Erinnerung an Grundbegriffe aus der Stochastik

2

Ziel der Statistik ist die Beschreibung und Untersuchung der Variabilität gemachter Beobachtungen. Die Grundidee ist, Variabilität durch mathematische Modelle des Zufalls zu beschreiben. Deren Grundbausteine sind Zufallsvariable. Daher wiederholen wir zunächst zentrale Begriffe aus der Stochastik, wobei Grundlagen wie Zufallsvariable, Verteilungen und das Integral als bekannt vorausgesetzt werden. Eine wahrscheinlichkeitstheoretische Behandlung ist zum Beispiel Georgii (2009) oder Klenke (2013) zu entnehmen. In Abschn. 2.1 wiederholen wir einige Kenngrößen von Zufallsvariablen wie etwa ihren Erwartungswert oder ihre Verteilungsfunktion. Da wir in der Statistik oft untersuchen, wie sich gewisse Größen verhalten, wenn wir mehr und mehr Beobachtungen heranziehen, widmen wir uns in Abschn. 2.2 Folgen von Zufallsvariablen.

2.1 Elementare Begriffe aus der Stochastik

2.1.1 Verteilungen und Quantile

Im Folgenden sei stets $\Gamma \subseteq \mathbb{R}$ und X eine Zufallsvariable mit Bildraum Γ. Die Zufallsvariable X wird beschrieben durch ihre *Verteilung*.

Wir erinnern an zwei Spezialfälle von Verteilungen auf Γ, nämlich *diskrete* und *absolutstetige* Verteilungen. Ist Γ höchstens abzählbar, zum Beispiel $\Gamma = \{0, 1\}$ oder $\Gamma = \mathbb{N}$, so nennen wir die Verteilung ν diskret. In diesem Fall können wir ν beschreiben durch *Gewichte* gegeben durch eine Funktion $g : \Gamma \to [0, 1]$, wobei

$$\nu(B) = \sum_{k \in B} g(k), \qquad \text{für alle } B \subseteq \Gamma. \tag{2.1}$$

M. Messer und G. Schneider, *Statistik*, https://doi.org/10.1007/978-3-662-59339-4_2

Ist Γ ein Intervall, zum Beispiel $\Gamma = \mathbb{R}$, $\Gamma = \mathbb{R}_0^+ := [0, \infty)$ oder $\Gamma = [0, 1]$, und gilt für die Verteilung ν

$$\nu(B) = \int_B f(x)dx := \int_a^b f(x)dx, \qquad \text{für alle } B = [a, b] \subseteq \Gamma \text{ mit } a < b, \tag{2.2}$$

für eine integrierbare Funktion $f : \Gamma \to \mathbb{R}_0^+$, so nennen wir die Verteilung ν absolutstetig und f *Dichte* von ν. Wir erinnern, dass für eine Verteilung $\nu(\Gamma) = 1$ gelten muss.

Beispiel 2.1 (Verteilungen)
Diskrete Verteilungen:

i. *Für $n \in \mathbb{N}\backslash\{0\}$ hat die uniforme Verteilung auf $\Gamma = \{1, \dots, n\}$, kurz $\nu = U\{1, 2, \dots, n\}$, die Gewichte*

$$g(k) = \frac{1}{n} \quad \text{(Abb. 2.1a).}$$

Die Gewichte hängen also nicht von k ab. Analog ist die uniforme Verteilung auf jeder anderen endlichen Menge Γ mit n paarweise verschiedenen Elementen definiert.

ii. *Die Binomialverteilung zu den Parametern $n \in \mathbb{N}\backslash\{0\}$ und $p \in [0, 1]$, kurz $\nu = b(n, p)$, ist definiert auf $\Gamma = \{0, 1, \dots, n\}$ mit Gewichten*

$$g(k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{(Abb. 2.1b).}$$

Für $n = 1$ entspricht die Binomialverteilung der Bernoulli-Verteilung , kurz $ber(p) = b(1, p)$.

iii. *Die Poisson-Verteilung zum Parameter $\lambda > 0$, kurz $\nu = poi(\lambda)$, ist definiert auf $\Gamma = \mathbb{N}$ mit Gewichten*

$$g(k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{(Abb. 2.1c).}$$

Absolutstetige Verteilungen:

iv. *Es sei $a < b$. Die uniforme Verteilung auf dem Intervall $[a, b]$, kurz $\nu = U[a, b]$, ist definiert auf $\Gamma = [a, b]$ mit der Dichte*

$$f(x) = \frac{1}{b - a} \quad \text{(Abb. 2.1d).}$$

Die Dichte hängt also nicht von x ab. Gerne formulieren wir den Raum Γ als $\mathbb{R}$ und setzen dann $f(x) = (1/(b - a))\mathbb{1}_{[a,b]}(x)$.

v. *Die Normalverteilung oder auch Gauß-Verteilung mit Parametern* $\mu \in \mathbb{R}$ *und* $\sigma^2 \in \mathbb{R}^+$, *kurz* $\nu = N(\mu, \sigma^2)$, *ist definiert auf* $\Gamma = \mathbb{R}$ *mit Dichte*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad \text{(Abb. 2.1e)}. \tag{2.3}$$

vi. *Die Gammaverteilung mit Parametern* $\alpha > 0$ *und* $\lambda > 0$, *kurz* $\nu = \gamma(\alpha, \lambda)$, *ist definiert auf dem Bildraum* $\mathbb{R}_0^+$ *mit Dichte*

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x) \quad \text{(Abb. 2.1f)}. \tag{2.4}$$

Dabei bezeichnet $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z}\, \mathrm{d}z$ *die Gammafunktion. Für* $\alpha = 1$ *entspricht die Gammaverteilung der Exponentialverteilung, kurz* $\exp(\lambda) = \gamma(1, \lambda)$. *Definieren wir die Gammaverteilung auf dem Bildraum* $\mathbb{R}$, *so fügen wir der Dichte den Faktor* $\mathbb{1}_{[0,\infty)}(x)$ *an.*

Es sei X eine Zufallsvariable mit Bildraum Γ und ν eine diskrete oder absolutstetige Verteilung auf Γ. Dann ist ν die Verteilung von X, falls für jedes *Ereignis* $\{X \in B\}$ unter sämtlichen Mengen B wie in (2.1) bzw. (2.2) gilt

$$\mathbb{P}(X \in B) := \mathbb{P}(\{X \in B\}) = \nu(B).$$

Wir schreiben dann kurz $X \sim \nu$ und sagen ‚X ist verteilt gemäß ν'. Intuitiv beschreibt also $\nu(B)$ die Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert in der Menge B annimmt. Gilt $\mathbb{P}(X \in B) = 1$, so sagen wir, dass das Ereignis $\{X \in B\}$ *mit Wahrscheinlichkeit* 1 eintritt. Wir nennen eine Zufallsvariable diskret bzw. stetig, wenn ihre Verteilung diskret bzw. absolutstetig ist.

