Ooit moet het toch echt niet meer sneller kunnen, zou je zeggen. Door technische verbeteringen (overdekte banen, nieuwe schaatspakken, klapschaatsen, enzovoort) worden steeds snellere tijden gerealiseerd. Maar zelfs als deze invloeden eruit gefilterd worden is het eind nog niet in zicht. **Ruud Koning** laat de *extreme waarden theorie* los op schaatstijden en voorspelt nog veel scherpere wereldrecords.

Snel, sneller, snelst: statistiek en 1500 m schaatsen

Inleiding

Citius, *Altius*, *Fortius:* sneller, hoger, sterker. Iedereen kent dit motto van het Internationaal Olympisch Comité. Toch roept het meteen een vraag op: kunnen we eigenlijk wel sneller gaan dan vorig jaar? Hoe vaak zijn records nog te verbeteren? Per slot van rekening zijn er grenzen aan de atletische vermogens van mensen! Toch is het mogelijk gebleken om regelmatig wereldrecords te verbeteren. De belangrijkste reden hiervoor is de technologische vooruitgang: sportkleding is nu beter dan vroeger, schoeisel is verbeterd, de klapschaats is beter dan de 'oude' Noor, trainings- en voedingsmethoden zijn verbeterd en doping is verbeterd (deze laatste vorm van technische vooruitgang is overigens niet toegestaan).

Het blijkt dat de wiskunde iets heeft te vertellen over de frequentie van recordverbeteringen. Het verwachte aantal records neemt slechts langzaam toe met het aantal evenementen, dus als een sport enige tijd bestaat zijn er weinig recordverbeteringen meer te verwachten.
Verder blijkt dat in de statistiek ook iets bekend is over hoe extremen zich gedragen, dus we kunnen niet alleen uitspraken doen over het aantal records, maar ook over recordtijden. Deze analyses zijn echter alleen toepasbaar op een stationaire situatie, waarin dus niets verandert. Gelukkig is er sprake van vooruitgang, dus in de loop van de tijd is men sneller gaan lopen, schaatsen, enzovoort.

Dit is in elk geval duidelijk zichtbaar in het leidende voorbeeld van dit artikel: de toptijden van 1500 m schaatsevenementen (mannen). Bovendien maakt het nogal wat uit voor de tijd of er wordt geschaatst op hoogte in Salt Lake City, of op het schuurpapierijs van Albertville. Hiermee blijken we rekening te kunnen houden, en we kunnen de effecten van schaatsen op hoogte, technische vooruitgang en toernooi-effecten van elkaar scheiden. Zo is het mogelijk om toch iets te zeggen over de kans dat de 1.43:95 van Parra binnenkort weer uit de boeken verdwijnt als mondiale toptijd.

Hoeveel records zijn te verwachten?

Vaak worden sportprestaties gemodelleerd als toevalsvariabelen. Dat is ook wel redelijk, want niemand is in staat om op deterministische wijze aan te geven hoe hard de snelste loper tijdens een bepaald evenement zal lopen of hoe scherp de winnende tijd op de Olympische 1500 m zal zijn. Iedereen die wel eens aan sport doet weet dat het op de ene dag nu eenmaal beter gaat dan op de andere! Als we sportprestaties beschrijven met behulp van toevalsvariabelen, doen we dat soort deterministische uitspraken niet. Wel proberen we bijvoorbeeld de kansverdeling van winnende tijden te bepalen of de kans dat een bestaand record wordt verbeterd.

Hoeveel nieuwe records kunnen we eigenlijk verwachten als er meer evenementen (of metingen) plaatsvinden? Om de gedachten te bepalen kijken we naar de 1500 m schaatstijden, waarbij we voor het gemak net doen alsof elke schaatser alleen schaatst. Naarmate er meer schaatsers de afstand hebben afgelegd, zal de snelste tijd scherper komen te staan. De eerste schaatser heeft natuurlijk even de snelste tijd in handen. Het verwachte aantal records bij één meting is dus één. Voor de tweede schaatser zijn er twee mogelijkheden: hij is sneller dan de eerste schaatser of hij is dat niet. Als hij sneller is, is er sprake van een recordverbetering en is het aantal records twee. Als we nu aannemen dat de schaatstijden onderling onafhankelijke trekkingen zijn uit een identieke verdeling, dan blijkt dat het aantal records R_n in n metingen de volgende verwachting en variantie heeft:

$$\mathrm{E}\, R_n = \sum_{i=1}^{n} \frac{1}{i} \qquad (1)$$

$$\mathrm{var}\, R_n = \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i^2} \right) \qquad (2)$$

