

# Einführung in die Wahrscheinlichkeitstheorie und Statistik

Dr. C.J. Luchsinger

## 3 Erwartungswerte

### Literatur Kapitel 3

- \* Statistik in Cartoons: Kapitel 4
- \* Kregel: 3.3 und 3.5 in § 3 und (Honours Program) 11.4 in § 11
- \* Storrer: 37, 38, 39, 40, 41, 49

### 3.1 Erwartungswert und Varianz einer diskreten und stetigen Zufallsgrösse

#### Ziele dieses Teils:

- \* Die StudentInnen kennen die Definition des Erwartungswerts / der Varianz von diskreten und stetigen Zufallsgrössen.
- \* Sie können einfache Erwartungswerte / Varianzen selber berechnen und kennen weitere Erwartungswerte / Varianzen von bekannten Zufallsgrössen (mehr dazu in Kapitel 4).
- \* Gefühl für Erwartungswerte / Varianzen (z.B. Schwerpunkt; "normalerweise" etwas wie "mittlerer Wert")

Falls wir Daten  $x_1, x_2, \dots, x_n$  haben (mehr dazu in 3.3), können wir einen sogenannten **Stichproben-Mittelwert** berechnen:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i.$$

Es ist denkbar, dass einige der  $x_i$  den gleichen Wert darstellen. Wir könnten also obiges  $\bar{x}$  umschreiben, indem wir über alle möglichen Werte von  $x$  summieren und mit  $n_x$  angeben, wie häufig Wert  $x$  vorgekommen ist. Damit erhalten wir:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{\text{alle } x} x n_x = \sum_{\text{alle } x} x \frac{n_x}{n}.$$

Warum haben wir das gemacht? Weil

$$\frac{n_x}{n}$$

die *relative* Häufigkeit darstellt: wie häufig kommt  $x$  in  $n$  Daten vor, geteilt durch  $n$ . Wir werden in Kapitel 5 sehen, dass diese Grösse gegen die wahre Wahrscheinlichkeit für dieses Ereignis konvergiert

$$P[X = x],$$

wenn  $n \rightarrow \infty$ . Deshalb definieren wir:

**Definition 3.1 [Erwartungswert einer diskreten und stetigen Zufallsgrösse]**

Der Erwartungswert  $E[X]$  einer Zufallsgrösse  $X$  ist definiert als

$$E[X] := \begin{cases} \sum_{x_i} x_i P[X = x_i] & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{falls } X \text{ stetig.} \end{cases}$$

Sei  $g(x)$  eine (messbare) Funktion von  $\mathbb{R}$  nach  $\mathbb{R}$ . Dann definieren\* wir:

$$E[g(X)] = \begin{cases} \sum_{x_i} g(x_i) P[X = x_i] & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{falls } X \text{ stetig.} \end{cases}$$

Diese Definitionen gelten, falls die Summe bzw. das Integral existiert (siehe Bemerkung 4 nachfolgend). Dabei wird jeweils über den gesamten Wertebereich der Zufallsgrösse summiert respektive integriert. Diese einfachen Definitionen reichen für unsere Vorlesungen aus. Es gibt allgemeinere Definitionen von  $E$  (siehe VlsG WT). Die "Definition\*" von  $E[g(X)]$  ist übrigens streng genommen ein kleines Resultat und keine Definition (vgl VlsG WT).

**Bemerkungen** 1. Eine andere Bezeichnung für Erwartungswert ist *Mittelwert*.

2. Die Zufallsgrösse muss den Erwartungswert nicht annehmen: "Reality differs from Expectations!" Dazu noch das einfache Beispiel von  $\text{Be}(p)$ , wo  $p \in (0, 1)$  und  $P[X = 1] = p = 1 - P[X = 0]$ . Berechnen Sie den Erwartungswert dieser Zufallsgrösse:

Er wird offenbar von  $X$  *nie* angenommen, weil  $X$  entweder 0 oder 1 ist.

3. Obschon wir im täglichen Leben oft mit Erwartungswerten argumentieren, ist es gar nicht so einfach, zu verstehen, was das genau ist. Definition 3.1 ist algebraisch (eine Summe) bzw. von der Analysis her (ein Integral) klar. Physikalisch ist der Mittelwert ein Schwerpunkt. Die Wahrscheinlichkeiten sind dann Gewichte bzw. Gewichtsverteilungen.

4. [dieser 4. Punkt Honours-Programm; nicht Prüfungsstoff] In Definition 3.1 haben wir eine Summe (bzw. Integral) über potentiell unendlich viele Summanden / unendliches Intervall. Es können 4 Fälle auftreten:

a) Summe/Integral ist  $\in (-\infty, \infty)$  "Normalfall"; Bsp. aus Bemerkung 2

b) Summe/Integral ist  $= +\infty$ ; Bsp. auf Übungsblatt 6.

c) Summe/Integral ist  $= -\infty$ ; Bsp. aus b) mit  $Y := -X$

d) Summe/Integral ist nicht definiert: negativer Teil und positiver Teil geben  $-\infty$  bzw.  $+\infty$ ; Bsp. auf Übungsblatt 6, wenn man Dichte an  $y$ -Achse spiegelt und damit im negativen Bereich fortsetzt (und die Normierungskonstante halbiert!).

## Beispiele I

1. Berechnen Sie in der Stunde den Erwartungswert der Anzahl Augen beim Wurf eines perfekten Würfels. Überlegen Sie sich zuerst, was es geben sollte.

2. Berechnen Sie in der Stunde den Erwartungswert einer  $U[-2, 1]$ -Zufallsgrösse. Überlegen Sie sich zuerst, was es geben sollte.