Beispielsweise finden wir, dass für $X \sim U\{1, 2, \ldots, 6\}$ gilt, dass $\mathbb{P}(X \in \{1, 2\}) = 1/6 + 1/6 = 1/3$. Beim Würfeln ist kein Ausgang bevorzugt, und folglich erhalten wir eine Augenzahl kleiner als drei mit Wahrscheinlichkeit 1/3. Oder es gilt zum Beispiel für $X \sim N(\mu, \sigma^2)$, dass $\mathbb{P}(X \leq \mu) = 1/2$, denn die Dichte der Normalverteilung liegt symmetrisch um μ.

Wahrscheinlichkeiten für Ereignisse lassen sich also im Falle diskreter oder absolutstetiger Verteilungen über Summen ihrer Gewichte bzw. Integrale ihrer Dichte ausdrücken. Ohne es explizit zu formulieren, können wir fortan immer an Verteilungen und Zufallsvariable denken, die diskret oder (absolut-)stetig sind. Für den allgemeineren Begriff der Verteilung sei der Leser zum Beispiel an Brokate und Kersting (2011) verwiesen.

Die *Verteilungsfunktion* F einer Verteilung ν auf $\Gamma \subseteq \mathbb{R}$ ist eine Abbildung von $\mathbb{R}$ nach $[0, 1]$ gegeben durch $F(x) := \nu((-\infty, x])$, für $x \in \mathbb{R}$. Ist $X \sim \nu$, so gilt also

$$F(x) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}. \tag{2.5}$$

F ist monoton wachsend, und es gilt $\lim_{x\to-\infty} F(x) = 0$ sowie $\lim_{x\to\infty} F(x) = 1$. Da das Intervall $(-\infty, x]$ rechts abgeschlossen ist, ist F zudem rechtsstetig. Ist ν eine absolutstetige Verteilung, so ist ihre Verteilungsfunktion F stetig. Eine Verteilungsfunktion F ist an einer Stelle x unstetig genau dann, wenn die Verteilung ν in x eine Punktmasse besitzt, d. h., wenn $\nu(\{x\}) > 0$ ist.

Beispiel 2.2 (Verteilungsfunktionen)

i. Die Verteilungsfunktion der Standardnormalverteilung $N(0,1)$ ist gegeben durch die Gauß'sche Fehlerfunktion

$$F(x) = \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz \qquad \text{(Abb. 2.2a)}.$$

ii. Für eine absolutstetige Verteilung mit Dichte f ersetze man den Integranden in i) durch $f(z)$.

iii. Die Verteilungsfunktion F der b(2, 1/2)-Verteilung ist gegeben durch

$$F(x) = \begin{cases} 0 & \textit{für}\, x < 0 \\ 1/4 & \textit{für}\, x \in [0,1) \\ 3/4 & \textit{für}\, x \in [1,2) \\ 1 & \textit{sonst} \end{cases} \qquad \text{(Abb. 2.2b)},$$

denn b(2, 1/2) ist diskret mit Gewichten $g(0) = g(2) = 1/4$ und $g(1) = 1/2$.

Der Begriff des *Quantils* spielt in der Statistik eine zentrale Rolle und kann betrachtet werden als eine Art Umkehrung der Verteilungsfunktion.

Definition 2.3 (Quantil einer Verteilung)
Es sei ν eine Verteilung auf $\Gamma \subseteq \mathbb{R}$ und $p \in (0,1)$. Eine reelle Zahl q_p heißt ein p-Quantil von ν, wenn gilt

$$\nu((-\infty, q_p]) \geq p \qquad \textit{und} \qquad \nu([q_p, \infty)) \geq 1 - p.$$

Ist X eine Zufallsvariable mit $X \sim \nu$, so schreiben sich die Ungleichungen als

$$\mathbb{P}(X \leq q_p) \geq p \qquad \text{und} \qquad \mathbb{P}(X \geq q_p) \geq 1 - p.$$

Interpretation: q_p teilt die Masse der Verteilung ν auf. Mindestens $(p \cdot 100)$ % der Masse liegen links (genauer: nicht rechts) von q_p, und mindestens $((1-p) \cdot 100)$ % der Masse liegen rechts (genauer: nicht links) von q_p.

Beispiel 2.4 (Quantile)

i. *Für* $\nu = N(0,1)$ *ist* $q_{1/2} = 0$. *Dieses 1/2-Quantil ist eindeutig, vgl. Abb. 2.2a. Auch in Abb. 2.1e erkennen wir, dass* $q_{1/2} = \mu$ *die Gaußsche Glockenkurve in zwei gleich große Flächen teilt.*

ii. *Für* $\nu = b(2, 1/2)$ *(Abb. 2.1b, schwarz) ist* $q_{1/2} = 1$ *eindeutig, und für ein 1/4-Quantil* $q_{1/4}$ *gilt* $q_{1/4} \in [0,1]$, *d. h., alle Werte in* $[0,1]$ *sind 1/4-Quantile, vgl. Abb. 2.2b.*

Allgemein gilt für eine Zufallsvariable X mit Verteilungsfunktion F und alle $p \in (0,1)$, dass die Menge Q_p aller p-Quantile ein Intervall $Q_p = [q_p^-, q_p^+]$ bildet mit

$$q_p^- := \sup\{x \in \mathbb{R} | F(x) < p\} \quad \text{und} \quad q_p^+ := \inf\{x \in \mathbb{R} | F(x) > p\}. \tag{2.6}$$