Deze verwachting en variantie zijn *onafhankelijk* van de verdeling van de schaatstijden zelf. Een iets preciezere formulering van dit resultaat en het bewijs staat in de Appendix. Veel informatie over de kansverdeling van het aantal recordverbeteringen staat in Glick (1978). Als het

aantal metingen groot is, kunnen we die verwachting en variantie eenvoudig benaderen. Aangezien

$$\sum_{i=1}^{n} \frac{1}{i} - \ln n = \gamma \approx 0.5772$$

(γ is de constante van Euler) en

$$\sum_{i=1}^{n} \frac{1}{i^2} \to \frac{\pi^2}{6}$$

$$\mathrm{E}\, R_n \approx 0.5772 + \ln n \qquad (3)$$

$$\mathrm{var}\, R_n \approx 0.5772 + \ln n - \frac{\pi^2}{6} \qquad (4)$$

Dit is een sombere boodschap voor topsporters. Het aantal records groeit slechts logaritmisch met het aantal pogingen! Verdubbeling van het aantal schaatsevenementen zou, naar verwachting, slechts leiden tot 0.7 (≈ ln 2) nieuwe records. Gelukkig zien we in de praktijk dat records regelmatig worden verbeterd. De aanname dat de tijden identieke kansverdelingen volgen is niet erg realistisch.

De kansverdeling van extremen

Het is niet alleen mogelijk om iets te zeggen over de frequentie van records, maar ook over de scherpte van die records! De statistische theorie die dit allemaal behandelt staat bekend onder de naam 'extreme value theory'[1]. In zekere zin is deze theorie vergelijkbaar met de centrale limietstelling. Stelt u zich voor dat we van een aantal onderling onafhankelijke metingen uit een identieke verdeling het gemiddelde berekenen. In dat geval volgt dat gemiddelde een normale verdeling als het aantal waarnemingen maar groot genoeg is. Iets preciezer geformuleerd: als $T_1, \ldots, T_n$ onderling onafhankelijk verdeelde toevalsvariabelen zijn uit een verdeling met verwachting μ en variantie σ^2, dan geldt, als het aantal waarnemingen groot genoeg is:

$$\sqrt{n}\,(\overline{T} - \mu) \sim N(0,\sigma^2).$$

De afwijkingen van het steekproefgemiddelde rond het populatiegemiddelde volgt een normale verdeling, als die afwijking tenminste wordt geschaald met een factor $\sqrt{n}$. Bestaan er nu ook dergelijke limietstellingen voor het maximum van een reeks toevalsvariabelen, waarbij we het steekproefmaximum centreren en misschien schalen? Het antwoord op de vraag is, merkwaardig genoeg, bevestigend. Als we bijvoorbeeld eens aannemen dat de T's een exponentiële verdeling volgen, met parameter 1 (dus $\Pr(T \le t) = 1 - \exp(-t)$). Het is niet erg interessant om naar de kansverdeling van $M_n = \max(T_1,\ldots,T_n)$ te kijken voor toenemende n: het maximum zal groter en groter worden. Is het dan wel mogelijk om iets te zeggen over de kansverdeling van het maximum als we van dat maximum iets aftrekken? Dat is inderdaad het geval, en hier is ln n een goede keuze:

$$\begin{aligned}
&\Pr(M_n - \ln n \le x) = \Pr(M_n \le x + \ln n) \\
&= \Pr(T_1 \le x + \ln n, \ldots, T_n \le x + \ln n) \\
&= \Pr(T_1 \le x + \ln n)^n = (1 - \exp(-\ln n)\exp(-x))^n \\
&= \left(1 - \frac{1}{n}\exp(-x)\right)^n \to \exp(-\exp(-x)).
\end{aligned}$$

