Kapitel 7

Grænseresultater for stokastiske variable

I dette kapitel behandles to af sandsynlighedsregningens vigtigste resultater, som begge drejer sig om, hvordan gennemsnittet af n stokastiske variable opfører sig, når n går mod uendelig.

7.1 De store tals lov

Store tals lov siger, at for n ukorrelerede stokastiske variable med middelværdi og varians vil gennemsnittet med stor sandsynlighed ligge nær middelværdien, når n er stor. Dette resultat vises ved hjælp af Chebychevs ulighed.

Sætning 7.1.1 *(Chebychevs ulighed). Lad X være en diskret eller kontinuert stokastisk variabel med middelværdi μ og varians σ^2. Da gælder, for ethvert $a > 0$, at*

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}. \tag{7.1.1}$$

Bevis: Definer for $a > 0$ mængden

$$A = \{x \in \mathbb{R} | \, |x - \mu| \geq a\}.$$

Da er

$$(x - \mu)^2 1_A(x) \geq a^2 1_A(x)$$

for alle $x \in \mathbb{R}$. Dermed er

$$\begin{aligned}\sigma^2 &= \mathrm{E}\left((X-\mu)^2\right) \\ &= \mathrm{E}\left((X-\mu)^2 1_A(X)\right) + \mathrm{E}\left((X-\mu)^2 1_{\mathbb{R}\setminus A}(X)\right) \\ &\geq \mathrm{E}\left((X-\mu)^2 1_A(X)\right) \\ &\geq \mathrm{E}\left(a^2 1_A(X)\right) = a^2 P(X \in A) = a^2 P(|X-\mu| \geq a).\end{aligned}$$

Det andet ulighedstegn følger af (5.2.6)

□

Eksempel 7.1.2 Hvis X har middelværdi 0 og varians 1 er $P(|X| \geq 10) \leq 0.01$.

□

Vi kan nu uden større besvær bevise de store tals lov.

Sætning 7.1.3 *(De store tals lov). Lad $X_1, X_2, \ldots$ være en følge af ukorrelerede diskrete eller kontinuerte stokastiske variable, som alle har samme middelværdi og varians. Hvis $\mu = E(X_i)$, gælder for ethvert $\epsilon > 0$ at*

$$P\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| < \epsilon\right) \to 1 \tag{7.1.2}$$

for $n \to \infty$.

Bevis: Gennemsnittet $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ har middelværdi μ og varians

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right) = \frac{\sigma^2}{n},$$

hvor $\sigma^2 = \mathrm{Var}(X_i)$. Ved at anvende Chebychevs ulighed (7.1.1) fås

$$\begin{aligned}1 &\geq P\left(|\bar{X}_n - \mu| < \epsilon\right) = 1 - P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \\ &\geq 1 - \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = 1 - \frac{\sigma^2}{n\epsilon^2} \to 1,\end{aligned}$$

for $n \to \infty$.

□

Den eneste grund til at det i Sætningerne 7.1.1 og 7.1.3 kræves, at de stokastiske variable er diskrete eller kontinuerte, er at vi kun har defineret middelværdi og varians for disse to typer af fordelinger. Chebychevs ulighed og store tals lov gælder for alle typer stokastisk variabel, som har middelværdi og varians.

Sætning 7.1.3 er den enkleste version af de store tals lov. Der findes andre varianter, som vi ikke vil bevise her (se dog Opgave 7.7). F. eks. er det ikke nødvendigt at antage, at de stokastiske variable har varians. Det er nok, at $\mathrm{E}(|X_i|) < \infty$, hvilket til gengæld klart nok er nødvendigt. I den situation antages uafhængighed i stedet for ukorrelerethed, men de stokastiske variable behøver faktisk ikke være ukorrelerede eller uafhængige; resultatet (7.1.2) gælder, hvis blot afhængigheden er tilstrækkeligt svag.

Vi har nu bevist, at *sandsynlighedsregningens frekvensfortolkning* holder. Lad nemlig $Y_1, Y_2, \ldots$ være en følge af uafhængige stokastiske variable med samme fordeling, og lad A være en delmængde af $\mathbb{R}$. Sæt $p = P(Y_i \in A)$. Da er $\mathrm{E}(1_A(Y_i)) = p$ og

$$H_n = \frac{1_A(Y_1) + \cdots + 1_A(Y_n)}{n}$$

er den relative hyppighed af udfald i A blandt de n første stokastiske variable. Da $1_A(Y_i)$ er begrænset, har den varians, så ifølge store tals lov gælder

$$P(|H_n - p| < \epsilon) \to 1$$

for $n \to \infty$ for ethvert $\epsilon > 0$.