Wenn wir von *dem* (eindeutigen) p-Quantil sprechen, so meinen wir den Mittelwert $\left(q_p^- + q_p^+\right)/2$. Insbesondere nennen wir ein Element aus Q_p einen *Median* der Verteilung,

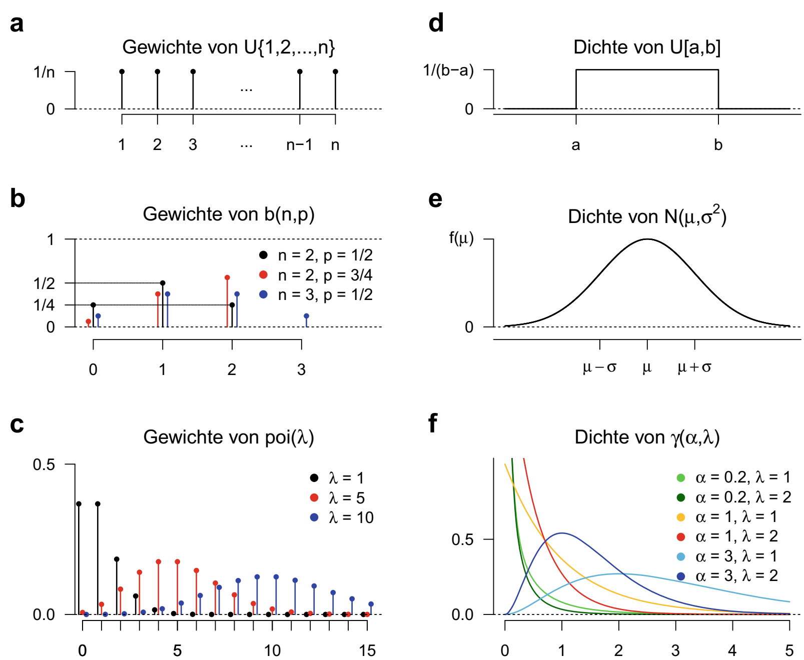

Abb. 2.1 Gewichte (**a–c**) und Dichten (**d–f**) der Verteilungen aus Beispiel 2.1

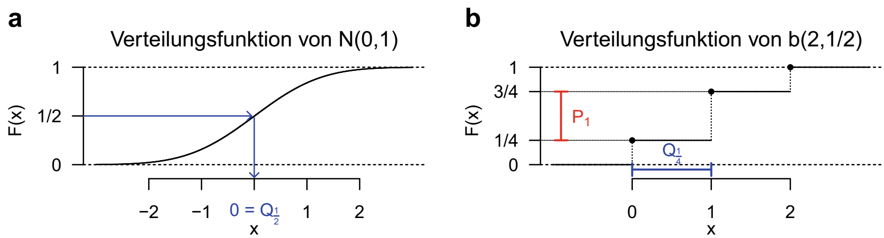

Abb. 2.2 **a** Verteilungsfunktion der $N(0, 1)$-Verteilung und ihr eindeutiges 1/2-Quantil $Q_{1/2} = \{0\}$ (blau), **b** Verteilungsfunktion der $b(2, 1/2)$-Verteilung, mit der Menge aller 1/4-Quantile (blau) und der Menge $\mathcal{P}_1$ aller p, für die die 1 ein p-Quantil ist (rot)

falls $p = 1/2$; ein *erstes Quartil,* falls $p = 1/4$; und ein *drittes Quartil,* falls $p = 3/4$. Weiter beschreibt das Intervall

$$\mathcal{P}_q := [\sup\{F(x)|x < q\}, F(q)] \cap (0, 1) \tag{2.7}$$

die Menge aller $p \in (0, 1)$, für die q ein p-Quantil ist, vgl. Abb. 2.2b. Es ist F in q stetig und streng monoton wachsend (umkehrbar) mit $F(q) = p$ genau dann, wenn $\mathcal{P}_q = \{p\}$ und $Q_p = \{q\}$.

2.1.2 Erwartungswert und Varianz

Wir erinnern an die Begriffe des *Erwartungswertes* und der *Varianz* von Zufallsvariablen. Im Rahmen statistischer Modellierung wird der Erwartungswert eine Kenngröße zur Beschreibung der *Lage,* sowie die Varianz eine Kenngröße zur Beschreibung der *Variabilit*ät von Beobachtungen darstellen. Häufig werden wir auch *stochastische Unabhängigkeit* oder *Unkorreliertheit* von Zufallsvariablen annehmen. Schließlich betrachten wir zwei prominente Ungleichungen im Kontext der Erwartungswerte, nämlich die *Markov-Ungleichung* und die *Jensen-Ungleichung,* welche häufig in Beweisen genutzt werden.

Der Erwartungswert ist über den Begriff des Integrals formuliert. Wir erinnern hier wieder an den Fall, dass die Zufallsvariablen diskret oder stetig sind.

Es sei X eine diskrete oder stetige Zufallsvariable mit Bildraum $\Gamma \subseteq \mathbb{R}$ und h eine Funktion von Γ nach $\mathbb{R}$ so, dass der Ausdruck

$$\int h(X)d\mathbb{P} := \begin{cases} \sum\limits_{k\in\Gamma} h(k)\mathbb{P}(X = k), & \text{falls X diskret} \\ \int\limits_{\Gamma} h(x)f(x)dx, & \text{falls X stetig,} \end{cases}$$

wohldefiniert ist. Insbesondere soll damit $\int |h(X)|d\mathbb{P} < \infty$ gelten. In diesem Falle nennen wir die Zufallsvariable $h(X)$ *integrierbar* und

$$\mathbb{E}[h(X)] := \int h(X)d\mathbb{P}$$

den Erwartungswert von $h(X)$. Intuitiv werden also sämtliche möglichen Ausgänge von X unter Anwendung der Funktion h betrachtet und mit der Wahrscheinlichkeit ihres Auftretens gewichtet aufsummiert. Den Erwartungswert von X erhält man durch Wahl von $h = id$, sofern X integrierbar ist.