Net zoals het geval is bij de centrale limietstelling heeft het maximum een welgedefinieerde kansverdeling, mits het maximum goed genormaliseerd is. De kansverdeling in dit geval is de Gumbel-verdeling.
Het bovenstaande voorbeeld is eenvoudig, omdat de kansverdeling van T eenvoudig is. Echter, er bestaat een algemener resultaat:

Stelling 1 *Als de verdeling van* $M_n = \max(T_1, \ldots, T_n)$ *bestaat, dan is die van de vorm*

$$\Pr\left(\frac{1}{c_n}(M_n - d_n) \le x\right) = \begin{cases} \exp\left(-\left[\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right) & \xi \ne 0 \\ \exp(-\exp(-x)) & \xi = 0 \end{cases} \qquad (5)$$

Onafhankelijk van de kansverdeling van de toevalsvariabelen waaruit we trekken, heeft het genormaliseerde maximum een bekende kansverdeling! Deze kansverdeling heet de 'gegeneraliseerde extreme waarde verdeling', omdat hij een generalisatie is van de drie klassieke extreme waarde verdelingen: de Weibull-, Gumbel- en Fréchetverdelingen. De verdeling waaruit we trekken bepaalt natuurlijk wel wat we moeten gebruiken voor c_n en d_n. Als de T's een exponentiële verdeling volgen, dan hebben we $c_n = 1$, $d_n = \ln n$, en $\xi = 0$. Trekken we uit een standaardnormale verdeling, dan hebben we

$$c_n \sim \frac{1}{\sqrt{2\ln n}},\ d_n \sim \frac{1}{\sqrt{2\ln n}} - \frac{\ln\ln n + \ln 4\pi}{2\sqrt{2\ln n}} \quad \text{en } \xi = 0$$

en komen de gegevens uit een uniforme (0,1)-verdeling en hebben we $c_n = \frac{1}{n}$, $d_n = 1$, en $\xi = -1$. In figuur 1 staan de kansdichtheden van M_n als we gegevens hebben die normaal verdeeld zijn. Duidelijk is dat de kansverdeling van het maximum verder naar rechts ligt als de steekproef groter is. Dat is niet zo verbazingwekkend. Verder is spreiding van de kansverdeling voor $n = 50$ groter dan die voor $n = 20$.

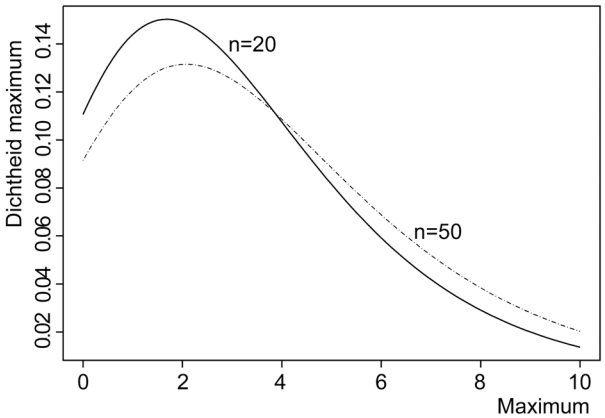

fig. 1 De kansdichtheid van het maximum uit een normaal verdeelde populatie

Deze stelling is buitengewoon belangrijk. De hoogte van de Delta-werken is erop gebaseerd (zie De Haan, 1990), banken bepalen zo hun reserves en de stelling vormt de basis voor het verzekeren van grote rampen, zoals een orkaan in het zuiden van de Verenigde Staten. Echter, ook de capaciteit van bijvoorbeeld het telefoonnetwerk wordt op basis van deze stelling bepaald: per slot van rekening wil je graag dat met grote kans aan de *maximale* belasting voldaan kan worden, en niet de *gemiddelde* belasting!

