week9: Der zentrale Grenzwertsatz

In den letzten beiden Wochen hatten wir in dem week7.pdf und dem week8.pdf die Monte Carlo Summen

$$\frac{1}{n}\sum_{i=1}^{n} F(x_i)$$

betrachtet. Dabei waren die $x_1, ..., x_n$ Zufallszahlen, die alle dieselbe Wahrscheinlichkeitsverteilung $p(x)$ hatten, so dass der Erwartungswert für alle $F(x_i)$ derselbe war, nämlich

$$\mathsf{E}\big[\, F(x_i)\,\big] \;=\; \int_{\mathbb{R}} F(x_i)\, p(x_i)\, dx_i \;=\; \int_{\mathbb{R}} F(x)\, p(x)\, dx \;=\; \mathsf{E}[F]$$

Die Varianz war ebenfalls für alle $F(x_i)$ dieselbe, das ist dann der Ausdruck

$$\mathsf{V}\big[\, F(x_i)\,\big] \;=\; \mathsf{E}\big[\, [F(x_i)]^2\,\big] \;-\; \big(\,\mathsf{E}\big[\, F(x_i)\,\big]\,\big)^2 \;=\; \mathsf{E}[F^2\,] \;-\; (\,\mathsf{E}[F]\,)^2 \;=\; \mathsf{V}[F]$$

mit

$$\mathsf{E}\big[\, [F(x_i)]^2\,\big] \;=\; \int_{\mathbb{R}} [F(x_i)]^2\, p(x_i)\, dx_i \;=\; \int_{\mathbb{R}} [F(x)]^2\, p(x)\, dx \;=\; \mathsf{E}[F^2]$$

Zur Formulierung des zentralen Grenzwertsatzes wollen wir jetzt etwas kompakter schreiben

$$F(x_i) \;=:\; X_i$$

mit

$$\begin{aligned} \mathsf{E}[X_i] \;&=\; \mathsf{E}[F] \;=:\; \mu \\ \mathsf{V}[X_i] \;&=\; \mathsf{V}[F] \;=:\; \sigma^2 \end{aligned}$$

Wir betrachten die Summen und die Mittelwerte (in week7 und 8 hatten wir immer von Monte Carlo Summe gesprochen, in der Terminologie, die wir jetzt benutzen wollen, wären das also genauer Monte Carlo Mittelwerte, die S_n aus week7 und 8 sind hier jetzt M_n's)

$$\begin{aligned} S_n \;&:=\; \sum_{i=1}^{n} X_i \\ M_n \;&:=\; \frac{1}{n}\sum_{i=1}^{n} X_i \;=\; \frac{S_n}{n} \end{aligned}$$

Wenn man stochastische Grössen hat, macht es Sinn, diese zu standardisieren: man zieht den Erwartungswert ab und teilt durch die Standardabweichung, das ist die Wurzel aus der Varianz. Wir wollen also die standardisierten Grössen

$$Z_n \;:=\; \frac{S_n \,-\, \mathsf{E}[\,S_n\,]}{\sqrt{\mathsf{V}[S_n]}} \tag{1}$$

$$\tilde{Z}_n \;:=\; \frac{M_n \,-\, \mathsf{E}[\,M_n\,]}{\sqrt{\mathsf{V}[M_n]}} \tag{2}$$

betrachten. Dann gilt zunächst mal das folgende

Lemma 9.1: Es seien $X_1, ..., X_n$ unabhängige und identisch verteilte Zufallsgrössen mit

$$\begin{aligned} \mathsf{E}[X_i] &= \mu \\ \mathsf{V}[X_i] &= \sigma^2 \end{aligned}$$

für alle $i = 1, ..., n$. Dann gilt

a)

$$\begin{aligned} \mathsf{E}[S_n] &= n\,\mu \\ \mathsf{E}[M_n] &= \mu \end{aligned}$$

und

$$\begin{aligned} \mathsf{V}[S_n] &= n\,\sigma^2 \\ \mathsf{V}[M_n] &= \frac{\sigma^2}{n} \end{aligned}$$

b) Die standardisierten Grössen Z_n und $\tilde{Z}_n$ aus (1) und (2) sind identisch, es gilt

$$Z_n \;=\; \tilde{Z}_n \;=\; \frac{1}{\sqrt{n}} \left\{ \frac{X_1 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma} \right\}$$