Der Erwartungswert von X hat die nette geometrische Bedeutung des Schwerpunkts der Verteilung von X. Betrachten wir etwa eine diskrete oder stetige Zufallsvariable X und stellen uns deren Gewichte bzw. Dichte auf einer Wippe liegend vor. Dann ist der Erwartungswert $\mathbb{E}[X]$ derjenige Punkt, der die Wippe im Gleichgewicht hält. In Abb. 2.1 erkennen wir so etwa den Erwartungswert einer $U\{1, 2, \ldots, n\}$-verteilten Zufallsvariablen als $(n + 1)/2$, oder den einer $N(\mu, \sigma^2)$-verteilten Zufallsvariablen als μ.

Wir betonen die *Linearität* des Erwartungswertes

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \tag{2.8}$$

für integrierbare Zufallsvariable X und Y und Konstanten $a, b \in \mathbb{R}$. Dies folgt aus der Tatsache, dass der Erwartungswert als Integral bzw. Summe per definitionem ein lineares Funktional beschreibt. Aufgrund der Monotonie des Integrals bzw. der Summe folgt zudem auch die Monotonie des Erwartungswertes

$$\mathbb{E}[X] \leq \mathbb{E}[Y], \quad \text{falls } \{X \leq Y\} \text{ mit Wahrscheinlichkeit 1.}$$

Beispiel 2.5 (Indikatorvariable)
Ein prominentes Beispiel einer Zufallsvariablen ist die Indikatorvariable: Für ein Ereignis A (zum Beispiel $A = \{X \in B\}$) ist dessen Indikatorvariable I_A gegeben durch

$$I_A := \begin{cases} 1, & \textit{falls Aeintritt} \\ 0, & \textit{sonst}. \end{cases}$$

Die Indikatorvariable I_A ist diskret mit Werten in $\Gamma = \{0, 1\}$. Es gilt $\mathbb{P}(I_A = 1) = \mathbb{P}(A) =: p$. Daher ist I_A Bernoulli-verteilt zum Erfolgsparameter p, und durch Einsetzen in die Definition des Erwartungswertes erhalten wir direkt

$$\mathbb{E}[I_A] = 1 \cdot \mathbb{P}(I_A = 1) + 0 \cdot \mathbb{P}(I_A = 0) = p,$$

d. h., der Erwartungswert der Indikatorvariable I_A gleicht der Erfolgswahrscheinlichkeit für das Eintreten des Ereignisses A.

Für $p > 0$ nennen wir eine Zufallsvariable X *p-fach integrierbar,* kurz $X \in \mathscr{L}^p$, falls $\int |X|^p d\mathbb{P} < \infty$. Für $p = 1$ entspricht dies gerade der Integrierbarkeit von X. Gilt die Bedingung für $p = 2$, so heißt X *quadratintegrierbar.* Ist $X \in \mathscr{L}^p$, so nennen wir $\mathbb{E}[X^p]$ das *p-te Moment.*

Für $p' \geq p \geq 1$ gilt die Implikation $X \in \mathscr{L}^{p'} \Rightarrow X \in \mathscr{L}^p$, denn es ist $\{|X|^{p'} \geq |X|^p\}$ genau dann, wenn $\{|X| \geq 1\}$, sodass aufgrund der Linearität und der Monotonie des Erwartungswertes gilt

$$\mathbb{E}\Big[|X|^p\Big] = \mathbb{E}\Big[I_{\{|X|\geq 1\}}|X|^p\Big] + \mathbb{E}\Big[I_{\{|X|<1\}}|X|^p\Big] \leq \mathbb{E}\Big[|X|^{p'}\Big] + 1.$$

Insbesondere folgt aus der Quadratintegrierbarkeit von X die Integrierbarkeit von X. Die Quadratintegrierbarkeit von X impliziert die Existenz der *Varianz* von X, definiert durch

$$\mathbb{V}ar(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2, \tag{2.9}$$

wobei die zweite Gleichung aus der Linearität des Erwartungswertes folgt. Die Varianz beschreibt also, um wie viel die Zufallsvariable erwartungsgemäß vom Schwerpunkt ihrer Verteilung (quadratisch) abweicht.

Die Wurzel aus der Varianz nennen wir die *Standardabweichung* $(\mathbb{V}ar(X))^{1/2}$ von X. Warum formulieren wir explizit sowohl die Varianz als auch die Standardabweichung? Das liegt im Grunde daran, dass sich auf der einen Seite mit der Varianz besonders charmant rechnen lässt (siehe zum Beispiel Gleichung (2.12) – im Falle unkorrelierter Zufallsvariablen schreibt sich die Varianz ihrer Summe als die Summe ihrer Varianzen – Pythagoras lässt grüßen!). Andererseits liegt die Bedeutung der Standardabweichung darin, dass sie durch das Wurzelziehen die gleichen Einheiten wie die Zufallsvariable X oder der Erwartungswert besitzt.

Wir erinnern an die *Kovarianz* zweier quadratintegrierbarer Zufallsvariablen X und Y als

$$\mathbb{C}ov(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{2.10}$$

Dass die Kovarianz existiert, d.h. $\mathbb{E}[|XY|] < \infty$ ist, folgt aus der Cauchy-Schwarz-Ungleichung $\mathbb{E}[|XY|]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$, siehe zum Beispiel Kersting und Wakolbinger (2010).

Zudem ist die Kovarianz eine symmetrische Bilinearform, sodass $\mathbb{C}ov(aX + bY, Z) = a\mathbb{C}ov(X, Z) + b\mathbb{C}ov(Y, Z)$, falls auch $Z \in \mathscr{L}^2$, und analog ließe sich die zweite Komponente bearbeiten. Wir nennen X und Y *unkorreliert,* falls $\mathbb{C}ov(X, Y) = 0$. Ist $\mathbb{C}ov(X, Y) \neq 0$, so nennen wir sie *korreliert.* Haben X und Y positive Varianz, so ist ihr *Korrelationskoeffizient* gegeben durch

$$\mathbb{C}or(X, Y) := \frac{\mathbb{C}ov(X, Y)}{\left[\mathbb{V}ar(X)\mathbb{V}ar(Y)\right]^{1/2}}. \tag{2.11}$$

Aufgrund der Cauchy-Schwarz-Ungleichung finden wir, dass $|\mathbb{C}or(X, Y)| \leq 1$. Die Korrelation hängt nicht von den Einheiten der Zufallsvariablen ab. Kovarianz und Korrelation sind Maße für den linearen Zusammenhang von X und Y. Ist der Zusammenhang beispielsweise perfekt linear, etwa $Y = aX + b$, so erkennen wir wegen der Bilinearität der Kovarianz, dass $|\mathbb{C}or(X, Y)| = 1$.