De theorie aan het werk

In deze paragraaf zullen we de theorie van extremen illustreren aan de hand van de ontwikkeling van toptijden op de 1500 m. De vooruitgang van het wereldrecord op deze afstand staat grafisch weergegeven in figuur 2. Het eerste officiële wereldrecord werd geschaatst op 11 januari 1893 door Jaap Eden op de natuurijsbaan in Groningen. Zijn tijd toen was 2.35, metingen van tienden of honderdsten van seconden waren nog niet mogelijk. Het huidige wereldrecord staat op 1.43:95 en is tijdens de Olympische Spelen van Salt Lake City geschaatst door Derek Parra. Uiteraard zijn de omstandigheden sinds 1893 veranderd: de schaatsen zijn verbeterd, de pakken zijn verbeterd, het ijs is beter, de trainingsmethoden zijn vooruitgegaan, en topprestaties kunnen tegenwoordig worden geleverd in indoorbanen op hoogte. Kortom, de kansverdeling van toptijden in 1893 is verschillend van die van toptijden anno 2002.

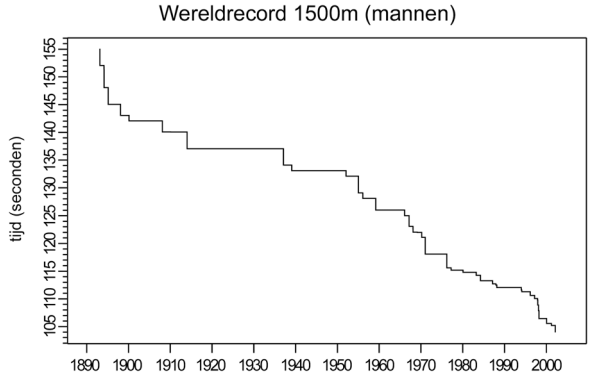

fig. 2 De ontwikkeling van het wereldrecord 1500 m

Teneinde het effect van al die factoren enigzins te kunnen kwantificeren, moeten we eerst meer in detail zien hoe de tijden zich hebben ontwikkeld. Voor dit onderzoek gebruiken we schaatstijden die zijn gerealiseerd na de invoering van de klapschaats.
De gegevensverzameling bestaat uit alle tijden gerealiseerd tijdens Olympische Spelen, wereldbekerwedstrijden en wereldkampioenschappen afstanden sinds het seizoen 1996/97[2]. In de figuren 3 en 4 zien we door middel van boxplots hoe de verdeling van de tijden zich heeft ontwikkeld. Bij de interpretatie van de figuren moet niet worden vergeten dat het seizoen 2002/2003 tijdens het schrijven van dit artikel nog bezig was.

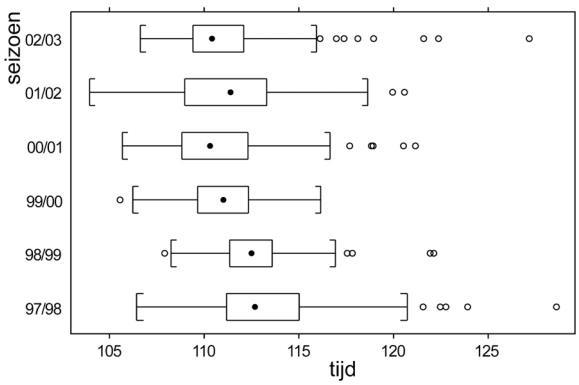

fig. 3 De verdeling van 1500 m tijden per seizoen

Allereerst zien we in figuur 3 dat de toptijden (de staafjes aan de linkerkant van de doos) een licht dalende trend volgen. Dat kan niet worden gezegd van de mediane tijd (de bolletjes in de doos): de gemiddelde schaatser is de afgelopen vijf seizoenen niet veel harder gaan schaatsen, terwijl het lijkt alsof de snelste schaatsers wel iets sneller zijn gaan schaatsen.