Beweis: a) Das haben wir im wesentlichen im week8.pdf in den Gleichungen (8) und (9) schon gemacht. Schreiben wir die Sachen eben nochmal hin, wir brauchen nur, dass der Erwartungswert und die Kovarianz linear sind:

$$\begin{aligned} \mathsf{E}[S_n] &= \mathsf{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathsf{E}[X_i] = \sum_{i=1}^{n} \mu = n\,\mu \\ \mathsf{E}[M_n] &= \mathsf{E}\Big[\frac{S_n}{n}\Big] = \frac{1}{n}\,\mathsf{E}[S_n] = \frac{1}{n}\,n\,\mu = \mu \end{aligned}$$

Und für die Varianzen:

$$\begin{aligned} \mathsf{V}[S_n] &= \mathsf{V}\Big[\sum_{k=1}^{n} X_k\Big] = \mathsf{Cov}\Big[\sum_{k=1}^{n} X_k\,,\,\sum_{\ell=1}^{n} X_\ell\Big] = \sum_{k,\ell=1}^{n} \mathsf{Cov}\big[\,X_k\,,\,X_\ell\,\big] \\ &\overset{\substack{X_k,\,X_\ell \\ \text{unabhängig}}}{=} \sum_{k=1}^{n} \mathsf{Cov}\big[\,X_k\,,\,X_k\,\big] = \sum_{k=1}^{n} \mathsf{V}\big[\,X_k\,\big] = \sum_{k=1}^{n} \sigma^2 = n\,\sigma^2 \end{aligned}$$

und damit

$$\mathsf{V}[M_n] \;=\; \mathsf{V}\Big[\frac{S_n}{n}\Big] \;=\; \frac{1}{n^2}\,\mathsf{V}[S_n] \;=\; \frac{1}{n^2}\,n\,\sigma^2 \;=\; \frac{\sigma^2}{n}$$

b) Wir haben

$$\begin{aligned} Z_n &= \frac{S_n - \mathsf{E}[S_n]}{\sqrt{\mathsf{V}[S_n]}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\frac{1}{n}S_n - \mu}{\frac{1}{n}\sqrt{n\sigma^2}} = \frac{\frac{1}{n}S_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \\ &= \frac{M_n - \mathsf{E}[M_n]}{\sqrt{\mathsf{V}[M_n]}} = \tilde{Z}_n \,, \end{aligned}$$

die beiden Grössen sind also identisch, wir können beides mit Z_n bezeichnen und brauchen kein $\tilde{Z}_n$. Wir können das Z_n dann auch folgendermassen schreiben:

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}} \frac{X_1 + \cdots + X_n - n\mu}{\sigma} = \frac{1}{\sqrt{n}} \left\{ \frac{X_1 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma} \right\}$$

und das Lemma ist bewiesen. ■

Definieren wir die Zufallsgrössen

$$Y_i := \frac{X_i - \mu}{\sigma} = \frac{X_i - \mathsf{E}[X_i]}{\sqrt{\mathsf{V}[X_i]}}$$

dann gilt

$$\mathsf{E}[\,Y_i\,] = \mathsf{E}\Big[\,\frac{X_i - \mu}{\sigma}\,\Big] = \frac{E[X_i] - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

und

$$\mathsf{V}[\,Y_i\,] = \mathsf{V}\Big[\,\frac{X_i - \mu}{\sigma}\,\Big] = \frac{1}{\sigma^2}\,\mathsf{V}\big[\,X_i - \mu\,\big] = \frac{1}{\sigma^2}\,\mathsf{V}[X_i] = \frac{1}{\sigma^2}\,\sigma^2 = 1$$

Die Y_i sind also standardisiert, sie haben Mittelwert 0 und Standardabweichung 1. Mit Hilfe der Y_i können wir das Z_n dann auch einfach schreiben als

$$Z_n = \frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \tag{3}$$

Wären die Y_i jetzt alle normalverteilt, dann können wir exakt vorhersagen, welche Verteilung (also wie die Histogramme aussehen wenn man viele von den Z_n erzeugt) die Z_n hätten, sie wären dann nämlich ebenfalls wieder normalverteilt. Das ergibt sich aus dem folgenden allgemeinen Theorem, welches wir auf dem neuen Übungsblatt 9 durch eine geignete R-Simulation überprüfen:

Theorem 9.1 (Summen von normalverteilten Zufallszahlen): Es seien $\phi_1\,,\,\cdots\,,\,\phi_n$ normalverteilte Zufallsvariablen mit Mittelwerten $\mu_1\,,\,\cdots\,,\,\mu_n$ und Standardabweichungen $\sigma_1\,,\,\cdots\,,\,\sigma_n$ Dann gilt: Die Summe

$$\phi := \phi_1 + \phi_2 + \cdots + \phi_n \tag{4}$$

ist normalverteilt mit Mittelwert

$$\mu = \mu_1 + \mu_2 + \cdots + \mu_n \tag{5}$$

und Varianz

$$\sigma^2 \;=\; \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2 \tag{6}$$

Das heisst genauer, für eine beliebige Funktion $f : \mathbb{R} \to \mathbb{R}$ gilt

$$\int_{\mathbb{R}^n} f(\phi_1 + \cdots + \phi_n) \prod_{k=1}^{n} e^{-\frac{(\phi_k - \mu_k)^2}{2\sigma_k^2}} \frac{d\phi_k}{\sqrt{2\pi\sigma_k^2}} \;=\; \int_{\mathbb{R}} f(\phi)\, e^{-\frac{(\phi-\mu)^2}{2\sigma^2}} \frac{d\phi}{\sqrt{2\pi\sigma^2}} \tag{7}$$

mit μ und σ^2 gegeben durch (5) und (6).

Wählen wir für das f dann etwa die Indikator-Funktion

$$f(\phi) \;:=\; \chi\big(\, x \,\le\, \phi \,\le\, x + dx \,\big) \;:=\; \begin{cases} 1 & \text{falls } \; x \,\le\, \phi \,\le\, x + dx \\ 0 & \text{sonst} \end{cases}$$

dann ergibt sich aus (7), wenn wir das dx sehr klein wählen,

$$\begin{aligned} &\mathsf{Prob}\Big[\; \phi_1 + \cdots + \phi_n \,\in\, [x, x+dx] \;\Big] \\ &\quad = \quad \int_{\mathbb{R}^n} \chi\Big(\, x \le \phi_1 + \cdots + \phi_n \le x + dx \,\Big) \prod_{k=1}^{n} e^{-\frac{(\phi_k - \mu_k)^2}{2\sigma_k^2}} \frac{d\phi_k}{\sqrt{2\pi\sigma_k^2}} \\ &\quad \overset{\text{Thm. 9.1}}{=} \quad \int_{\mathbb{R}} \chi\big(\, x \,\le\, \phi \,\le\, x + dx \,\big)\, e^{-\frac{(\phi-\mu)^2}{2\sigma^2}} \frac{d\phi}{\sqrt{2\pi\sigma^2}} \\ &\quad \overset{dx \text{ klein}}{\approx} \quad e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{dx}{\sqrt{2\pi\sigma^2}} \end{aligned}$$

die Summe $\phi_1 + \cdots + \phi_n$ ist also wieder normalverteilt mit Mittelwert μ und Standardabweichung σ gegeben durch die Formeln (5) und (6).

Kehren wir zu den Z_n aus (1), (2) und (3) zurück,

$$\begin{aligned} Z_n \;&=\; \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \;=\; \frac{M_n - \mu}{\sqrt{\sigma^2/n}} \\ &=\; \frac{1}{\sqrt{n}} \left\{ \frac{X_1 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma} \right\} \;=\; \frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \end{aligned} \tag{8}$$

