De Wet van Grote Aantallen

In hoofdstuk 5 hebben wij de discrete begrippen uit hoofdstukken 1 en 2 (met aftelbare uitkomstenruimten) gegeneraliseerd naar de uitkomstenruimte $\mathbb{R}^d$. Hiervoor moesten we hoofdzakelijk bedenken hoe we het $\sum$-teken konden vervangen door de $\int$-teken. Bovendien is het een extra eis aan een deelverzameling $A \subseteq \mathbb{R}^d$ dat de integraal

$$\int_A f(x)\mathrm{d}x \tag{1}$$

bestaat. In het discrete geval is iedere deelverzameling $A \subseteq \Omega$ een gebeurtenis — we kunnen immers de kans

$$P(A) \quad = \quad \sum_{x \in A} p(x)$$

zonder moeite berekenen — maar om soortgelijk een kans $P(A)$ aan $A \subseteq \mathbb{R}^d$ toe te kennen moet de integraal (1) wel bestaan, en alleen dan noemen we A een gebeurtenis. Meer details hierover in het college 'maat en integratie'.

Hoofdstuk 3 ging over de stochastische wandeling. Ook hierover kunnen we d.m.v. de uitbreiding naar continue uitkomstenruimten meer uitspraken doen, zie § 6.4.

De inhoud van hoofdstuk 4 was limietstellingen. De centrale limietstelling komen we in hoofdstuk 8 weer tegen, op dit moment beperken we ons tot de wet van grote aantallen. **Opgave 5.11.2 :** formuleer en bewijs de stellingen 4.1.1 en 4.1.4 voor continue toevalsvariabelen.

Stelling 5.11.2.A (Zwakke wet van grote aantallen) *Gegeven onafhankelijke en identiek verdeelde (i.i.d.) toevalsvariabelen* $X_1, X_2, \ldots, X_n : \mathbb{R}^d \longrightarrow \mathbb{R}$ *met* $E(X_1) = \mu \in \mathbb{R}$ *en* $\mathrm{var}(X_1) = \sigma^2 \in \mathbb{R}$, *dus allebei eindig. Dan geldt voor* $S_n = X_1 + \ldots X_n$

$$\forall_{\delta>0} \qquad P\left(\left|\frac{S_n}{n} - \mu\right| > \delta\right) \quad \leq \quad \frac{\sigma^2}{n\delta^2} \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0\ .$$

Het bewijs is letterlijk hetzelfde als in het discrete geval, vervang alleen de verwijzing naar gevolg 2.3.23 (de ongelijkheid van Chebyshev) door een verwijzing naar opgave 5.11.1.

Lemma 5.11.2.B *Zij $p > 0$ en $Z : \mathbb{R}^d \longrightarrow [0,\infty[$ een toevalsvariabele die geen negatieve waarden aanneemt. Dan is*

$$E(Z^p) \quad = \quad \int_0^\infty p z^{p-1} P(Z > z)\,\mathrm{d}z\ .$$

Bewijs : We beperken ons tot continue toevalsvariabelen, schrijven $f = f_Z$ voor de kansdichtheid van Z en berekenen eerst

$$P(Z > z) \quad = \quad \int_z^\infty f(\zeta)\,\mathrm{d}\zeta \quad = \quad \int_{-\infty}^\infty f(\zeta)\mathbf{1}_{z\le\zeta}(z)\,\mathrm{d}\zeta$$

(want tussen $-\infty$ en z integreren we de nulfunctie). Dan volgt

$$\begin{aligned}
\int_0^\infty p z^{p-1} P(Z > z)\,\mathrm{d}z \quad &= \quad \int_0^\infty p z^{p-1} \int_{-\infty}^\infty f(\zeta)\mathbf{1}_{z\le\zeta}(z)\,\mathrm{d}\zeta\,\mathrm{d}z \\
&= \quad \int_{-\infty}^\infty f(\zeta) \int_0^\infty p z^{p-1}\mathbf{1}_{z\le\zeta}(z)\,\mathrm{d}z\,\mathrm{d}\zeta \\
&= \quad \int_{-\infty}^\infty f(\zeta) \int_0^\zeta p z^{p-1}\,\mathrm{d}z\,\mathrm{d}\zeta \\
&= \quad \int_{-\infty}^\infty f(\zeta)\zeta^p\,\mathrm{d}\zeta \quad = \quad E(Z^p)\ .
\end{aligned}$$