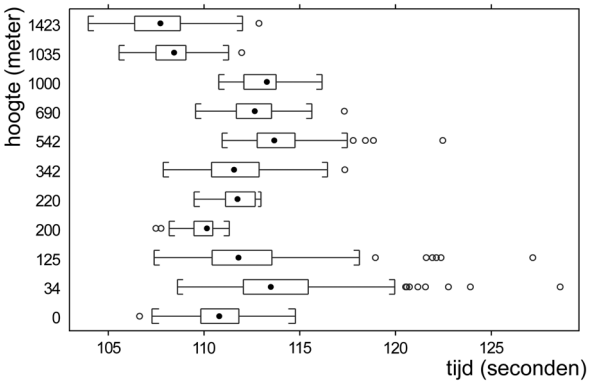

fig. 4 De verdeling van 1500 m schaatstijden per hoogte

Het effect van hoogte op schaatstijden zien we in figuur 4. Op de verticale as staat de hoogte van de baan, en voor elke baan is een boxplot gemaakt van alle tijden die op die baan zijn gerealiseerd. De banen op 1035 m en 1423 m springen er duidelijk uit: dat zijn Calgary en Salt Lake City. Er lijkt geen groot verschil te zijn tussen de tijden die geschaatst worden op de andere banen. Klaarblijkelijk weegt het hoogtevoordeel van de banen in Inzell (690 m) en Baselga di Pine (1000 m) niet op tegen het feit dat dit buitenbanen zijn.
Als we de parameters van de extreme waarde verdeling voor schaatstijden willen schatten, moeten we rekening houden met deze heterogeniteit: niet elke waarneming komt uit dezelfde kansverdeling. Bovendien kunnen we niet alle waarnemingen gebruiken, want de extreme waarden theorie gaat over extremen en niet over alle waarnemingen. Om die reden kijken we in het onderstaande alleen maar naar de vijf snelste tijden per evenement. Teneinde rekening te houden met de heterogeniteit, volgen we een aanpak die econometristen niet vreemd zal

voorkomen. De figuren 3 en 4 suggereren dat de locatie van de verdeling van de toptijden varieert over tijd en hoogte. We nemen aan dat die variabiliteit constant is, en dat we de locatie kunnen modelleren.We nemen aan dat de gemeten toptijden de volgende kansverdeling volgen:

$$Pr\,(T \le t) = \begin{cases} \exp\left(-\left[\frac{t-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right) & \xi \neq 0 \\ \exp(-\exp(-t)) & \xi = 0 \end{cases}$$

$\mu = \mu_0 + \mu_1 H + \mu_2 D + \mu_3 O + \mu_4 S98/99 + \mu_5 S99/00 + \mu_6 S01/02 + \mu_7 S02/03$

De parameter μ van de extremen waarde verdeling hangt af van de omstandigheden waaronder de tijd werd gerealiseerd: het seizoen (*S*98/99 tot en met *S*02/03), de baanhoogte *H* en het toernooi (*D* en *O*). Zo is de μ voor een wedstrijd die op hoogte is geschaatst, geen Wereldkampioenschappen afstanden of Olympische Spelen, in het seizoen 2000/01, gelijk aan $\mu_0 + \mu_1 + \mu_6$. Dezelfde wedstrijd één seizoen later heeft een μ gelijk aan $\mu_0 + \mu_1 + \mu_7$. Op deze wijze laten we toe dat toptijden variëren per seizoen. De effecten staan vermeld in de tabel hieronder.

Schattingsresultaten

	schatting	stdfout
μ	110.13	0.15
hoogte	– 1.94	0.30
afstanden	– 0.76	0.25
olympisch	– 1.26	0.27
s98/99	– 1.85	0.19
s99/00	– 1.03	0.25
s00/01	– 2.74	0.22
s01/02	– 0.52	0.04
s02/03	– 0.58	0.10
σ	– 2.74	0.22
ξ	– 0.70	0.23

We zien in de tabel dat de waarde van μ voor een wereldbekerwedstrijd op een laaglandbaan 110.13 is. Tijdens speciale toernooien wordt harder geschaatst: deze parameter daalt met 0.76 als het een wereldkampioenschap afstanden betreft, en met 1.26 als het om de Olympische Spelen gaat. In de tabel zien we verder dat schaatsen op hoogte leidt tot duidelijk lagere tijden, en dat elk seizoen gemiddeld genomen sneller wordt geschaatst dan in het referentiepunt 1996/97. Grafisch worden de resultaten weergegeven in figuur 5. De meest rechter kansverdeling is die van de winnende tijd tijdens een wereldbekerwedstrijd in het seizoen 1997/98. De kansverdeling die daar links van ligt, is die van de winnende tijd tijdens eenzelfde wedstrijd op een hooglandbaan. De dun gestippelde lijn geeft de kansverdeling van een winnende tijd van een wereldbekerwedstrijd in het seizoen 2001/02, op een laaglandbaan. Deze verdeling ligt links van de verdeling uit 1997/98, en dat geeft de mate van technische vooruitgang weer. Dit verschil is de grafische weergave van de