Meist ist es bei der Berechnung von Erwartungswerten von immensem Vorteil, wenn man schon weiss, was es geben sollte. Klassisches Beispiel dazu ist die Poisson-Zufallsgrösse. Wenn man weiss, dass der Erwartungswert  $\lambda$  sein muss, ist es ganz einfach:

3. Erwartungswert einer  $\text{Po}(\lambda)$ -Zufallsgrösse,  $\lambda > 0$ : Eine Poisson-Zufallsgrösse hat die Verteilung:

$$P[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0.$$

Damit steigen wir folgendermassen ein:

$$\begin{aligned} E[X] &= \sum_{k \geq 0} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k \geq 0} k \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k \geq 1} k \frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda} \sum_{k \geq 1} k \frac{\lambda^{k-1}}{k!} \\ &= \lambda e^{-\lambda} \sum_{k \geq 1} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{k \geq 0} \frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda \end{aligned}$$

4. Auf Übungsblatt 6 sind weitere Erwartungswerte zu berechnen. Wir fügen hier noch ein gerechnetes Beispiel an, nämlich der Erwartungswert einer  $\mathcal{N}(\mu, \sigma^2)$ -Zufallsgrösse. Je nach Zeit wird das Beispiel in der Vorlesung auch besprochen. Der Erwartungswert sollte definitionsgemäss  $\mu$  sein. Die Dichte ist (vgl. Aufgabe auf Blatt 4):

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Damit steigen wir folgendermassen ein:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu) + \mu}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx + \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx + \mu \\ &= 0 + \mu \\ &= \mu. \end{aligned}$$

Offenbar ist der Erwartungswert *ein* Mass für die **Lage** der Zufallsgrösse ("wo werden Werte etwa erwartet?"). Wir werden jetzt *ein* Mass für die **Streuung** der Zufallsgrösse um diesen Erwartungswert kennenlernen.

**Definition 3.2 [Varianz/Standardabweichung einer diskreten und stetigen Zufallsgrösse]** Sei  $E[X^2] < \infty$ . Mit  $\mu_X := E[X]$  definieren wir die Varianz  $V[X]$  einer Zufallsgrösse  $X$  als

$$V[X] := E[(X - \mu_X)^2] = \begin{cases} \sum_{x_i} (x_i - \mu_X)^2 P[X = x_i] & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx & \text{falls } X \text{ stetig.} \end{cases}$$

Dabei wird auch hier über den gesamten Wertebereich der Zufallsgrösse summiert respektive integriert. Die Standardabweichung  $sd$  (**S**tandard **D**eviation) ist die Wurzel aus der Varianz:

$$sd[X] := \sqrt{V[X]}.$$

Man beachte: der Ausdruck

$$E[(X - \mu_X)^2]$$

besteht aus 3 (!) Teilen. Welchen und weshalb?

**Bemerkung zu Definition 3.2:** Varianz bzw. Standardabweichung sind *zwei* Masse für die Streuung einer Zufallsgrösse. Es gibt aber viele weitere Masse für die Streuung. Otto Normalverbraucher würde unter Standardabweichung übrigens eher den Ausdruck:

$$E[|X - \mu_X|] \tag{3.1}$$

vermuten. Dies ist die absolute ("|"), erwartete ("E") Abweichung vom Erwartungswert ("X -  $\mu_X$ "): Mean absolute deviation. In den Übungen ist ein einfaches Beispiel anzugeben, das zeigt, dass

$$E[|X - \mu_X|] = sd[X]$$

im Allgemeinen *nicht* gilt. In späteren Semestern wird man in der Analysis übrigens lernen, dass wegen der Hölder-Ungleichung immer gilt:

$$E[|X - \mu_X|] \leq sd[X].$$

Man verwendet aus mathematischen Gründen (einfachere Rechnungen) in der Statistik eher die Varianz anstatt die Mean absolute deviation.

## Beispiele II

5. Berechnen Sie die Varianz einer  $\text{Be}(p)$ -Zufallsgrösse, wo  $p \in (0, 1)$ :

6. Berechnen Sie die Varianz einer  $U[0, 1]$ -Zufallsgrösse:

7. Was vermuten Sie: wie wird die Varianz einer  $U[0, 3]$ -Zufallsgrösse sein (Auflösung nach Lemma 3.7)?

8. Wir fügen hier noch ein gerechnetes Beispiel an, nämlich die Varianz einer  $\mathcal{N}(\mu, \sigma^2)$ -Zufallsgrösse. Je nach Zeit wird das Beispiel in der Vorlesung auch besprochen. Die Varianz sollte definitionsgemäss  $\sigma^2$  sein. Die Dichte ist (vgl. Aufgabe auf Blatt 4):

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Damit steigen wir folgendermassen ein (partielle Integration im 5. Schritt):

$$\begin{aligned}
 V[X] &:= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\
 &= \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}y^2} dy \\
 &= \int_{-\infty}^{\infty} y \left( y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}y^2} \right) dy \\
 &= -y \frac{\sigma^2}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}y^2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{\sigma^2}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}y^2} dy \\
 &= \sigma^2
 \end{aligned}$$

In den Übungen ist die Varianz einer exponentialverteilten Zufallsgrösse zu berechnen.

**Definition 3.3 [Erwartungswert bei 2 involvierten Zufallsgrössen]** *Der Erwartungswert  $E[g(X, Y)]$  (z.B.  $g(x, y) = xy$ ) von 2 Zufallsgrössen  $X$  und  $Y$  auf dem gleichen Wahrscheinlichkeitsraum ist definiert als*

$$E[g(X, Y)] := \begin{cases} \sum_{x_i} \sum_{y_j} g(x_i, y_j) P[X = x_i, Y = y_j] & \text{falls } X, Y \text{ diskret} \\ \int_x \int_y g(x, y) f(x, y) dy dx & \text{falls } X, Y \text{ stetig.} \end{cases}$$

*Dabei wird auch hier über den gesamten Wertebereich der Zufallsgrössen summiert respektive integriert. Analog verfährt man bei mehr als 2 Zufallsgrössen.*

Wir rechnen zu Definition 3.3 noch ein diskretes Beispiel vor. Es ist deshalb sehr wichtig, weil sehr viel der bisherigen Theorie darin vorkommt!

**Aufgabe:** Sei  $\Omega := \{\omega_1, \omega_2, \omega_3\}$ .  $P$  sei derart, dass  $P[\{\omega_1\}] = 0.5$ ,  $P[\{\omega_2\}] = 0.3$  und  $P[\{\omega_3\}] = 0.2$ .  $X(\omega_1) = 3$ ,  $X(\omega_2) = 4$ ,  $X(\omega_3) = 5$ ;  $Y(\omega_1) = 2$ ,  $Y(\omega_2) = 2$ ,  $Y(\omega_3) = 7$ . Gesucht ist  $E[X^2Y + 1/Y]$ .

**Lösung:** Wie ist  $g$ ?  $g(a, b) = a^2b + 1/b$ .

Welche Werte nimmt  $(X, Y)$  an?  $(3, 2), (4, 2), (5, 7)$ .