□

Stelling 5.11.2.C (Algemene zwakke wet van grote aantallen) *Gegeven onafhankelijke en identiek verdeelde (i.i.d.) toevalsvariabelen $X_1, X_2, \ldots : \mathbb{R}^d \longrightarrow \mathbb{R}$ met $E(X_1) = \mu \in \mathbb{R}$, dus eindig. Dan is*

$$\forall_{\delta>0} \qquad P\left(\left|\frac{S_n}{n} - \mu\right| > \delta\right) \ \overset{n\to\infty}{\longrightarrow}\ 0\ .$$

Bewijs : We houden n eerst vast en nemen de limiet pas aan het eind. Dan hebben we te maken met een eindig aantal toevalsvariabelen $X_1, \ldots, X_n$, $S_n = X_1 + \ldots X_n$, en in het geval van eindige variantie kunnen we stelling 5.11.2.A toepassen. Het idee is dan ook om over te stappen naar toevalsvariabelen $Y_1, \ldots, Y_n$ die wél eindige variantie hebben. Dit gebeurt d.m.v. truncatie :

$$Y_k(\omega) \quad := \quad X_k(\omega)\cdot\mathbf{1}_{\{|X_k|\le n\}}(\omega)\,, \qquad \text{voor } k = 1,\ldots,n \tag{2}$$

en we hebben uiteraard

$$\begin{aligned}
\mathrm{var}(Y_k) &= E(Y_k^2) - (E(Y_k))^2 \;\leq\; E(Y_k^2) \\
&\overset{5.11.2.\mathrm{B}}{=} \int_0^\infty 2xP(|Y_k| > x)\,\mathrm{d}x \;=\; \int_0^\infty 2xP(|X_k|\mathbf{1}_{\{|X_k|\leq n\}} > x)\,\mathrm{d}x \\
&= \int_0^n 2xP(|X_k|\mathbf{1}_{\{|X_k|\leq n\}} > x)\,\mathrm{d}x \;\leq\; n^2 \;<\; \infty\,,
\end{aligned}$$

maar we hebben een betere afschatting nodig. Immers, de toevalsvariabelen $Y_1, \ldots, Y_n$ hangen volgens de definitie (2) direct van n af, we schrijven dan ook liever Y_k^n, en als we hun variantie $\mathrm{var}(Y_k^n) = \sigma_n^2$ niet beter kunnen afschatten dan gaat het quotient $\frac{\sigma_n^2}{n\delta^2}$ uit stelling 5.11.2.A naar oneindig i.p.v. naar nul. Daarom kijken we naar de integrand

$$\begin{aligned}
xP(|X_k| > x) &= xE(\mathbf{1}_{\{|X_k|>x\}}) \;=\; E(x\mathbf{1}_{\{|X_k|>x\}}) \;\leq\; E(|X_k|\mathbf{1}_{\{|X_k|>x\}}) \\
&= \int_{-\infty}^\infty y f_{|X_k|\mathbf{1}_{\{|X_k|>x\}}}(y)\,\mathrm{d}y \;=\; \int_x^\infty y f_{|X_k|}(y)\,\mathrm{d}y \qquad (3)\\
&= E(|X_1|) \;-\; \int_{-\infty}^x y f_{|X_k|}(y)\,\mathrm{d}y \;\overset{x\to\infty}{\longrightarrow}\; 0
\end{aligned}$$

(hier gebruiken we aan het eind dat $E(|X_1|)$ eindig is). I.h.b. heeft $xP(|X_k| > x)$ een maximum M. De afschatting $\sigma_n^2 \leq 2Mn$ is al beter, maar nog niet goed genoeg. Het volgende blijkt te zullen voldoen : gegeven $\eta > 0$ bestaat er (vanwege (3)) een $\xi(\eta) \in \mathbb{R}$ met $xP(|X_k| > x) < \eta \;\forall_{x\geq\xi(\eta)}$ en dus

$$\begin{aligned}
\sigma_n^2 &= \int_0^n 2xP(|X_k|\mathbf{1}_{\{|X_k|\leq n\}} > x)\,\mathrm{d}x \;\leq\; 2\int_0^n xP(|X_k| > x)\,\mathrm{d}x \\
&\leq 2\int_0^{\xi(\eta)} xP(|X_k| > x)\,\mathrm{d}x \;+\; 2\int_{\xi(\eta)}^n xP(|X_k| > x)\,\mathrm{d}x \qquad (4)\\
&\leq 2M\xi(\eta) \;+\; 2\eta n\,.
\end{aligned}$$