Det skal også lige nævnes, at vi her har stiftet bekendtskab med et af sandsynlighedsregningens vigtigste begreber, *konvergens i sandsynlighed.* Hvis $Y_1, Y_2, \ldots$ er en følge af stokastiske variable, som opfylder, at der findes en stokastisk variabel Y, så $P(|Y_n - Y| < \epsilon) \to 1$ for $n \to \infty$ for ethvert $\epsilon > 0$, så siger man, at følgen $\{Y_n\}$ konvergerer i sandsynlighed mod Y. Store tals lov siger altså, at gennemsnittet af n uafhængige stokastiske variable med samme middelværdi og varians konvergerer i sandsynlighed mod den fælles middelværdi. Her er Y lig med en konstant, nemlig middelværdien.

7.2 Den centrale grænseværdisætning

Betragt en følge af uafhængige diskrete eller kontinuerte stokastiske variable $X_1, X_2, \ldots$, som alle har samme fordeling med middelværdi μ og varians σ^2. De store tals lov fortæller os da, at gennemsnittet af de n første stokastiske variable

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

er tæt på μ, når n er stor. Gennemsnittet er jo en stokastisk variabel, så det er naturligt at spørge, hvad dets fordeling er, når n er stor. Variansen er meget lille, så det varierer kun lidt i nærheden af μ. Imidlertid kan man studere gennemsnittets stokastiske variation ved at betragte den standardiserede variabel

$$U_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}. \tag{7.2.1}$$

Da $\mathrm{E}(\bar{X}_n) = \mu$ og $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$, har U_n middelværdi nul og varians en. Vi har ved at gange med $\sqrt{n}/\sigma$ forstørret den stokastiske variation af $\bar{X}_n$, så vi kan studere den nøjere. Der gælder følgende resultat.

Sætning 7.2.1 *(Den centrale grænseværdisætning). Lad $X_1, X_2, \ldots$ være en følge af uafhængige, identisk fordelte kontinuerte eller diskrete stokastiske variable med middelværdi μ og varians σ^2. Da vil fordelingen af U_n, defineret ved (7.2.1), konvergere mod en standard normalfordeling, i den forstand at for alle $u \in \mathbb{R}$ vil*

$$P(U_n \leq u) \to \Phi(u), \tag{7.2.2}$$

når $n \to \infty$. Som sædvanlig betegner Φ fordelingsfunktionen for standard normalfordelingen givet ved (5.3.2).

Vi har på dette kursus ikke de nødvendige tekniske hjælpemidler til at bevise den centrale grænseværdisætning. Med de rette hjælpemidler er sætningen faktisk ikke svær at vise.

Der er mange gode grunde til, at normalfordelingen dukker op her. Hvis X_i-erne er normalfordelte, følger det af Sætning 6.3.12, at U_n er standard normalfordelt for alle n. Hvis sætningen skal være sand, må

$P(U_n \leq u)$ derfor nødvendigvis konvergere mod $\Phi(u)$. Bemærk, at vi kan slutte af Sætning 7.2.1, at standard normalfordelingen er den eneste fordeling med varians, som har den egenskab, at U_n har samme fordeling som X_i-erne for alle n. Det kan faktisk bevises, at den er den eneste fordeling overhovedet med denne egenskab.

Hvis vi definerer $S_n = X_1 + \cdots + X_n$, kan resultatet (7.2.2) også skrives på formen

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq u\right) \to \Phi(u), \tag{7.2.3}$$

når $n \to \infty$. Vi kan altså også opfatte den centrale grænseværdisætning som et udsagn om fordelingen af summen S_n, når n er stor.

Den centrale grænseværdisætning giver en forklaring på, hvorfor det i praksis så ofte viser sig, at observationer kan antages at være normalfordelte. Vi kan f. eks. tænke os, at man skal måle en størrelse, som har værdien ξ, men at en række tilfældige forhold indvirker på målingen, så den ikke bliver helt nøjagtig. Det, der faktisk måles, kan derfor antages at være $Y = \xi + X_1 + \cdots + X_n$, hvor X_i-erne er virkningen af de forskellige kilder til målefejl. Hvis X_i-erne opfylder betingelserne i den centrale grænseværdisætning, kan fordelingen af målingen Y tilnærmes med en normalfordeling med middelværdi $\xi + n\mu$ og varians $n\sigma^2$. Ved gode målinger er μ lig nul eller i hvert fald meget lille.

Der findes andre udgaver af den centrale grænseværdisætning med væsentligt svagere betingelser, som gør overstående argument for, at målefejl ofte er normalfordelte, mere overbevisende. De stokastiske variable X_i behøver således ikke at have samme fordeling; ikke engang samme middelværdi og varians. De behøver heller ikke at være uafhængige. Det væsentlige er, at de alle er små i forhold til summen S_n, og at afhængigheden mellem dem ikke er for stor.

Den form for konvergens, som udtrykkes ved (7.2.2) kaldes *konvergens i fordeling*. Et andet eksempel på denne form for konvergens så vi i Sætning 4.1.2. Udforskningen af konvergens i fordeling spiller en central rolle i sandsynlighedsregningen og dens anvendelser i statistik.