Mit welchen Wahrscheinlichkeiten?  $0.5, 0.3, 0.2$ ; d.h. z.B.

$$P[X = 3, Y = 2] := P[\{\omega | X(\omega) = 3, Y(\omega) = 2\}] = P[\{\omega_1\}] = 0.5.$$

In den Summen in Definition 3.3 kann man auch die Wertekombination  $(3, 7)$  suchen und sogar berücksichtigen! Man kann sogar alle 3 mal 3 Kombinationen nehmen (ohne Unterscheidung, dass  $Y$  2 mal den gleichen Wert annimmt). Jedoch kommen nur 3 Fälle mit positiver Wahrscheinlichkeit vor  $((3, 2), (4, 2), (5, 7))$ , der Rest wird dann in unterer Summe mit 0 multipliziert.

Wir berechnen jetzt die Summe aus Definition 3.3, wobei wir nur noch die Kombinationen nehmen, welche positive Wahrscheinlichkeit haben:

$$E[X^2Y + 1/Y] = (18 + 0.5)0.5 + (32 + 0.5)0.3 + (175 + 1/7)0.2 \doteq 54.03.$$

### 3.2 Einige wichtige Rechenregeln

**Lemma 3.4 [Absolutbetrag "vorher" und "nachher", Linearität, Indikatorfunktion]** *Mit obiger Notation gelten (falls Summen und Integrale existieren):*

a)  $|E[X]| \leq E[|X|]$ .

b) *Linearität des Erwartungswertes:*

$$E[aX + bY] = aE[X] + bE[Y]$$

*und damit unter anderem  $E[b] = b$  und  $E[0] = 0$  (setze z.B.  $a = 0, Y = 1$ ). Umgangssprachlich nennt man Teil b): "Konstanten herausziehen" und "Summen auseinanderziehen".*

c) *Mit  $I$  Indikatorfunktion (d.h.  $I_A(\omega) = 1$  gdw  $\omega \in A$ , 0 sonst, wo  $A \in \mathcal{A}$ ) gilt:*

$$E[I_A] = P[I_A = 1] = P[A].$$

*Umgangssprachlich: "Erwartungswert von Indikator ist Wahrscheinlichkeit!"*

**Bemerkungen zu Lemma 3.4:** 1. a) ist deshalb klar, weil sich auf der linken Seite positive und negative  $X(\omega)$  gegenseitig aufheben können, bevor der absolute Wert genommen wird. Auf der rechten Seite werden zuerst die absoluten Werte genommen, womit dies nicht mehr passieren kann. Damit wird die rechte Seite " $\geq$ ". Diese Ungleichung wird in der Analysis (und Wahrscheinlichkeitstheorie) noch in den verschiedensten Varianten vorkommen.

2. Wir haben gesehen, dass der Erwartungswert einer  $\text{Be}(p)$ -Zufallsgrösse gleich  $p$  ist. Die  $\text{Bin}(n, p)$ -Zufallsgrösse ist ja eine Summe von  $n$   $\text{Be}(p)$ -Zufallsgrössen (sogar unabhängige Summanden!). Wegen Lemma 3.4 b) muss deshalb der Erwartungswert einer  $\text{Bin}(n, p)$ -Zufallsgrösse gleich  $np$  sein (vgl. auch die direkte Berechnung in den Übungen).

(teilweise) Beweis von Lemma 3.4 (mehr in Vlsg WT):

a) Wir machen den stetigen Fall; diskret als kleine, freiwillige Übung.

$$|E[X]| = \left| \int x f(x) dx \right| \leq \int |x| f(x) dx = E[|X|].$$

b) Wir machen den stetigen Fall; diskret als kleine, freiwillige Übung. Mit Definition 3.3 und Bemerkung 2 aus Sektion 2.4 folgt:

$$\begin{aligned} E[aX + bY] &= \int \int (ax + by) f_{X,Y}(x, y) dy dx \\ &= \int \int ax f_{X,Y}(x, y) dy dx + \int \int by f_{X,Y}(x, y) dy dx \\ &= \int ax \int f_{X,Y}(x, y) dy dx + \int by \int f_{X,Y}(x, y) dx dy \\ &= \int ax f_X(x) dx + \int by f_Y(y) dy \\ &= a \int x f_X(x) dx + b \int y f_Y(y) dy \\ &= aE[X] + bE[Y] \end{aligned}$$

c)

$$E[I_A] = 0P[I_A = 0] + 1P[I_A = 1] = 0P[A^c] + 1P[A] = P[A].$$

□

Für lineare  $g$  gilt offenbar

$$E[g(X)] = g(E[X]);$$

dies ist aber für beliebige  $g$  im Allgemeinen falsch. Immerhin gilt die sogenannte Jensen-Ungleichung für konvexe  $g$ :

**Lemma 3.5 [Jensen-Ungleichung]** Für konvexe  $g$  gilt:  $E[g(X)] \geq g(E[X])$ . [math. exakt:  $g$  muss borelsch sein und sowohl  $E[|g(X)|] < \infty$  wie auch  $E[|X|] < \infty$ .]

**Bemerkungen zu Lemma 3.5:** 1. Eine "Anwendung" ist Lemma 3.7 b) wo  $g(x) = x^2$ .

2. In den Übungen ist zu zeigen, dass  $E[1/X] = 1/E[X]$  im Allgemeinen *nicht* gilt. Sei jetzt  $X$  eine stetige Zufallsgrösse auf  $(0, \infty)$ . Versuchen Sie mit Hilfe von Lemma 3.5 herauszufinden, ob in diesem Fall  $E[1/X] \leq 1/E[X]$  oder  $E[1/X] \geq 1/E[X]$  gilt.

3. Wie steht es mit konkaven  $g$ ?

**Beweis von Lemma 3.5:**

$g$  konvex bedeutet, dass für jedes  $x$  und  $a$  gilt ("Tangente an  $a$ "):

$$g(x) \geq g(a) + (x - a)g'(a).$$

Wir wählen jetzt speziell:  $a := E[X]$ . Dann gilt:

$$g(x) \geq g(E[X]) + (x - E[X])g'(E[X]).$$

Da dies für *jedes*  $x$  gilt, können wir fortfahren mit:

$$g(X) \geq g(E[X]) + (X - E[X])g'(E[X]).$$

Wir nehmen hiervon den Erwartungswert:

$$E[g(X)] \geq g(E[X]) + E[(X - E[X])]g'(E[X]) = g(E[X]),$$

weil  $E[(X - E[X])] = 0$ .

