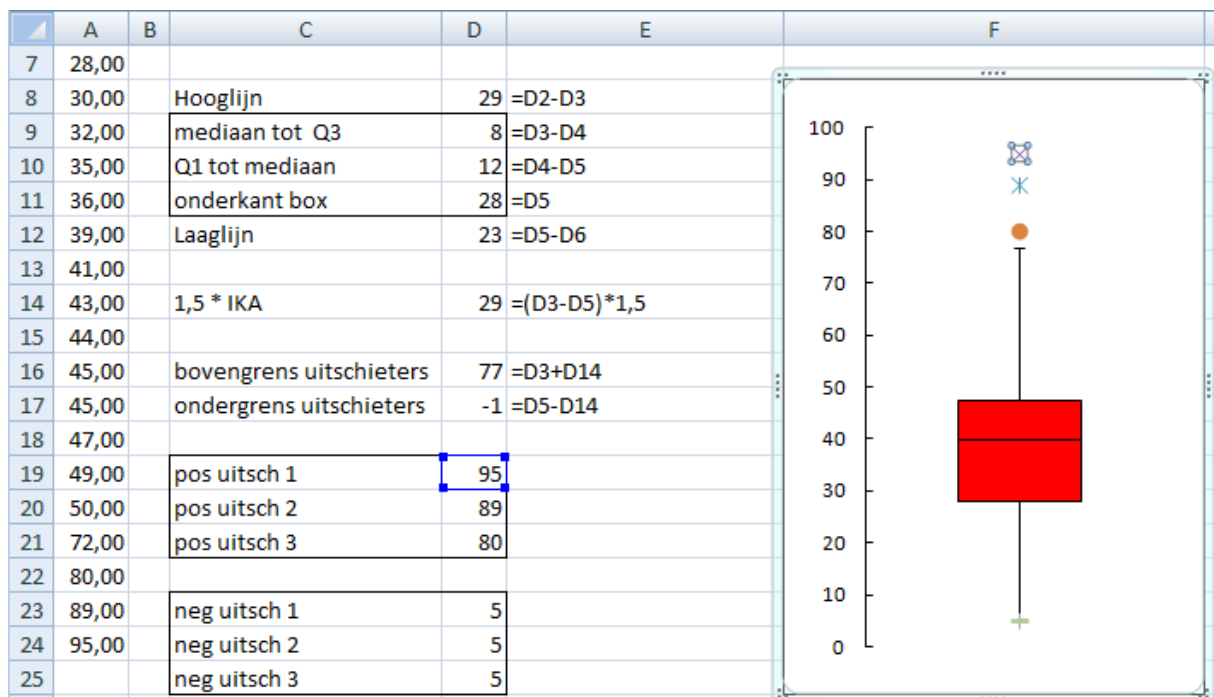
Afbeelding 6. We zetten de uitschieters in een apart tabelletje bij elkaar. Ze komen voorlopig bovenop de blokkentoren.

Om de uitschieters boven en onder de hoog- en laaglijn te krijgen, kiezen we voor hun waarden het type Spreidingsgrafiek. Klik op een van de reeksen die u zojuist boven op de kolom hebt gekregen.

- Als u de muisaanwijzer op een van die vlakjes houdt, verschijnt er Reeks met een nummer. Klik daarop.
- Is een vlak zo klein dat u er niet gemakkelijk op kunt klikken, vergroot dan het beeld door de Ctrl-toets ingedrukt te houden en aan het wiel van de muis te draaien. Lukt het nog niet, klik dan op de grafiek, klik op de tab **Opmaak**; helemaal links in beeld verschijnt een keuzelijst. Kies met die keuzelijst om te beginnen **Reeks6**.

Klik op de tab **Ontwerpen**, op **Ander grafiek type**, klik op de knop **Spreiding** en kies het eerste type, **Spreiding met alleen markeringen**.

Herhaal deze stappen voor alle zes vlakjes van de uitschieters. Hierna worden de reeksen 6 tot en met 11 weergegeven met kleine markeringen: dit zijn de uitschieters en die komen nu boven en onder de box.