Om X_k en S_n volledig door Y_k^n en $T_n = Y_1^n + \ldots Y_n^n$ te kunnen vervangen, moeten we ook met $\nu_n = E(Y_k^n)$ werken. Let op dat de truncatie zodanig is gekozen dat $Y_k^n \to X_k$ als $n \to \infty$. Hierdoor is

$$\nu_n \;=\; \int_{-\infty}^n x f_{X_1}(x)\,\mathrm{d}x \;\overset{n\to\infty}{\longrightarrow}\; \int_{-\infty}^\infty x f_{X_1}(x)\,\mathrm{d}x \;=\; \mu\,.$$

Verder geldt

$$\begin{aligned}
P\left(\left|\frac{S_n}{n}-\nu_n\right|>\delta\right) &= P\left(\left|\frac{S_n}{n}-\nu_n\right|>\delta\,,\,S_n=T_n\right) \;+\; P\left(\left|\frac{S_n}{n}-\nu_n\right|>\delta\,,\,S_n\neq T_n\right)\\
&\leq P\left(\left|\frac{T_n}{n}-\nu_n\right|>\delta\right) \;+\; P\left(\left|\frac{S_n}{n}-\nu_n\right|>\delta\,,\,S_n\neq T_n\right)\\
&\leq \frac{\sigma_n^2}{n\delta^2} \;+\; P(S_n\neq T_n)
\end{aligned}$$

met

$$P(S_n\neq T_n) \;\leq\; P(\exists_{k\leq n}\, X_k\neq Y_k^n) \;\leq\; nP(|X_k|>n) \;\xrightarrow{n\to\infty}\; 0$$

vanwege (3) en

$$\frac{\sigma_n^2}{n\delta^2} \;\leq\; \frac{2M\xi(\eta)}{n\delta^2} \;+\; \frac{2\eta}{\delta^2}$$

vanwege (4).

Om te laten zien dat deze laatste som ook daadwerkelijk naar nul gaat, gebruiken we de definitie van de limiet uit 'inleiding analyse'. Gegeven $\varepsilon>0$. Dan kiezen we (nu pas!) $\eta=\frac{1}{6}\delta^2\varepsilon$ en $N_\varepsilon\geq\frac{6M\xi(\eta)}{\delta^2\varepsilon}$. Hiermee volgt voor alle $n\geq N_\varepsilon$

$$\ldots \;\leq\; \frac{2M\xi(\eta)}{\frac{6M\xi(\eta)}{\delta^2\varepsilon}\cdot\delta^2} \;+\; \frac{2\cdot\frac{\delta^2\varepsilon}{6}}{\delta^2} \;=\; \frac{\varepsilon}{3}\;+\;\frac{\varepsilon}{3} \;<\; \varepsilon\,.$$

Voor de kleine lettertjes gebruiken we de 'overgebleven' $\frac{\varepsilon}{3}$ en vervangen in bovenstaande redenering δ door $\tilde{\delta}=2\delta$, en kiezen

$$N_\varepsilon \;\geq\; \max\left\{\frac{6M\xi(\eta)}{\tilde{\delta}^2\varepsilon}\,,\,N_\delta\,,\,N_{\frac{\varepsilon}{3}}\right\}\,.$$

Hier is $N_{\frac{\varepsilon}{3}}$ zodanig gekozen dat

$$\forall_{n\geq N_{\frac{\varepsilon}{3}}} \qquad nP(|X_k|>n) \;<\; \frac{\varepsilon}{3}$$

en dus

$$P\left(\left|\frac{S_n}{n}-\nu_n\right|>\tilde{\delta}\right) \;<\; \varepsilon\,.$$

Verder is N_δ zodanig gekozen dat

$$\forall_{n\geq N_\delta} \qquad |\,\nu_n-\mu\,| \;<\; \delta$$

en dus

$$\left|\frac{S_n}{n}\;-\;\nu_n\right| \;>\; \tilde{\delta} \quad\Rightarrow\quad \left|\frac{S_n}{n}\;-\;\mu\right| \;>\; \delta$$

waarmee uiteindelijk

$$\forall_{n\geq N_\varepsilon} \qquad P\left(\left|\frac{S_n}{n}-\mu\right|>\delta\right) \;\leq\; \varepsilon\,. \qquad \square$$