Aufgrund der Bilinearität folgern wir außerdem für die Varianz einer Summe

$$\mathbb{V}ar(X+Y) = \mathbb{C}ov(X+Y, X+Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y) + 2\mathbb{C}ov(X, Y).$$

Sind X und Y unkorreliert, so vereinfacht sich die Varianz ihrer Summe zu

$$\mathbb{V}ar(X+Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y). \tag{2.12}$$

Zwei Zufallsvariable X und Y heißen (stochastisch) *unabhängig,* falls für sämtliche Ereignisse $\{X \in B_1\}$ und $\{Y \in B_2\}$ gilt, dass $\mathbb{P}(\{X \in B_1\} \cap \{Y \in B_2\}) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2)$. Sind X und Y unabhängig und zusätzlich quadratintegierbar, dann gilt $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ (siehe Georgii 2009), sodass die Unabhängigkeit aufgrund (2.10) die Unkorreliertheit impliziert. Die Umkehrung gilt i. Allg. nicht.

Beispiel 2.6 (Erwartungswert und Varianz)

i. *Indikatorvariable: Aus Beispiel 2.5 wissen wir schon, dass $I_A \sim ber(p)$, mit $p = \mathbb{P}(A)$, sowie Erwartungswert $\mathbb{E}[I_A] = p$. Was ist die Varianz einer $ber(p)$-verteilten Zufallsvariablen? Da $I_A^2 = I_A$, folgt durch Einsetzen in die Definition (2.9)*

$$\mathbb{V}ar(I_A) = \mathbb{E}[I_A^2] - \mathbb{E}[I_A]^2 = p - p^2 = p(1-p).$$

ii. *Binomialverteilung: Es sei $X \sim b(n, p)$. Dann ist $\mathbb{E}[X] = np$ und $\mathbb{V}ar(X) = np(1-p)$.*
Dies folgt aus der Tatsache, dass X so verteilt ist wie die Summe einer unabhängigen Münzwurffolge, genauer $X \sim \sum_{i=1}^n X_i$, wobei $X_1, \ldots, X_n$ unabhängig und identisch verteilt, mit $X_1 \sim ber(p)$. Bezüglich der Summe lässt sich dann die Linearität des Erwartungswertes (2.8), sowie (2.12) ausnutzen:

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np \qquad \textit{und} \qquad \mathbb{V}ar(X) = \sum_{i=1}^n \mathbb{V}ar(X_i) = np(1-p).$$

iii. *Exponentialverteilung: Es sei $X \sim exp(\lambda)$, mit Parameter $\lambda > 0$. Dann folgt mit partieller Integration*

$$\begin{aligned}\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx = -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty \frac{1}{\lambda} e^{-\lambda x} dx = 0 + \int_0^\infty e^{-\lambda x} dx \\ = -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}.\end{aligned}$$

Analog kann man zeigen, dass $\mathbb{V}ar(X) = 1/\lambda^2$.

iv. *Uniforme Verteilung: Sei $X \sim U[a, b]$, mit $a < b$. Dann gilt*

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \textit{und} \qquad \mathbb{V}ar(X) = \frac{(b-a)^2}{12}.$$

Denn nach dem Transformationssatz für Dichten (vgl. Krengel 2005) ist $Y := (X - a)/(b - a) \sim U[0, 1]$ *und hat dann die Dichte* $f(y) = \mathbb{1}_{[0,1]}(y)$, *sodass* $\mathbb{E}[Y] = \int_0^1 y dy = 1/2$ *und* $\mathbb{V}ar(Y) = \int_0^1 y^2 dy - 1/4 = 1/3 - 1/4 = 1/12$. *Damit ist* $\mathbb{E}[X] = (b - a)\mathbb{E}[Y] + a = (a + b)/2$ *und* $\mathbb{V}ar(X) = (b - a)^2 \mathbb{V}ar(Y) = (b - a)^2/12$.

Im Rahmen des Erwartungswertes formulieren wir zwei prominente *Ungleichungen,* die Markov- und die Jensen-Ungleichung. Mit der Markov-Ungleichung lassen sich Wahrscheinlichkeiten durch den Erwartungswert abschätzen. Die Jensen-Ungleichung liefert eine Abschätzung speziell hinsichtlich Funktionen von Zufallsvariablen und deren Erwartungswerten.

Lemma 2.7 (Markov-Ungleichung)
Es sei X *eine Zufallsvariable mit Bildraum* $\Gamma \subseteq \mathbb{R}$ *und* h *eine monoton wachsende Funktion von* Γ *nach* $\mathbb{R}$ *so, dass* $h(X)$ *integrierbar ist. Dann gilt für alle* $c \in \mathbb{R}$ *die Ungleichung*

$$h(c)\mathbb{P}(X \geq c) \leq \mathbb{E}[h(X)]. \tag{2.13}$$

Beweis

$$h(c)\mathbb{P}(X \geq c) = h(c) \int I_{\{X \geq c\}} d\mathbb{P} \leq \int I_{\{X \geq c\}} h(X) d\mathbb{P} \leq \mathbb{E}[h(X)].$$

Ist Y eine quadratintegrierbare Zufallsvariable, so ergibt sich durch Wahl von $X = |Y - \mathbb{E}[Y]| \in \mathbb{R}^+$, sowie $h(x) = x^2$, die bekannte *Chebyschev-Ungleichung* für $\varepsilon > 0$ als

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \leq \frac{\mathbb{V}ar(Y)}{\varepsilon^2}. \tag{2.14}$$

Lemma 2.8 (Jensen-Ungleichung)
Es sei X *eine integrierbare Zufallsvariable mit Bildraum* $\Gamma \subseteq \mathbb{R}$, *und es bezeichne* $\mu := \mathbb{E}[X]$. *Weiter sei* h *eine konvexe Funktion von* Γ *nach* $\mathbb{R}$ *so, dass* $h(X)$ *integrierbar ist. Dann gilt die Ungleichung*

$$h(\mathbb{E}[X]) \leq \mathbb{E}[h(X)]. \tag{2.15}$$

Ist zudem h *bei* μ *zweimal differenzierbar mit* $h''(\mu) > 0$, *dann gilt Gleichheit in (2.15) genau dann, wenn* X *mit Wahrscheinlichkeit* 1 *konstant ist* ($\Leftrightarrow \mathbb{V}ar(X) = 0$).