□

**Lemma 3.6 [Alternative Berechnung des Erwartungswerts]** *Mit obiger Notation gelten (falls Summen und Integrale existieren):*

*Falls  $X$  Werte auf  $\{0, 1, 2, 3, \dots\}$  annimmt, dann gilt:*

$$E[X] = \sum_{n \geq 1} P[X \geq n].$$

*Falls  $X$  eine stetige Zufallsgrösse auf  $(0, \infty)$  mit Verteilungsfunktion  $F(t)$  ist, dann gilt analog:*

$$E[X] = \int_0^{\infty} (1 - F(t)) dt = \int_0^{\infty} P[X \geq t] dt.$$

**(unvollständiger) Beweis von Lemma 3.6:** Diese beiden überraschend eleganten Resultate haben im diskreten wie im stetigen Fall eine sehr einfache Beweisskizze; bei beiden Fällen gibt es dann aber leider eine schwierige Passage, welche Umordnungssätze benötigt, welche uns (noch) nicht zur Verfügung stehen (mehr in Vlsg WT). Wir begnügen uns deshalb mit einleuchtenden Beweisskizzen:

diskret:

Wir beweisen jetzt mit Lemma 3.6, dass  $E[Ge(p)] = 1/p$ .

stetig:

$$\begin{aligned} E[X] &:= \int_0^\infty s f(s) ds \\ &= \int_0^\infty \left[ \int_0^\infty I_{[0,s]}(t) dt \right] f(s) ds \\ &= \int_0^\infty \left[ \int_0^\infty I_{[0,s]}(t) f(s) ds \right] dt \\ &= \int_0^\infty \left[ \int_0^\infty I_{[s \geq t]}(s) f(s) ds \right] dt \\ &= \int_0^\infty P[X \geq t] dt \end{aligned}$$

Im dritten Schritt haben wir die Integrationsreihenfolge vertauscht. Dies können wir bis jetzt nicht begründen.

□

**Lemma 3.7 [elementare Rechenregeln der Varianz]** *Mit obiger Notation gelten:*

- a)  $V[aX + b] = a^2 V[X]$  für  $a, b$  reelle Zahlen ("Konstante quadratisch raus").
- b)  $V[X] = E[X^2] - (E[X])^2$

**Bemerkungen zu Lemma 3.7:** 1. Wir haben uns in Beispiel 7 gefragt, wie wohl die Varianz einer  $U[0, 3]$ -Zufallsgrösse sein muss. Sei  $X$  eine  $U[0, 1]$ -Zufallsgrösse. Man kann einfach via Verteilungsfunktion (Kapitel 2.6) zeigen, dass dann  $Y := 3X$  eine  $U[0, 3]$ -Zufallsgrösse ist. Deshalb muss die Varianz von  $Y$  wegen Lemma 3.7 a) 9 mal grösser sein als diejenige von  $X$  (also  $9/12 = 3/4$ ). Überprüfen Sie, dass  $3X$  eine  $U[0, 3]$ -Zufallsgrösse ist.

2. In den Übungen wird mit Hilfe von Lemma 3.7 b) die Varianz einer exponentialverteilten Zufallsgrösse berechnet.

3. Wir haben in den Übungen  $E[X^2] = \lambda^2 + \lambda$  einer Poisson( $\lambda$ )-verteilten Zufallsgrösse  $X$  berechnet. In Beispiel 3 berechneten wir den Erwartungswert  $E[X] = \lambda$ . Wegen Lemma 3.7 b) ist also die Varianz hier  $V[X] = \lambda^2 + \lambda - \lambda^2 = \lambda (= E[X])$ .

4. Wegen Lemma 3.7 a) folgt mit  $a \in \mathbb{R}$ :

$$sd[aX] = |a|sd[X];$$

im Gegensatz zur Varianz kann man bei der Standardabweichung einen konstanten Faktor einfach herausnehmen (Absolutbetrag!).

**Beweis von Lemma 3.7:** In den Übungen Blatt 7.

□

**Lemma 3.8 [Unabhängigkeit von Zufallsgrössen und Erwartungswerte]** *Mit obiger Notation gelten (falls Summen und Integrale existieren):*

a) *Bei 2 Zufallsgrössen  $X$  und  $Y$  auf dem gleichen Wahrscheinlichkeitsraum gilt im Fall von  $X \perp\!\!\!\perp Y$ , dass*

$$E[XY] = E[X]E[Y]$$

*(„Produkte auseinandernehmen“).*

b) *Sei  $(X_i)_{i=1}^n$  eine Folge von unabhängigen Zufallsgrössen (die gleiche Verteilung wird nicht gefordert!). Dann gilt:*

$$V\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n V[X_i]$$

*(„Varianz der Summe ist die Summe der Varianzen“).*

**Bemerkung zu Lemma 3.8:** Wir haben in Beispiel 5 gesehen, dass die Varianz einer  $\text{Be}(p)$ -Zufallsgrösse gleich  $p(1-p)$  ist. Die  $\text{Bin}(n, p)$ -Zufallsgrösse ist ja eine Summe von  $n$   $\text{Be}(p)$ -Zufallsgrössen (sogar unabhängig!). Wegen Lemma 3.8 b) muss deshalb die Varianz einer  $\text{Bin}(n, p)$ -Zufallsgrösse gleich  $np(1-p)$  sein.

**Deppenliste Rechenregeln:**

**Beweis von Lemma 3.8:** a) Wir machen den stetigen Fall; diskret als kleine, freiwillige Übung. Mit Definition 3.3 ( $g(x, y) := xy$ ) folgt:

$$\begin{aligned} E[XY] &= \int \int xy f_{X,Y}(x, y) dy dx = \int \int xy f_X(x) f_Y(y) dy dx \\ &= \int x f_X(x) dx \int y f_Y(y) dy = E[X]E[Y]. \end{aligned}$$

b) Teil b) sollte jeder durcharbeiten, weil darin wichtige Rechenregeln für Erwartungswerte (v.a. Lemma 3.4 b)) immer wieder eingesetzt werden.