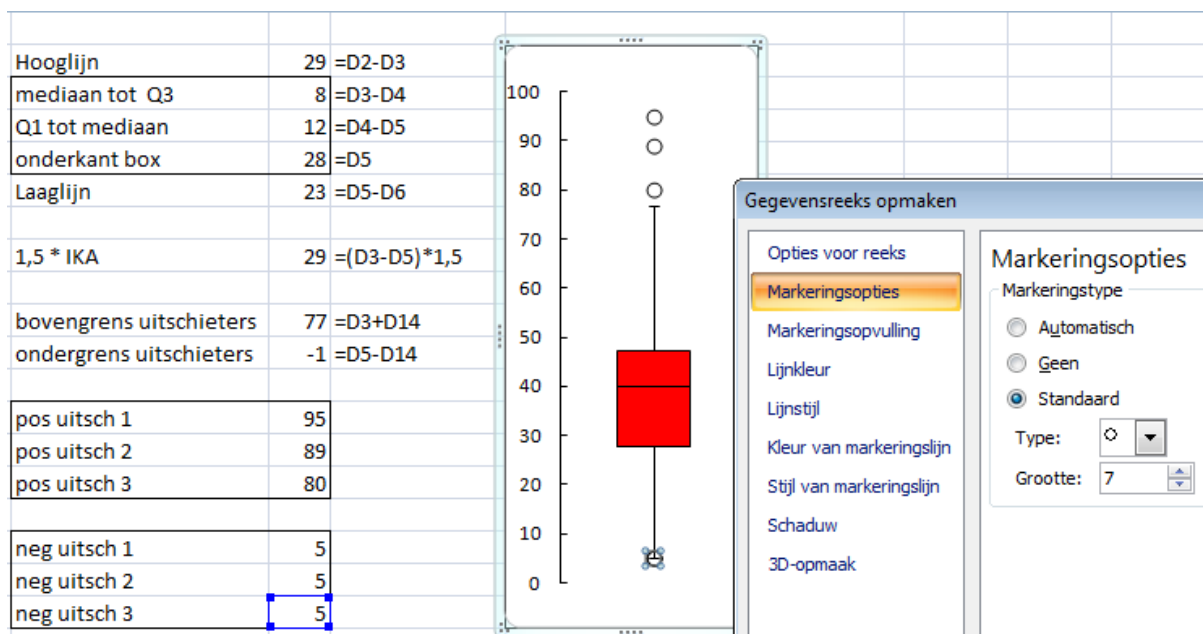
Afbeelding 7. Om de uitschieters weer te geven, geven we de reeksen 6 tot en met 11 weer als Spreidingsgrafiek.

Uitschieters als cirkels weergegeven

Uitschieters worden gewoonlijk als cirkels in de grafiek weergegeven. Dat kan in Excel ook. Rechtsklik op een markering en kies **Gegevensreeks opmaken**; dit opent een venster. Kies in het tabblad **Markeringsopties** voor **Standaard** en kies als Type het rondje, met als grootte 7. Klik op **Markeringsopvulling** en kies **Geen opvulling**. Klik op **Kleur van markeringslijn**, kies **Ononderbroken streep** en kies via **Kleur** zwart.

- Tip: klik nog niet op **Sluiten**, want u kunt de volgende markering gewoon aanklikken terwijl dit venster open staat en die dan met dit venster opmaken.

Selecteer de volgende markering en herhaal deze stappen voor alle zes reeksen.



Afbeelding 8. Via de opmaak geeft u de uitschieters in de boxplot weer als witte cirkels.

Geeft de grafiek de uitschieters goed weer, dan is het tijd om hun waarde automatisch te berekenen. Neemt u de volgende formules over in uw werkblad. Deze vervangen de waarden die u daar zopas zelf had getypt.