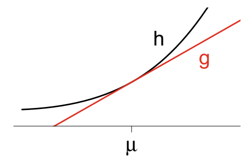

Abb. 2.3 Zur Jensen-Ungleichung

Beweisidee Die Jensen-Ungleichung rührt daher, dass sich eine konvexe Funktion h an jedem Punkt durch eine lineare Funktion ‚stützen' lässt, d. h., dass $h(x) \geq c(x-\mu)+h(\mu) =: g(x)$, für ein $c \in \mathbb{R}$, vgl. Abb. 2.3. Setzen wir nun X ein und bilden den Erwartungswert, so folgt die Ungleichung.

Für die Aussage über die Gleichheit bemerken wir hier, dass h bei μ nach Voraussetzung echt links gekrümmt ist, was aufgrund der Konvexität zur Folge hat, dass $g(X) < h(X)$ mit positiver Wahrscheinlichkeit, sofern nicht X mit Wahrscheinlichkeit 1 konstant μ ist.

2.2 Konvergenz von Zufallsvariablen

Im Folgenden erinnern wir an *Folgen* von Zufallsvariablen $X_1, X_2, \ldots$ Wir diskutieren zunächst drei *Konvergenzbegriffe,* also Arten der Annäherung der Folgenglieder gegen eine Zufallsvariable X, die uns im Rahmen der Statistik immer wieder begegnen werden. Wir betrachten die Konvergenzbegriffe exemplarisch am *Gesetz der großen Zahlen* sowie dem *Zentralen Grenzwertsatz.* Dann führen wir einen vierten Konvergenzbegriff ein und diskutieren Implikationen der Konvergenzarten.

Definition 2.9 (Konvergenzbegriffe)
Es sei $X, X_1, X_2, \ldots$ eine Folge von Zufallsvariablen.
Konvergenz mit Wahrscheinlichkeit 1*:*
Die Folge $(X_n)_{n=1,2,\ldots}$ konvergiert mit Wahrscheinlichkeit 1 *gegen X, kurz $X_n \to X$ mit Wahrscheinlichkeit* 1*, wenn*

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Stochastische Konvergenz:
Die Folge $(X_n)_{n=1,2,\ldots}$ konvergiert stochastisch gegen X, kurz $X_n \xrightarrow{\mathbb{P}} X$, wenn

$$\forall\, \varepsilon > 0: \qquad \mathbb{P}(|X_n - X| > \varepsilon) \longrightarrow 0 \qquad \textit{für } n \to \infty.$$

Konvergenz in Verteilung:
Es bezeichne $F, F_1, F_2, \ldots$ die Folge der Verteilungsfunktionen der $X, X_1, X_2 \ldots$ Die Folge $(X_n)_{n=1,2,\ldots}$ konvergiert in Verteilung gegen X, kurz $X_n \xrightarrow{d} X$, falls

$$F_n(z) \longrightarrow F(z) \quad \text{für } n \to \infty,$$

an allen Stellen $z \in \mathbb{R}$, an denen F stetig ist.

In Lemma 2.14 ist festgehalten, dass die Konvergenz mit Wahrscheinlichkeit 1 die stochastische Konvergenz impliziert, welche wiederum die Konvergenz in Verteilung nach sich zieht. Der Buchstabe d in $X_n \xrightarrow{d} X$ steht für *distribution,* also Verteilung. Die definierende Eigenschaft ist hier über die Verteilungen gegeben, und wir schreiben daher auch häufig $X_n \xrightarrow{d} \nu$, falls ν der der Verteilungsfunktion F assoziierten Verteilung entspricht.

Ein prominentes Beispiel für die Konvergenz mit Wahrscheinlichkeit 1 ist das *Starke Gesetz der großen Zahlen.* Es sei $X_1, X_2, \ldots$ eine Folge von Zufallsvariablen. Dann ist der *Mittelwert* der ersten n Beobachtungen bezeichnet durch

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

Satz 2.10 (Starkes Gesetz der großen Zahlen)
Es sei $X_1, X_2, \ldots$ eine Folge unabhängiger und identisch verteilter, integrierbarer Zufallsvariablen. Dann gilt für $n \to \infty$

$$\bar{X}_n \longrightarrow \mathbb{E}[X_1] \quad \text{mit Wahrscheinlichkeit } 1$$

Ziehen wir also immer wieder unabhängig aus der Verteilung von X_1, so nähert sich der Mittelwert $\bar{X}_n$ dem Erwartungswert der Verteilung an. Besonders toll ist, dass wir keine explizite Annahme an die Verteilung der X_i gemacht haben. Wir erkennen auch sofort die Notwendigkeit der Integrierbarkeit, denn sonst würde der Grenzwert gar nicht existieren. In Abb. 2.4 ist das Starke Gesetz der großen Zahlen visualisiert. Wir betrachten jeweils $X_1, \ldots, X_{100}$ unabhängige und identisch verteilte, integrierbare Zufallsvariable. In beiden Abbildungen geht es um je fünf Verteilungen von X_1. Diese sind farblich gekennzeichnet. Für jede Verteilung ist eine Realisierung der Mittelwertfolge $\bar{X}_1, \ldots, \bar{X}_{100}$ dargestellt. In Abb. 2.4a sind die fünf Verteilungen durch die $poi(\lambda)$-Verteilung zu den Parametern $\lambda = 1, 2, \ldots, 5$ gegeben, d. h. der Erwartungswert ist jeweils der Parameter λ. In Abb. 2.4b

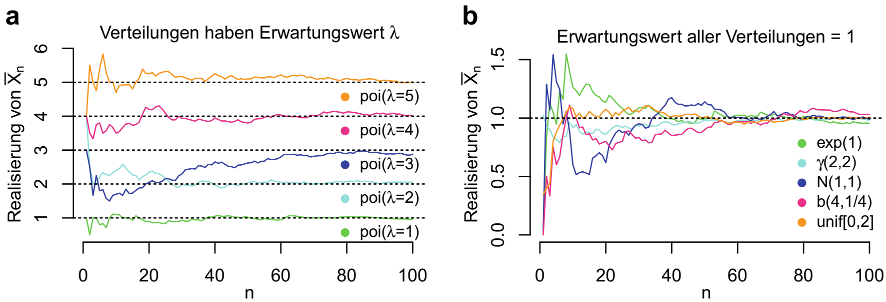

Abb. 2.4 Visualisierung des Starken Gesetzes der Großen Zahlen. Realisierungen von $\bar{X}_n$ als Funktion von n, **a** Poisson-verteilte Beobachtungen mit verschiedenen Erwartungswerten, **b** verschiedene Verteilungen mit Erwartungswert 1

haben die betrachteten Verteilungen alle Erwartungswert 1. In beiden Grafiken erkennen wir, wie sich jede realisierte Mittelwertfolge dem Erwartungswert der Verteilung von X_1 annähert.