$$\begin{aligned} V\left[\sum_{i=1}^n X_i\right] &:= E\left[\left(\sum_{i=1}^n X_i - E\left[\sum_{i=1}^n X_i\right]\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n E[X_i]\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^n (X_i - E[X_i])\right)^2\right] \\ &= E\left[\sum_{i=1}^n (X_i - E[X_i])^2 + \sum_{i \neq j} (X_i - E[X_i])(X_j - E[X_j])\right] \\ &= E\left[\sum_{i=1}^n (X_i - E[X_i])^2\right] + E\left[\sum_{i \neq j} (X_i - E[X_i])(X_j - E[X_j])\right] \\ &= E\left[\sum_{i=1}^n (X_i - E[X_i])^2\right] + \sum_{i \neq j} E[(X_i - E[X_i])(X_j - E[X_j])] \\ &= E\left[\sum_{i=1}^n (X_i - E[X_i])^2\right] + \sum_{i \neq j} (E[X_i X_j] - E[X_i]E[X_j]) \\ &= E\left[\sum_{i=1}^n (X_i - E[X_i])^2\right] + \sum_{i \neq j} (E[X_i]E[X_j] - E[X_i]E[X_j]) \\ &= E\left[\sum_{i=1}^n (X_i - E[X_i])^2\right] \\ &= \sum_{i=1}^n E[(X_i - E[X_i])^2] \\ &=: \sum_{i=1}^n V[X_i] \end{aligned}$$

Man beachte, dass wir *nie* gefordert haben, dass die  $X_i$ 's die gleiche Verteilung haben!

□

### 3.3 Einschub: Stichproben und deren Erwartungswerte/Varianzen

1. Dieser kleine Einschub ist ein wichtiger Vorgriff auf den Statistik-Teil ab Kapitel 6. Er ist auch wichtig für die Übungen mit dem Statistikpaket R. Folgenden Zusammenhang sollte man sich vergegenwärtigen:

- \* Bis jetzt haben wir immer *Wahrscheinlichkeitstheorie* gemacht.
- \* Die *Wahrscheinlichkeitstheorie* zeichnet sich dadurch aus, dass wir immer sicher wissen, wie das Modell ist (z.B. "Sei  $X$  eine  $\mathcal{N}(0,1)$ -Zufallsgrösse.") Wir müssen uns in der *Wahrscheinlichkeitstheorie* *nie* Gedanken machen, ob dieses Modell überhaupt "stimmt".
- \* Ich setze ab hier voraus, dass Sie weder Gott sind noch Seinen göttlichen Plan (inklusive Verteilungen) kennen; zudem setze ich voraus, dass "Gott würfelt".
- \* In der Statistik gilt folgendes: **Wir haben nur die Daten**  $d = (x_1, x_2, x_3, \dots, x_n)$ !!! und wissen nicht, aus welcher Verteilung die eigentlich stammen. Diese Daten können Würfelauagen sein bei  $n$  Würfeln, Blutdruckmessungen bei verschiedenen Personen oder bei der gleichen Person zu verschiedenen Zeitpunkten, Aktienkurse etc.

2. So ist die Lage. Was jetzt folgt, ist weder göttlich, noch zwingend als Analysemethode. Es ist ein Vorschlag unter vielen. Es gibt aber gute Gründe, am Anfang und unter anderem folgende Untersuchungen zu machen:

- \* Sortieren der Daten nach der Grösse, **R**: `sort(d)`
- \* grösster und kleinster Wert, beide zusammen, Median, **R**: `max(d)`, `min(d)`, `range(d)`, `median(d)`
- \* Histogramm, **R**: `hist(d)`
- \* arithmetisches Mittel, Stichproben-Varianz, **R**: `mean(d)`, `var(d)`

Beim letzten Punkt (arithmetisches Mittel  $\bar{x}$ , Stichproben-Varianz  $s^2$ ) wollen wir ein bisschen verweilen. Per Definitionem gilt hierfür:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$

und

$$s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

In den Übungen ist zu zeigen, dass gilt:

a)

$$\sum_{i=1}^n (x_i - \bar{x}) = 0,$$

b)

$$s^2 = \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2.$$

Bei der Definition von  $\bar{x}$  wird offensichtlich, dass wir hier eigentlich einen Erwartungswert berechnet haben, indem wir jedem Datenpunkt  $x_i$  das Gewicht (die Wahrscheinlichkeit)

$$\frac{1}{n}$$

gegeben haben. Dasselbe haben wir bei der Stichproben-Varianz  $s^2$  auch gemacht (in R wird `var(d)` leicht anders berechnet, man teilt dort durch  $(n - 1)$  - mehr dazu in Kapitel 7).

3. Wir sind von der Theorie zu den Daten gegangen und wollen jetzt unser Augenmerk auf ein Zwischending lenken: Wir können doch in einem Statistikpaket wie R einfach eine 100-er Stichprobe von Daten nehmen, von denen wir wissen, dass sie aus einer  $\mathcal{N}(0, 1)$ -Zufallsgrösse stammen: `rnorm(100)`. Wir setzen hier voraus, dass R *für unsere einfachen Berechnungen* einen guten Zufallsgenerator hat. Der gewaltige Vorteil von Simulationen für die Statistik ist der, dass wir wissen, dass z.B. der Erwartungswert 0 war. `mean(d)` wird aber gerade bei einer  $\mathcal{N}(0, 1)$ -Zufallsgrösse niemals genau 0 sein. Wir können aber so feststellen, wie gut ein Schätzer wie das arithmetische Mittel für die Schätzung des Erwartungswertes ist. Mehr dazu in Kapitel 7. Man mache sich klar, dass wir mit realen Daten nicht wissen, wie die Verteilung eigentlich aussieht und was der Erwartungswert ist (wir sind ja nicht Gott);

**wenn wir `mean(d)` eingeben, kommt einfach eine reelle Zahl heraus... . So what?**

#### 4. Jargon:

\* In der Statistik mit realen Daten aus der Welt sprechen wir von *Daten oder Stichproben*; bei Statistikpaketen oder wenn wir die Theorie anhand eines Beispiels veranschaulichen wollen, sprechen wir von *Realisationen*: "Sei  $x_1, x_2, \dots, x_n$  eine Realisation vom Umfang  $n$  aus einer  $\mathcal{N}(0, 1)$ -Zufallsgrösse".

\* Daten werden immer mit kleinen Buchstaben angegeben. Meist werden wir die dazugehörige Zufallsgrösse (aus der die Realisation stammt oder wir gehen zumindest mal davon aus) mit den dazugehörigen Grossbuchstaben bezeichnen:  $X_i$  und  $x_i$ .

