Toegepaste statistiek in het waterbeheer (2)

Enige inleidende begrippen uit de statistiek

Voordat wat geschreven wordt over de toepassing van statistiek in het waterkwaliteitsbeheer moeten de lezers en de auteurs over dezelfde taal beschikken. Daarom nemen Richard Duin (RIKZ) en Jaap van Steenwijk (RIZA) in de tweede aflevering van deze serie de tijd om enkele begrippen uit de statistiek toe te lichten. Ze gaan kort in op wat zij verstaan onder begrippen als populatie, gemiddelde, mediaan, percentiel en een normale verdeling, onder het motto "leuker kunnen we het niet maken, wel makkelijker".

Als we geïnteresseerd zijn in de jaargemiddelde nitraatconcentratie (in milligrammen per liter) in de Rijn bij Lobith, dan nemen we elke week een monster en analyseren het. De monsters vormen een steekproef van alle mogelijke nitraatconcentraties, hetgeen de populatie van de nitraatconcentratie in de Rijn bij Lobith is voor dat jaar. Met deze steekproef hebben we een schatting van het populatiegemiddelde (μ). De schatter is het gemiddelde van de steekproef en wordt aangeduid met $x_{gem.}$ Afhankelijk van natuurlijke variatie en toevallige fouten hebben de meetgegevens een spreiding. Een maat voor die spreiding in de populatie is de standaardafwijking: sigma (σ). Deze is ook bekend in zijn gekwadrateerde vorm: σ^2, de variantie. Ook deze parameter moet geschat worden. Deze schatter is de standaardafwijking (S) van de meetgegevens.

Terugkerend naar het voorbeeld. De Rijn voert gemiddeld 2000 m^3 water per seconde af. Per jaar is dat ongeveer 63 x 10^{12} liters. De steekproef bedraagt 52 monsters van een liter dus één op de 12 x 10^{12}. Als we zo een premier kiezen, is de kans op een 'Jan Peter' erg klein. Met statistiek kunnen we echter wel een schatter berekenen voor het gemiddelde (zeg maar Jan Modaal) en de standaardafwijking zegt dan ook iets over de spreiding van de gegevens binnen dat jaar.

Een andere centrummaat is de mediaan, die minder gevoelig is voor uitschieters dan het gemiddelde. De mediaan bepalen we door alle metingen naar grootte te rangschikken van laag naar hoog. De middelste waarde (bij een even aantal het gemiddelde van de middelste twee waarden) is dan de mediaan. Vaak zoeken we niet een centrummaat, maar is het de vraag of niet meer dan tien procent van de waarnemingen een bepaalde waarde overschrijdt. Anders gezegd: bij welke concentratie ligt 90 procent van de metingen lager: de 90-percentiel. De vaststelling daarvan is analoog aan die van de mediaan, want de mediaan is eigenlijk de 50-percentiel van de gegevens.

Nu is het ene watersysteem het andere niet. Ze verschillen bijvoorbeeld qua gemiddelde nitraatconcentratie, maar ook de spreiding kan variëren. Indien de gegevens normaal verdeeld zijn, liggen de meeste waarden tussen het gemiddelde +/- twee maal de standaardafwijking. Trekken we van alle concentraties de waarde van het gemiddelde af, dan wordt het gemiddelde 0. Delen we vervolgens alle dan verkregen waarden door de standaardafwijking, dan ontstaat de standaard normale curve met een standaardafwijking gelijk aan 1 (zie de grafiek). Elke normale verdeling kan tot deze standaard normale verdeling worden herleid. De zo verkregen waarden heten de Z-scores. Een bekende Z-score is die van het 95 procent betrouwbaarheidsinterval: 1,96. De Z-score heeft als voordeel dat voor elke normale verdeling één tabel volstaat om betrouwbaarheidsintervallen te maken. Zo hoort de Z-score 1,96 bij zowel een rechter als een linker overschrijdingskans van 0,025 procent. Het 95 procent betrouwbaarheidsinterval is daarmee: $\mu \pm 1{,}96\ \sigma/\sqrt{n}$. Enkele andere Z-scores staan in de tabel.

W.S. Gosset, chemicus van de Guinness Brouwerij, werd gevraagd om de kwaliteit van het bier te testen. Hij ontdekte dat bij een klein aantal monsters de kans groter was om een bier af te keuren terwijl de kwaliteit goed was, dan hij op grond van de normale verdeling verwachtte. De Z-score bleek te laag bij kleine steekproeven! Gosset construeerde een verdeling die afhankelijk is van het aantal monsters (meer precies het aantal vrijheidsgraden (df = n - 1). Hij publiceerde onder de naam 'Student', omdat het niet gebruikelijk was resultaten van zijn werkgever te publiceren. De verdeling staat dan ook bekend als Student-t verdeling die voor grote waarden van 'n' (orde n > 50) gelijk is aan de normale verdeling (zie de tabel).

Met deze basisinformatie zullen we de volgende stukjes over toepassing van statistiek in het waterbeheer te lijf gaan.

Voor meer informatie kunt u per e-mail contact opnemen met Jaap van Steenwijk: j.steenwijk@riza.rws.minvenw.nl

Z-scores en Student -t waarden bij 11, 21 en 31 monsters.

P(Z>z) eenzijdig	Z-score	t (df = 10)	t (df = 20)	t (df = 30)
10 %	1,28	1,37	1,32	1,31
5 %	1,65	1,81	1,72	1,77
2,5 %	1,96	2,23	2,09	2,02

Elke willekeurige normale verdeling is om te zetten in de standaard normale verdeling.