Eine schwächere Version von Satz 2.10 ist das *Schwache Gesetz der großen Zahlen.* Dies besagt, dass der Mittelwert $\bar{X}_n$ stochastisch gegen den Erwartungswert $\mathbb{E}[X_1]$ konvergiert. Und da Konvergenz mit Wahrscheinlichkeit 1 die stochastische Konvergenz impliziert, folgt das Schwache Gesetz der großen Zahlen direkt aus Satz 2.10. Wüssten wir lediglich um das Schwache Gesetz der großen Zahlen, so könnten wir in Abb. 2.4 *nicht* kommentieren, dass sich (mit Wahrscheinlichkeit 1) jeder ‚Pfad' schließlich in einer vorgegebenen Umgebung von $\mathbb{E}[X_1]$ befinden wird. Wir wüssten lediglich, dass die Wahrscheinlichkeit, außerhalb einer solchen Umgebung zu liegen, mit wachsendem n gegen null strebt, was aber noch nichts über das Verhalten eines jeden Pfades aussagt. Die Gültigkeit der Aussage des Schwachen Gesetzes der Großen Zahlen unter der stärkeren Annahme der Quadratintegrierbarkeit der X_i erkennen wir direkt aus der Chebyschev-Ungleichung (2.14)

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}ar(\bar{X}_n)}{\varepsilon^2} = \frac{\mathbb{V}ar(X_1)}{n\varepsilon^2} \longrightarrow 0 \qquad \text{für } n \to \infty,$$

wobei wir Gleichung (2.12) ausgenutzt haben. Interessanterweise würde es hier sogar ausreichen, die Unkorreliertheit anstatt der Unabhängigkeit der Zufallsvariablen zu fordern. Für einen Beweis von Satz 2.10 siehe zum Beispiel Feller (1968).

Ein bekanntes Beispiel für Verteilungskonvergenz ist der *Zentrale Grenzwertsatz.* Wieder geht es um den Mittelwert. Die Aussage ist, dass der Mittelwert nach geeigneter Reskalierung in Verteilung gegen die Normalverteilung konvergiert.

Satz 2.11 (Zentraler Grenzwertsatz)
Es seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte, quadratintegrierbare Zufallsvariable mit $\mu := \mathbb{E}[X_1]$ und $\sigma^2 := \mathbb{V}ar(X_1) > 0$. Dann gilt für $n \to \infty$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1).$$

Ein Beweis findet sich zum Beispiel in Kersting und Wakolbinger (2010). Ein Beispiel sehen wir in Abb. 2.5 an $X_1 \sim exp(1)$. Wir ziehen mehrere Stichproben der Größe $n = 10$ von unabhängigen $exp(1)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ (Abb. 2.5a, pro Zeile je zehn Realisierungen durch graue Punkte dargestellt) und bestimmen jeweils deren Mittelwert $\bar{X}_n$ (rot). In Abb. 2.5b sehen wir nur noch die Mittelwerte (rot) und deren Dichte (orange) für wachsendes n. Es ist $\mathbb{E}[\bar{X}_n] = \mu = 1$ und $\mathbb{V}ar(\bar{X}_n) = \sigma^2/n$, mit $\sigma^2 = 1$, und wir sehen, wie sich die Dichte von $\bar{X}_n$ der Dichte der $N(\mu, \sigma^2/n)$-Verteilung (blau) annähert. Nach dem Starken Gesetz der großen Zahlen konvergiert allerdings $\bar{X}_n$ gegen 1 mit Wahrscheinlichkeit 1. Die Dichte von $\bar{X}_n$ zieht sich immer mehr bei $\mu = 1$ zusammen. In Abb. 2.5c erkennen wir die Bedeutung des Zentralen Grenzwertsatzes: Betrachten wir $\bar{X}_n^* := (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$, skalieren wir also den (zentrierten) Mittelwert noch mit $\sigma/\sqrt{n}$, so ist dies gerade die richtige Größenordnung, damit sich die Verteilung von $\bar{X}_n^*$ (orange) der Standardnormalverteilung (blau) nähert.

Wir erinnern an das Schwache Gesetz der großen Zahlen, das im Grunde besagt, dass die Wahrscheinlichkeit der Abweichung des Mittelwertes $\bar{X}$ von dem Erwartungswert $\mathbb{E}[X_1]$ klein wird. Der Zentrale Grenzwertsatz gibt nun sogar Auskunft über die Größenordnung dieser Abweichung, denn wir wissen nun näherungsweise um die Wahrscheinlichkeit, um

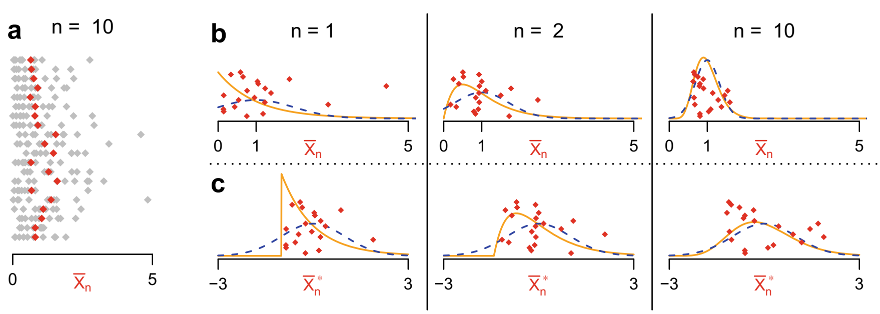

Abb. 2.5 Illustration des Zentralen Grenzwertsatzes am Beispiel standard-exponentialverteilter Zufallsvariablen. **a** Stichproben mit ihren Mittelwerten (rot), **b** Dichte (orange) der Mittelwerte (rot) und Normalverteilung (blau, gestrichelt), **c** analog für skalierte Mittelwerte

mehr als $\sigma/\sqrt{n}$ abzuweichen. Zusammenfassend betonen wir die Gutartigkeit des Mittelwertes. Er konvergiert mit Wahrscheinlichkeit 1 gegen den Erwartungswert und nach geeigneter Reskalierung gegen die Normalverteilung. Der Mittelwert wird uns immer wieder über den Weg laufen.