Wären die X_i nun alle normalverteilt, dann wären die Y_i alle standard-normalverteilt und nach dem Theorem 9.1 wäre dann auch das Z_n normalverteilt mit Varianz

$$\mathsf{V}[Z_n] \;=\; \big(\frac{1}{\sqrt{n}}\big)^2\, \mathsf{V}\big[\, Y_1 + \cdots + Y_n \,\big] \;=\; \frac{1}{n}\, \big(\, \underbrace{\sigma_1^2}_{=1} + \cdots + \underbrace{\sigma_n^2}_{=1} \,\big) \;=\; \frac{1}{n}\, n \;=\; 1$$

die Z_n wären also standard-normalverteilte Zufallszahlen. Wenn die X_i nun eine beliebige Verteilung haben, etwa alle X_i sind gleichverteilt auf dem Intervall $[0, 1]$, dann ist klar, dass

die Z_n nicht normalverteilt sein können, weil die S_n's dann zum Beispiel immer im endlichen Intervall $[0, n]$ liegen und die Z_n's dann auch immer in einem endlichen Intervall liegen müssen. Bei normalverteilten Zahlen kommt da, zumindest theoretisch, aber immer ganz $\mathbb{R}$ in Frage.

Also die exakte Verteilung der Z_n kann irgendetwas kompliziertes sein, was sich möglicherweise auch nur sehr schwer oder überhaupt nicht explizit berechnen lassen tut. Nun ist es aber so, dass für grössere Werte von n (n gleich 20 oder 50 reicht da gelegentlich schon aus, werden wir uns dann konkret in der nächsten Veranstaltung in R anschauen) sich die Verteilung der Z_n durch die Standard-Normalverteilung approximieren lässt, das ist jetzt die Aussage des zentralen Grenzwertsatzes:

Theorem 9.2 (Zentraler Grenzwertsatz): Es seien $X_1, \cdots, X_n$ eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit Erwartungswert μ und Standardabweichung σ und die normierten oder standardisierten Zufallsgrössen Z_n seien gegeben durch die Gleichung (8). Dann gilt: Im Limes $n \to \infty$ sind die Z_n standard-normalverteilt. Das heisst genauer, für ein beliebiges $f : \mathbb{R} \to \mathbb{R}$ gilt:

$$\lim_{n\to\infty} \mathsf{E}\big[\, f(Z_n) \,\big] \;=\; \int_{\mathbb{R}} f(z)\; e^{-\frac{z^2}{2}}\; \frac{dz}{\sqrt{2\pi}} \tag{9}$$

Beweis: Werden wir in der nächsten Veranstaltung mit einer R-Simulation verifizieren. ■

Wählen wir für das f wieder die Indikator-Funktion

$$f(z) \;:=\; \chi\big(\, x \,\leq\, z \,\leq\, x + dx \,\big)$$

dann ergibt sich aus (9)

$$\lim_{n\to\infty} \mathsf{Prob}\Big[\; Z_n \,\in\, [x, x + dx] \;\Big] \;=\; e^{-\frac{x^2}{2}}\; \frac{dx}{\sqrt{2\pi}}$$

Im Limes $n \to \infty$ sind die Z_n also standard-normalverteilt. Wir notieren uns noch die folgende

Folgerung 9.1: Mit $X_i = F(x_i)$ und

$$M_n \;=\; \frac{1}{n} \sum_{i=1}^{n} F(x_i)$$

und

$$Z_n \;=\; \frac{M_n \,-\, \mu}{\sqrt{\sigma^2/n}} \;=\; \frac{M_n \,-\, \mathsf{E}[F]}{\sqrt{\mathsf{V}[F]/n}}$$

können wir auch schreiben:

$$M_n \;=\; \mathsf{E}[F] \,+\, \sqrt{\frac{\mathsf{V}[F]}{n}}\; Z_n \tag{10}$$

Da standard-normalverteilte Zufallszahlen typischerweise so zwischen -5 und +5 liegen, oder, machen wir es etwas genauer (wir schauen uns das auch nochmal in Aufgabe 3 auf dem Übungsblatt 9 an),

$$\lim_{n\to\infty} \mathsf{Prob}\Big[\ Z_n \in [-5,+5]\ \Big] \quad = \quad \int_{-5}^{+5} e^{-\frac{x^2}{2}}\ \frac{dx}{\sqrt{2\pi}} \quad = \quad \Phi(+5) - \Phi(-5) \quad \approx \quad 0.9999994$$

lässt sich aus der Gleichung (10) also ablesen, dass der Monte Carlo Fehler typischerweise von der Grössenordnung

$$\text{Monte Carlo Fehler} \quad = \quad O\Big(\sqrt{\frac{\mathsf{V}[F]}{n}}\Big)$$

ist.