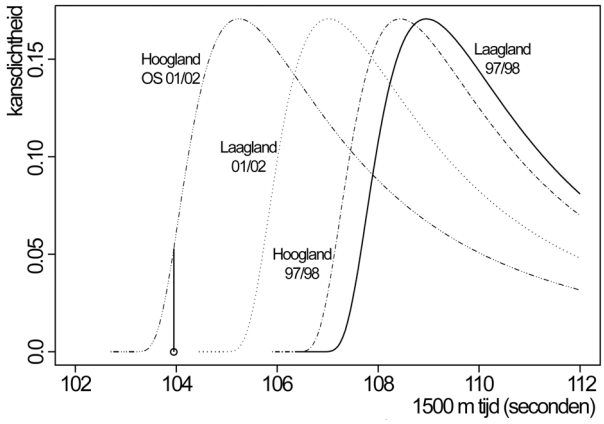

fig. 5 De extreme waarde verdeling van 1500 m schaatstijden

geschatte coëfficiënt -0.52 in tabel 1. De meest linker kansverdeling ten slotte is die van de winnende tijd tijdens een van de Olympische Spelen, als de wedstrijd op een hooglandbaan wordt geschaatst. Het huidige wereldrecord van Derek Parra staat aangegeven in de figuur, en duidelijk is te zien dat dit een topprestatie was: het is een tijd uit de linkerstaart van de kansverdeling van toptijden.

Hoe groot is nu de kans dat het wereldrecord wordt verbeterd? Laten we aannemen dat er *W* wedstrijden per jaar zijn die onder dezelfde optimale condities worden georganiseerd als de Olympische Spelen van 2002. Dit zijn bijvoorbeeld speciale recordwedstrijden, die tijdens slecht weer op hoogte worden gehouden. De kans dat de winnende tijd sneller is dan het huidige wereldrecord is 0.0106. Als we verder aannemen dat er *K* rijders zijn die mogelijk een nieuw wereldrecord zouden kunnen rijden is de kans dat het wereldrecord dan wordt gebroken $1 - (1 - 0.0106)^{W \cdot K}$. Redelijke waarden voor *W* en *K* zijn 2 en 4, en dan is de kans op een verbetering ongeveer 0.082. Uiteraard is deze kans groter als de kansverdeling van winnende tijden verder opschuift naar links, als gevolg van nóg betere pakken, schaatsen, enzovoort.
In elk geval geeft de statistische aanpak in deze paragraaf aan dat de grens van wereldrecords op de 1500 m nog niet is bereikt: de verdeling van winnende tijden is niet constant in de loop van de tijd, en de sombere conclusie dat het aantal records slechts logaritmisch met het aantal pogingen toeneemt, gaat niet op.

Conclusies

In dit artikel hebben we met behulp van statistiek beargumenteerd dat de grens van de mogelijkheden op de 1500 m schaatsen nog niet in zicht is. Van jaar tot jaar verbeteren de schaatsers zich, en onder een ideale constellatie van factoren is het zeker mogelijk om het huidige record scherper te stellen. De techniek die gebruikt is, is extreme waarde theorie. Dit is een van de moderne stukken wiskundig gereedschap dat wordt gebruikt bij met name de analyse van financiële gegevens. Modellen waar

de variabiliteit van financiële gegevens wordt gemodelleerd, worden ook meer en meer gebruikt bij het management van banken, verzekeraars en pensioenfondsen.

Ruud Koning, Vakgroep Econometrie, Fac. der Econ. Wetenschappen, Postbus 800, 9700 AV Groningen.
email: r.h.koning@eco.rug.nl; www.rhkoning.com

Ik bedank Eliena van der Velde voor commentaar en Jeroen Heijmans voor het beschikbaar stellen van de gegevens.

Literatuur

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Heidelberg: Springer.

Haan, L. de (1990). Fighting the arch-enemy with mathematics. *Statistica neerlandica 44*, 45-68.