Weniger geläufig in den Einführungsveranstaltungen zur Stochastik ist das folgende Resultat, häufig auch *Satz von Slutsky* genannt, das gewissermaßen einen Zusammenhang zwischen zwei Folgen von Zufallsvariablen herstellt. Es spielt in der Anwendung eine entscheidende Rolle, da die Verteilungsparameter in der Praxis fast nie bekannt sind, sondern geschätzt werden müssen. Durch den Satz von Slutsky kann man die unbekannten Verteilungsparameter durch geschätzte Größen ersetzen, die stochastisch gegen den wahren Parameter konvergieren.

Satz 2.12 (Satz von Slutsky)
Es seien $X, X_1, X_2, \ldots, Y_1, Y_2, \ldots$ Zufallsvariable und $c \in \mathbb{R}$ konstant und es gelten für $n \to \infty$ die Konvergenzen $X_n \xrightarrow{d} X$ und $Y_n \xrightarrow{\mathbb{P}} c$. Dann folgt auch

$$i)\quad X_n + Y_n \xrightarrow{d} X + c, \qquad\qquad ii)\quad X_n Y_n \xrightarrow{d} cX.$$

Ein Beweis findet sich zum Beispiel in Feller (1971). Wir erinnern noch an einen vierten Konvergenzbegriff.

Definition 2.13 (Konvergenz im p-ten Mittel)
Es sei $p > 0$ und $X, X_1, X_2, \ldots$ eine Folge p-fach integrierbarer Zufallsvariablen. Die Folge X_n konvergiert im p-ten Mittel gegen X, kurz $X_n \xrightarrow{\mathscr{L}^p} X$, falls

$$\mathbb{E}\left[(X_n - X)^p\right] \longrightarrow 0 \qquad \text{für } n \to \infty.$$

Schließlich wollen wir noch die nach Definition 2.9 behaupteten Implikationen der Konvergenzbegriffe zusammenfassen. Wir werden diese dann praktisch im Rahmen der Statistik benötigen, zum Beispiel, um Funktionen von Beobachtungen, etwa anhand des Satzes von Slutsky, miteinander zu kombinieren. Wir formulieren folgendes Lemma.

Lemma 2.14 (Implikationen der Konvergenzbegriffe)
Es sei $X, X_1, X_2, \ldots$ eine Folge von Zufallsvariablen, $p > 0$. Dann gilt für $n \to \infty$

$$\begin{array}{lcl} X_n \longrightarrow X \text{ mit Wahrscheinlichkeit } 1 & \overset{(1)}{\Longrightarrow} & \\ & & X_n \xrightarrow{\mathbb{P}} X \overset{(3)}{\Longrightarrow} X_n \xrightarrow{d} X. \\ X_n \xrightarrow{\mathscr{L}^p} X & \overset{(2)}{\Longrightarrow} & \end{array}$$

Ist der Grenzwert konstant, $\{X = c\}$ mit Wahrscheinlichkeit 1, so gilt auch die Umkehrung von (3).

Beweisideen ‚$\overset{(1)}{\Longrightarrow}$': Siehe zum Beispiel Bauer (2002). ‚$\overset{(2)}{\Longrightarrow}$': Die Aussage folgt direkt aus der Markov-Ungleichung unter Betrachtung der Folgenglieder $|X_n - X| \in \mathbb{R}_0^+$ und der Funktion $h(z) = z^p$, welche auf $\mathbb{R}_0^+$ streng monoton wächst. Für $\varepsilon > 0$ gilt dann für $n \to \infty$

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \longrightarrow 0.$$

‚$\overset{(3)}{\Longrightarrow}$': Das ist eine unmittelbare Folgerung des Satzes von Slutsky (Satz 2.12). Dort setze $X_n := Z \xrightarrow{d} Z$ und $Y_n := Z_n - Z \xrightarrow{\mathbb{P}} 0$. Dann folgt $Z_n = X_n + Y_n \xrightarrow{d} Z$, für $n \to \infty$. ‚$\overset{(3)}{\Longleftarrow}$ (falls $\{X = c\}$ mit Wahrscheinlichkeit 1)': Per definitionem gilt, dass die Folge der Verteilungsfunktionen F_n von X_n punktweise konvergiert gegen $F(z) = \mathbb{1}_{[c,\infty)}(z)$ für $z \neq c$. Sei $\varepsilon > 0$, dann sind $c \pm \varepsilon$ Stetigkeitsstellen von F, und es gilt für $n \to \infty$, dass

$$\begin{aligned} \mathbb{P}(|X_n - c| > \varepsilon) &= \mathbb{P}(X_n < c - \varepsilon) + \mathbb{P}(X_n > c + \varepsilon) \\ &\leq F_n(c-\varepsilon) + 1 - F_n(c+\varepsilon) \longrightarrow F(c-\varepsilon) + 1 - F(c+\varepsilon) = 0. \end{aligned}$$

Dass die Umkehrung von (3) i. Allg. nicht gilt, zeigt folgendes Beispiel. Es seien $X, X_1, X_2, \ldots$ unabhängig und identisch verteilt, mit $X \sim ber(1/2)$ (Münzwurffolge). Insbesondere ist X nicht konstant. Dann gilt $X_n \xrightarrow{d} X$, da die Verteilungsfunktionen identisch sind, aber für jedes $\varepsilon \in (0, 1)$ gilt $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n \neq X) = 1/2$ für alle n.