\* Wenn nicht anders vereinbart, werden Stichproben/Realisationen vom Umfang  $n$  so behandelt, dass man jedem Datenpunkt die Wahrscheinlichkeit  $1/n$  zuweist und davon ausgeht, dass die  $n$  Daten unabhängig voneinander generiert wurden. Nochmals: Dies wird (mit gutem Grund) vereinbart und ist keineswegs zwingend! Wir haben dann also:

$$(X_1(\omega), X_2(\omega), \dots, X_n(\omega)) = (x_1, x_2, \dots, x_n),$$

wenn  $X_i$  z.B. die Anzahl Augen beim  $i$ -ten Wurf ist und die Welt im Zustand  $\omega$  ist und die  $X_i$ 's voneinander unabhängig sind.

\* \* \*

Wir wollen den Aspekt mit den Gewichten ( $1/n$ ) noch kurz illustrieren. Wenn man ja eine  $\mathcal{N}(0, 1)$ -Zufallsgrösse hat, so kommen Werte um 0 ja häufiger vor, als Werte um 2, 3 oder gar 4. Aber alle Werte der Realisation werden ja mit  $1/n$  gewichtet. Ist das sinnvoll?

### 3.4 Kovarianz und Korrelation

Die folgende Eigenschaft zweier Zufallsgrößen ist beispielsweise in der Finanzmathematik extrem wichtig (Portfolio-Management, dort v.a. der Mean-Variance-Ansatz von Markovitz). Betrachten wir deshalb ein Beispiel aus der Finanzwelt:

**Definition 3.9 [Kovarianz, Korrelationskoeffizient]** Seien  $X, Y$  zwei Zufallsgrößen auf dem gleichen Wahrscheinlichkeitsraum mit  $E[X^2] < \infty, E[Y^2] < \infty$ .

a) Den Ausdruck  $Cov(X, Y) := E[(X - E[X])(Y - E[Y])]$  nennen wir Kovarianz von  $X$  und  $Y$ .

b) Falls  $V[X] > 0, V[Y] > 0$ , so bezeichnen wir

$$Cor(X, Y) := \frac{Cov(X, Y)}{\sqrt{V[X]V[Y]}}$$

als Korrelationskoeffizient von  $X$  und  $Y$ .

**Bemerkungen zu Definition 3.9** 1. Die Befehle in **R** lauten "cov(x,y)" resp. "cor(x,y)".

2. In den Übungen sind ein paar einfache Turnübungen mit diesen Ausdrücken zu machen.

Es gelten:

$$\text{a) } \text{Cov}(X, Y) := E[(X - E[X])(Y - E[Y])] = E[X(Y - E[Y])] = E[(X - E[X])Y] = E[XY] - E[X]E[Y]$$

$$\text{b) } \text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y) \text{ für } a, b, c, d \in \mathbb{R} \text{ (Lageninvarianz dank Zentrierung!)}$$

$$\text{c) } |\text{Cor}(aX + b, cY + d)| = |\text{Cor}(X, Y)| \text{ für } a, b, c, d \in \mathbb{R} \text{ wo } a \neq 0, c \neq 0 \text{ (Skaleninvarianz dank Normierung!)}$$

$$\text{d) } \text{Cov}(X, X) = V[X]$$

3. Wenn  $\text{Cov}(X, Y) = \text{Cor}(X, Y) = 0$ , so nennen wir 2 Zufallsgrößen unkorreliert. Wenn  $\text{Cor}(X, Y) > 0$  sagen wir,  $X$  und  $Y$  seien positiv korreliert und wenn  $\text{Cor}(X, Y) < 0$  sagen wir,  $X$  und  $Y$  seien negativ korreliert.

4. Bei  $\text{Cov}(X, Y) = 0$  gilt offenbar wegen Turnübung a) auch  $E[XY] = E[X]E[Y]$ . Wir haben in Lemma 3.8 a) gesehen, dass wenn 2 Zufallsgrößen  $X, Y$  unabhängig voneinander sind, dann gilt auch  $E[XY] = E[X]E[Y]$ . Offenbar folgt also aus Unabhängigkeit Unkorreliertheit:

$$\text{unabhängig} \Rightarrow \text{unkorreliert}$$

5. Das folgende Gegenbeispiel zeigt, dass Unabhängigkeit von  $X$  und  $Y$  stärker ist als Unkorreliertheit. Sei  $P[(X, Y) = (-1, 1)] = 1/3, P[(X, Y) = (0, -2)] = 1/3, P[(X, Y) = (1, 1)] = 1/3$ . Zeigen Sie:  $\text{Cov}(X, Y) = 0$ ; zeigen Sie mit Definition 2.6, dass  $X$  und  $Y$  nicht unabhängig voneinander sind.

6. Wenn wir den Beweis von Lemma 3.8 b) nochmals durchgehen, sehen wir, dass für die Gültigkeit der Formel

$$V\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n V[X_i]$$

die Unabhängigkeit der Zufallsgrößen nicht *zwingend* ist: Dank Bemerkung 2 a) haben wir mit  $Cov(X, Y) = 0$  auch  $E[XY] = E[X]E[Y]$  und damit *reicht bereits die Unkorreliertheit!*

\* \* \*

Bevor wir in Lemma 3.10 zwei wichtige Eigenschaften der Korrelationen präsentieren, wollen wir noch auf eine umgangssprachliche Fehlleistung eingehen und die mathematisch exakte Formulierung erarbeiten:

In der Alltagssprache kommt es häufig vor, dass man sagt, "zwei Ereignisse seien miteinander korreliert", beispielsweise das Auftreten von Erdbeben in der Türkei und in Griechenland. Wir haben aber in Definition 3.9 gesehen, dass Korrelationen etwas mit Erwartungswerten von Zufallsgrößen zu tun haben. Ereignisse sind aber selber keine Zufallsgrößen. Wir können uns aber Lemma 3.4 c) bedienen: (zur Erinnerung:  $E[I_A] = P[A]$ ). Sei also  $T$  das Ereignis, es gibt ein Erdbeben in der Türkei und  $G$  das Ereignis, es gibt ein Erdbeben in Griechenland (jeweils innert einer Woche). Man will sagen, "wenn es in der Türkei ein Erdbeben gibt (bedingte Wahrscheinlichkeit), so ist es wahrscheinlicher als sonst (unbedingte Wahrscheinlichkeit), dass es auch ein Erdbeben in Griechenland gibt":

$$P[G|T] > P[G].$$

Daraus folgt aber

$$P[G \cap T] > P[G]P[T].$$

Wenn wir jetzt noch Lemma 3.4 c) zu Hilfe nehmen, so erhalten wir (falls unklar: Krengel 3.4 lesen)

$$E[I_G I_T] > E[I_G]E[I_T].$$

Das heisst aber

$$Cov(I_G, I_T) = E[I_G I_T] - E[I_G]E[I_T] > 0,$$

also positive Kovarianz und damit auch positive Korrelation. Voilà.