Glick, N. (1978). Breaking records and breaking boards. *American mathematical monthly, 85*(1), 2-26.

Noten

[1] Een toegankelijke inleiding tot deze theorie is Coles, 2001.

[2] De gegevens zijn beschikbaar gesteld door Jeroen Heijmans, zie: http://weasel.student.utwente.nl/~speedskating/

Appendix

Stelling 2 *Laat $T_1, \ldots, T_n$ n onderling onafhankelijke toevalsvariabelen zijn, die allemaal dezelfde kansverdeling hebben. De verwachting en variantie van het aantal records R_n in deze reeks zijn:*

$$\mathrm{E}\, R_n = \sum_{i=1}^{n} \frac{1}{i} \qquad (6)$$

$$\mathrm{var}\, R_n = \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i^2}\right) \qquad (7)$$

Het bewijs van deze stelling is verrassend eenvoudig. Laat Y_i aangeven of de laatste waarneming de grootste is:

$$Y_i = \begin{cases} 1 & \text{"als} X_i = max(X_1, \ldots, X_i) \\ 0 & \text{"anders} \end{cases}$$

Y_i geeft dus aan of de laatste waarneming een recordverbetering is. We weten natuurlijk niet welke van de i waarnemingen de grootste is, dus de kans dat de laatste dat is, is $1/i$. De verwachting van Y_i is dan:

$$\mathrm{E}\, Y_i = 1 \times \Pr(Y_i = 1) + 0 \times \Pr(Y_i = 0) = 1 \times \frac{1}{i} = \frac{1}{i}.$$

Het aantal records in een reeks van n waarnemingen is gelijk aan de som van het aantal recordverbeteringen:

$$R_n = \sum_{i=1}^{n} Y_i$$

en dus is het verwachte aantal recordverbeteringen in die reeks:

$$\mathrm{E}\, R_n = \sum_{i=1}^{n} EY_i = \sum_{i=1}^{n} \frac{1}{i}.$$

Verder zijn Y_i en Y_j ongecorreleerd (neem $i < j$). Dat betekent dat $\mathrm{E}(Y_i\, Y_j) = \mathrm{E}(Y_i)\, \mathrm{E}(Y_j)$. Dit volgt uit:

$$\begin{aligned}
&\mathrm{E}(Y_i\, Y_j) = \Pr(Y_i = 1 \text{ en } Y_j = 1) \\
&= \Pr(X_i = \max(X_1, \ldots, X_i) \text{ en } X_j = \max(X_1, \ldots, X_j)) \\
&= \Pr(X_i = \max(X_1, \ldots, X_i) < \max(X_{i+1}, \ldots, X_j) = X_j) \\
&= \Pr(X_i = \max(X_1, \ldots, X_i)) \\
&\quad \times \Pr(\max(X_1, \ldots, X_i) < \max(X_{i+1}, \ldots, X_j)) \\
&\quad \times \Pr(X_j = \max(X_{i+1}, \ldots, X_j)) \\
&= \frac{1}{i} \times \frac{j-i}{j} \times \frac{1}{j-i} = \frac{1}{i} \times \frac{1}{j} \\
&= \Pr(Y_i = 1)\, \Pr(Y_j = 1) = \mathrm{E}(Y_i)\, \mathrm{E}(Y_j).
\end{aligned}$$

De variantie van Y_i is $\mathrm{var} Y_i = \mathrm{E}\, Y_i^2 - (\mathrm{E}\, Y_i)^2 = \frac{1}{i} - \frac{1}{i^2}$.

Omdat Y_i en Y_j ongecorreleerd zijn, is de variantie van de som gelijk aan de som van de varianties, en is de variantie van R_n:

$$\mathrm{var}\, R_n = \sum_{i=1}^{n} \mathrm{var}\, Y_i = \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i^2}\right).$$

De kansverdeling van R_n is lastig te bepalen en hangt af van de kansverdeling van X. Uiteraard kunnen we aannemen dat R_n bij benadering een normale verdeling volgt met bovenstaande verwachting en variantie. Het is niet moeilijk om met een simulatiestudie na te gaan wat de kwaliteit van die benadering is.