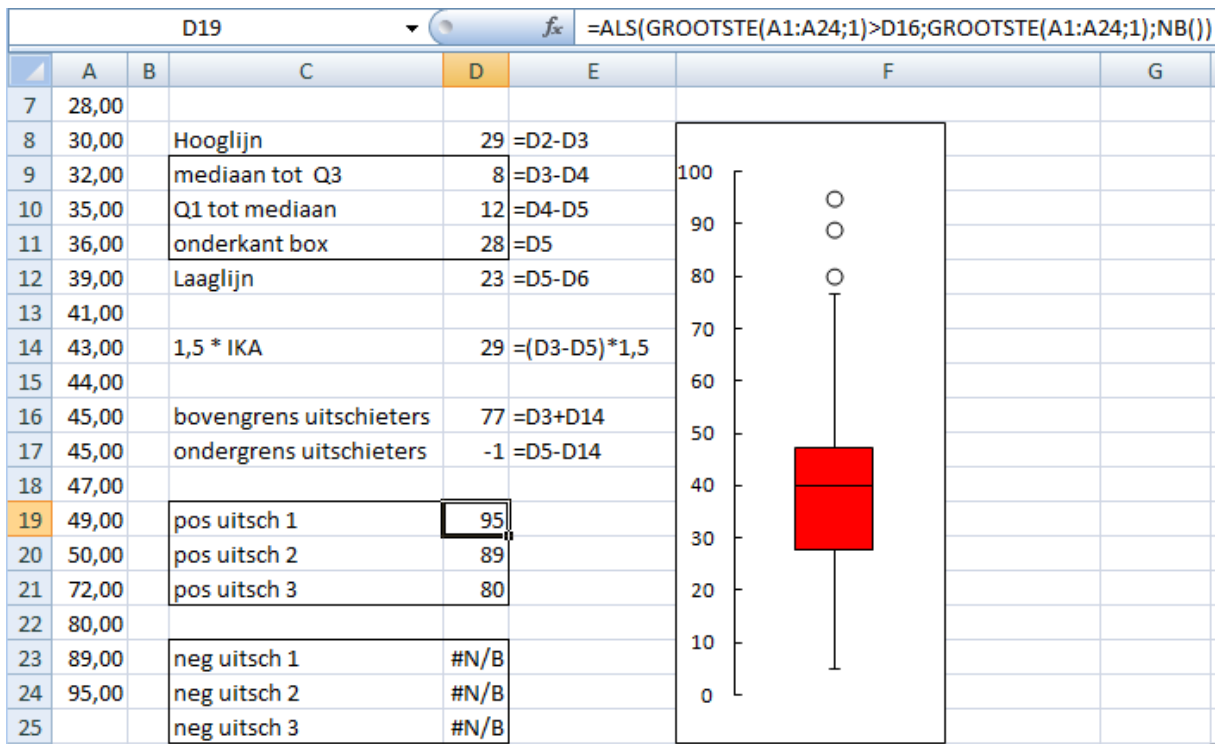
Rij	Kolom C	Formules in kolom D
19	pos uitsch 1	=ALS (GROOTSTE (A1 : A24 ; 1) > D16 ; GROOTSTE (A1 : A24 ; 1) ; NB ())
20	pos uitsch 2	=ALS (GROOTSTE (A1 : A24 ; 2) > D16 ; GROOTSTE (A1 : A24 ; 2) ; NB ())
21	pos uitsch 3	=ALS (GROOTSTE (A1 : A24 ; 3) > D16 ; GROOTSTE (A1 : A24 ; 3) ; NB ())
22		
23	neg uitsch 1	=ALS (KLEINSTE (A1 : A24 ; 1) < D17 ; KLEINSTE (A1 : A24 ; 1) ; NB ())
24	neg uitsch 2	=ALS (KLEINSTE (A1 : A24 ; 2) < D17 ; KLEINSTE (A1 : A24 ; 2) ; NB ())
25	neg uitsch 3	=ALS (KLEINSTE (A1 : A24 ; 3) < D17 ; KLEINSTE (A1 : A24 ; 3) ; NB ())

De formule in D19 zegt: als de grootste waarde van A1 tot en met A24 groter is dan de bovengrens voor de uitschieters (die in D16 staat), geef dan de grootste waarde; geef anders NB(); dat staat voor 'niet beschikbaar' en daardoor geeft de grafiek niets weer.

De formules in D20 en D21 kijken of er eventueel een tweede en derde uitschieter is.

De formules in de onderste helft van de tabel kijken of de kleinste drie waarden van A1 tot en met A24 kleiner zijn dan de ondergrens voor de uitschieters (die grens staat in D17).

Als er in de gegevens waarden voorkomen die buiten deze grenzen vallen, worden ze weergegeven als uitschieters. Met de getallen in het voorbeeld ziet u de cirkels voor de bovenste uitschieters 95, 89 en 80 (die steken boven de grens van 77 uit). We hebben geen waarden kleiner dan de ondergrens van -1, dus daar zien we geen cirkels; in het staatje met formules staat daarvoor drie keer #NB.



Afbeelding 9. De uitschieters worden automatisch berekend en weergegeven als witte cirkels. Deze uitgebreide boxplot geeft door de uitschieters een beter beeld van de werkelijkheid.

In dit voorbeeld bent u voorbereid op drie positieve en drie negatieve uitschieters. Komen er bij u meer uitschieters voor, dan:

- maakt u in uw werkblad ruimte voor langere lijstjes;
- typt u eerst zelf getallen groter dan de bovengrens dan wel de ondergrens van de uitschieters;
- voegt u die cellen toe aan de grafiek;
- geeft u de uitschieters weer als een spreidingsgrafiek;
- maakt u ze op als cirkels;
- kopieert u voor de vierde positieve uitschieter de formule en past u deze zo aan dat er twee keer GROOTSTE(A1:A24;4 staat, voor de vijfde komt er na de puntkomma een 5 enzovoort; voor de vierde negatieve uitschieter moet in het deel KLEINSTE(A1:A24;3 van de formule na de puntkomma een 4 staan (twee keer), voor de vijfde moet er na de puntkomma een 5 enzovoort.