Im folgenden Lemma werden wir im Teil b) ein Resultat präsentieren, welches viel zum Verständnis der Korrelationen beiträgt.

**Lemma 3.10 [Der Korrelationskoeffizient als Mass der (linearen) Gleichläufigkeit]** Seien  $X, Y$  zwei Zufallsgrößen auf dem gleichen Wahrscheinlichkeitsraum mit  $E[X^2] < \infty, E[Y^2] < \infty$ . Dann gelten:

a)  $|Cor(X, Y)| \leq 1$ , wegen Normierung!

b)  $|Cor(X, Y)| = 1 \Leftrightarrow \exists a, b \in \mathbb{R} : X(\omega) = a + bY(\omega) \forall \omega \in \Omega \setminus N, P[N] = 0$ .

**Bemerkungen zu Lemma 3.10** 1. In b) finden wir den Ausdruck " $\Omega \setminus N, P[N] = 0$ ". Dies ist rein technisch (fast sicher, bis auf eine Nullmenge  $N$ ). Wenn wir eine (endliche) Stichprobe haben, ist es sogar ohne diesen Zusatz gültig.

2. Man kann es nicht oft genug betonen: es geht in b) nur um (lineare) Gleichläufigkeit zwischen  $X$  und  $Y$ . Wenn man Bemerkung 5 zu Definition 3.9 anschaut, sieht man, dass dort die Korrelation 0 ist. Das heisst, dass es keinerlei (lineare) Gleichläufigkeit gibt. Aber  $Y$  ist 100 % von  $X$  abhängig! Des weiteren folgt weder aus (linearer) Gleichläufigkeit noch aus stochastischer Abhängigkeit zwingend ein *kausaler* Zusammenhang:

3. Bevor wir zum Beweis schreiten, noch ein paar typische Bilder und deren Korrelationen:

**Beweis von Lemma 3.10** a) (E(Bemerkung 2 c) zu Definition 3.9) sei  $E[X] = E[Y] = 0$  und  $V[X] > 0, V[Y] > 0$ . Wir definieren auf Vorrat

$$\lambda := \frac{E[XY]}{V[Y]}.$$

Dann steigen wir folgendermassen ein:

$$\begin{aligned} 0 &\leq E[(X - \lambda Y)(X - \lambda Y)] \\ 0 &\leq E[X^2] - 2\lambda E[XY] + \lambda^2 E[Y^2]. \end{aligned}$$

Jetzt setzen wir  $\lambda$  ein:

$$0 \leq E[X^2] - 2 \frac{E[XY]E[XY]}{V[Y]} + \frac{E[XY]^2 E[Y^2]}{V[Y]^2}.$$

Da  $E[Y] = 0$  ist  $V[Y] = E[Y^2]$ ; damit folgt:

$$0 \leq E[X^2] - \frac{E[XY]^2}{V[Y]}.$$

Das ist aber äquivalent zu:

$$E[XY]^2 \leq E[X^2]V[Y] = E[X^2]E[Y^2].$$

Wenn man jetzt noch die Wurzeln zieht, hat man a) bewiesen. Dieser Beweis wird in der Linearen Algebra allgemein bei der Behandlung des Skalarproduktes nochmals vorkommen. Auch in der Analysis ist dies ein prominentes Resultat. Der Name dieses Teilresultats ist "Ungleichung von Cauchy-Schwarz" oder kurz "CS-ung".

b) " $\Leftarrow$ ": Wir dürfen also  $X = a + bY$  benutzen. Wir benutzen auch Bemerkung 2 c) zu Definition 3.9) und dürfen  $V[X] > 0, V[Y] > 0$  voraussetzen.

$$\begin{aligned} |Cor(X, Y)| &= |Cor(a + bY, Y)| \\ &= \left| \frac{E[(a + bY - a - E[bY])(Y - E[Y])]}{\sqrt{V[bY]V[Y]}} \right| \\ &= \left| \frac{E[(bY - E[bY])(Y - E[Y])]}{\sqrt{V[bY]V[Y]}} \right| \\ &= \left| \frac{E[(Y - E[Y])(Y - E[Y])]}{\sqrt{V[Y]V[Y]}} \right| \\ &= 1. \end{aligned}$$

” $\Rightarrow$ “: Wir dürfen  $a$  und  $b$  frei wählen. Wir wählen (dies muss man wissen, sonst ist der Beweis schwierig)

$$a := E[X] - \frac{\text{Cov}(X, Y)}{V[Y]}E[Y], \quad b := \frac{\text{Cov}(X, Y)}{V[Y]}.$$

Dann steigen wir folgendermassen ein (Nachrechnen als freiwillige, mühsame Hausaufgabe, bei der man sich leicht verrechnen kann):

$$E[(X - (a + bY))^2] = V[X] \left( 1 - (\text{Cor}(X, Y))^2 \right)$$

Wenn jetzt aber die  $|\text{Cor}(X, Y)| = 1$  ist, dann gilt  $E[(X - (a + bY))^2] = 0$ . Wenn wir noch substituieren  $Z := X - (a + bY)$ , so heisst dies

$$E[Z^2] = 0.$$

*Der folgende Schritt ist noch nicht mathematisch für uns begründbar, kann jedoch nachvollzogen werden:* Dies heisst aber  $P[Z = 0] = 1$ , und somit  $X = a + bY \forall \omega \in \Omega \setminus N, P[N] = 0$ .

□

### 3.5 Bedingte Verteilungen und Definition des bedingten Erwartungswertes gegeben $Y = y_0$

Um Verwechslungen zu vermeiden sei vorausgeschickt, dass eine allgemeine Definition bedingter Erwartungswerte sehr aufwendig ist. Wir bedingen hier lediglich auf ein *Ereignis* (gegeben  $Y = y_0$ ) - wir bedingen hier nicht auf eine Zufallsgrösse ( $E[X|Y]$ , mehr dazu in einer allfälligen Vorlesung über Mass- und Wahrscheinlichkeitstheorie). Zudem behandeln wir hier nur die diskreten Zufallsgrössen und harmlose stetige Zufallsgrössen.

Wir müssen als Vorbereitung auf die bedingten Erwartungswerte noch bedingte Verteilungen einführen: Für bedingte Wahrscheinlichkeiten gilt

$$P[A|B] := \frac{P[A \cap B]}{P[B]}.$$

Wir können dann im Fall **diskreter Zufallsgrössen** folgendermassen fortfahren: Mit  $A := \{X = x\}, B := \{Y = y_0\}$  ist  $P[A|B]$  die Wahrscheinlichkeit, dass  $X = x$  ist, wenn  $Y = y_0$  ist (falls  $X \perp\!\!\!\perp Y$ , ist dies einfach  $P[X = x]$ ):

$$P[X = x|Y = y_0] := \frac{P[\{X = x\} \cap \{Y = y_0\}]}{P[\{Y = y_0\}]} := \frac{P[\{\omega|X(\omega) = x\} \cap \{\omega|Y(\omega) = y_0\}]}{P[\{\omega|Y(\omega) = y_0\}]}.$$

Wir können jetzt noch  $x$  variieren und *erhalten damit die Verteilung von ganz  $X|Y = y_0$*  (zur Erinnerung an 1.4.5: bedingte Wahrscheinlichkeiten sind auch Wahrscheinlichkeiten).

Die Schwierigkeiten liegen im **stetigen Fall**: bei stetigen Zufallsgrössen  $X, Y$  gilt ja  $P[X = x] = P[Y = y_0] = 0$ , wenn  $x, y_0$  konkrete, feste, reelle Zahlen sind. Damit hätten wir oben  $0/0$ , was nicht definiert ist. Da sich Wahrscheinlichkeitsfunktionen und Dichtefunktionen von der Bedeutung her entsprechen, ist man trotzdem versucht, bedingte Dichtefunktionen folgendermassen zu *definieren* (und in vielen Lehrbüchern geschieht dies auch diskussionslos):

$$f_{X|Y=y_0}(x) := \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)}, \quad (3.2)$$

falls  $f_Y(y_0) > 0$ . (3.2) können wir als Definition stehen lassen (sie ist sinnvoll), wollen aber trotzdem mit einem Limesresultat noch begründen, dass dies die Verteilung von  $X|Y = y_0$  angibt. Diese Begründung ist nicht Prüfungsstoff sondern Honours-Programm.

Wir repetieren zuerst Definition 2.4 und betrachten dann den Ausdruck

$$P[X \leq a | Y \in [y_0 - h, y_0 + h]] = \frac{P[X \leq a, Y \in [y_0 - h, y_0 + h]]}{P[Y \in [y_0 - h, y_0 + h]]}.$$

Jetzt sind Zähler und Nenner nicht mehr 0 - gehen aber in unteren Rechnungen gegen 0! Wir werden jetzt  $h \searrow 0$  gehen lassen und benutzen zwei mal den Mittelwertsatz der Integralrechnung (Forster Analysis I, Satz 8, § 18). Weiter setzen wir voraus, dass die Dichten  $f_X(x)$ ,  $f_Y(y)$  und  $f_{X,Y}(x, y)$  stetige, beschränkte Funktionen sind (es geht auch mit weniger starken Forderungen, aber wir haben das immer erfüllt).

$$\begin{aligned} \lim_{h \searrow 0} P[X \leq a | Y \in [y_0 - h, y_0 + h]] &= \lim_{h \searrow 0} \frac{P[X \leq a, Y \in [y_0 - h, y_0 + h]]}{P[Y \in [y_0 - h, y_0 + h]]} \\ &= \lim_{h \searrow 0} \frac{\int_{-\infty}^a \int_{y_0-h}^{y_0+h} f_{X,Y}(x, y) dy dx}{\int_{y_0-h}^{y_0+h} f_Y(y) dy} \\ &= \lim_{h \searrow 0} \frac{\int_{-\infty}^a 2h f_{X,Y}(x, \eta(x, h)) dx}{2h f_Y(\xi(h))} \\ &= \lim_{h \searrow 0} \frac{\int_{-\infty}^a f_{X,Y}(x, \eta(x, h)) dx}{f_Y(\xi(h))} \\ &= \frac{\int_{-\infty}^a f_{X,Y}(x, y_0) dx}{f_Y(y_0)} \\ &= \int_{-\infty}^a \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)} dx \end{aligned}$$

Also können wir wegen Definition 2.4 den Ausdruck (3.2) als bedingte Dichte von  $X|Y = y_0$  auffassen. Damit können wir einen wichtigen Begriff einführen:

**Definition 3.11 [bedingter Erwartungswert gegeben  $Y = y_0$ ]** Wir definieren den bedingten Erwartungswert  $E[X|Y = y_0]$  als:

$$E[X|Y = y_0] := \begin{cases} \sum_{x_i} x_i P[X = x_i | Y = y_0] & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} x f_{X|Y=y_0}(x) dx & \text{falls } X \text{ stetig.} \end{cases}$$

Offenbar hängt dieser bedingte Erwartungswert von  $y_0$  ab. Bedingte Wahrscheinlichkeiten sind selber auch Wahrscheinlichkeiten (ebenso sind bedingte Dichten selber auch Dichten!); wir haben also einfach einen gewöhnlichen Erwartungswert von  $X$  berechnet bedingt auf das Ereignis, dass  $Y = y_0$  ist.

Das nachfolgende Resultat ist ein Pendant auf der Ebene der Erwartungswerte zu Lemma 1.7 (FTW) auf der Ebene der Wahrscheinlichkeiten:

**Lemma 3.12 [Formel vom totalen Erwartungswert FTE]** *Mit obigen Bezeichnungen gilt*

$$E[X] = \begin{cases} \sum_{y_j} E[X|Y = y_j]P[Y = y_j] & \text{falls } X, Y \text{ diskret} \\ \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)dy & \text{falls } X, Y \text{ stetig.} \end{cases}$$

**Beweis von Lemma 3.12; diskreter Fall:**

$$\begin{aligned} E[X] &:= \sum_{x_i} x_i P[X = x_i] \\ &= \sum_{x_i} x_i \sum_{y_j} P[X = x_i, Y = y_j] \\ &= \sum_{x_i} x_i \sum_{y_j} P[X = x_i|Y = y_j]P[Y = y_j] \\ &= \sum_{x_i} \sum_{y_j} x_i P[X = x_i|Y = y_j]P[Y = y_j] \\ &= \sum_{y_j} \sum_{x_i} x_i P[X = x_i|Y = y_j]P[Y = y_j] \\ &=: \sum_{y_j} E[X|Y = y_j]P[Y = y_j] \end{aligned}$$

□