Mathe III

Mehrdimensionale Zufallsvariablen

Ralf Herbrich

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Überblick

1. Gemeinsame Verteilung und Randverteilungen
2. Unabhängigkeit von Zufallsvariablen
3. Summen und andere Funktionen von Zufallsvariablen
4. Erwartungswert
5. Kovarianz und Korrelation

Überblick

1. Gemeinsame Verteilung und Randverteilungen
2. Unabhängigkeit von Zufallsvariablen
3. Summen und andere Funktionen von Zufallsvariablen
4. **Erwartungswert**
5. Kovarianz und Korrelation

Erwartungswert von mehreren Zufallsvariablen

- **Definition (Erwartungswert).** Gegeben zwei reelle Zufallsvariablen X und Y und eine Funktion $g\colon \mathbb{R}^2 \to \mathbb{R}$. Dann ist der Erwartungswert von g definiert als

$$E[g(X,Y)] = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} g(x,y)\cdot p_{X,Y}(x,y)$$

Im diskreten Fall

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} g(x,y)\cdot p_{X,Y}(x,y)\,dxdy$$

Im stetigen Fall

- **Satz (Linearität des Erwartungswertes).** Seien X und Y reelle Zufallsvariablen mit je einem existierenden Erwartungswert. Seien außerdem $a,b,c \in \mathbb{R}$. Dann gilt

$$E[a\cdot X + b\cdot Y + c] = a\cdot E[X] + b\cdot E[Y] + c.$$

  - **Beweis**: Folgt direkt aus der Definition des Erwartungswertes

$$E[a\cdot X + b\cdot Y + c] = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}(a\cdot x + b\cdot y + c)\cdot p_{X,Y}(x,y)$$

$$= a\cdot \underbrace{\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} x\cdot p_{X,Y}(x,y)}_{=E[X]} + b\cdot \underbrace{\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} y\cdot p_{X,Y}(x,y)}_{=E[Y]} + c\cdot \underbrace{\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p_{X,Y}(x,y)}_{=1}$$

Beispiel: Erwartungswert der Summe

- **Beispiel (Erwartungswert)**. Eine Assistenz hat N Briefe geschrieben und sie in die richtig frankierten Umschläge gesteckt. Allerdings fallen der Assistenz alle Umschläge runter und die Briefe und Umschläge sind durcheinander. Wenn man davon ausgeht, dass jeder Brief mit gleicher Wahrscheinlichkeit in jedem Umschlag landen kann, was ist die erwartete Anzahl Briefe im richtigen Umschlag?

- **Lösung**: Wir definieren N Zufallsvariablen $X_1, \dots, X_N$ mit der Semantik

$$X_i = \begin{cases} 1 & \text{wenn der } i - \text{te Brief im richtigen Umschlag ist} \\ 0 & \text{sonst} \end{cases}$$

Dann wissen wir, dass $P(X_i = 1) = {}^1/_N$ und damit $E[X_i] = {}^1/_N$.

Folglich gilt damit für den Erwartungswert der Anzahl richtiger Briefe

$$E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i] = N \cdot \frac{1}{N} = 1$$

- **Bemerkung (Erwartungswert)**. Die Aussage gilt, egal wie groß N ist!

Beispiel: Erwartungswert der Summe

- **Beispiel (Erwartungswert).** Nehmen wir an, dass es 20 verschiedene Arten von Coupons gibt und wir 10 zufällig ausgewählte Coupons ziehen (d.h., jeder unabhängig mit gleicher Wahrscheinlichkeit von ${}^{1}/_{20}$ eine bestimmte Art hat). Was ist die erwartete Anzahl von verschiedenen Coupons?
- **Lösung**: Wir definieren 20 Zufallsvariablen $X_1, \ldots, X_{20}$ mit der Semantik

$$X_i = \begin{cases} 1 & \text{wenn mindestens einer der Coupons vom Typ } i \text{ ist} \\ 0 & \text{sonst} \end{cases}$$

Dann wissen wir, dass

$$P(X_i = 1) = 1 - P(X_i = 0) = 1 - \left(\frac{19}{20}\right)^{10}$$

Folglich gilt damit für den Erwartungswert von $\sum_{i=1}^{20} X_i$

$$E\left[\sum_{i=1}^{20} X_i\right] = \sum_{i=1}^{20} E[X_i] = 20 \cdot \left(1 - \left(\frac{19}{20}\right)^{10}\right) \approx 8.025$$

Erwartungswert von mehreren Zufallsvariablen

- **Satz (Produktregel für Erwartungswert).** Seien X und Y reelle Zufallsvariablen mit je einem existierenden Erwartungswert. Wenn X und Y unabhängig sind, dann gilt
$$E[X \cdot Y] = E[X] \cdot E[Y]$$
- **Beweis**: Folgt aus den Eigenschaften für die (Zähl)dichte von unabhängigen Zufallsvariablen

$$\begin{aligned}
E[X \cdot Y] &= \int_{\mathbb{R}} \int_{\mathbb{R}} x \cdot y \cdot p_{X,Y}(x,y)\, dxdy \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x \cdot y \cdot \underbrace{p_X(x) \cdot p_Y(y)}_{\text{Unabhängigkeit}}\, dxdy \\
&= \int_{\mathbb{R}} y \cdot p_Y(y) \cdot \underbrace{\left(\int_{\mathbb{R}} x \cdot p_X(x)\, dx \right)}_{\text{konstant für alle } y} dy \\
&= \left(\int_{\mathbb{R}} x \cdot p_X(x)\, dx \right) \cdot \left(\int_{\mathbb{R}} y \cdot p_Y(y) dy \right)
\end{aligned}$$

Überblick

1. Gemeinsame Verteilung und Randverteilungen
2. Unabhängigkeit von Zufallsvariablen
3. Summen und andere Funktionen von Zufallsvariablen
4. Erwartungswert
5. **Kovarianz und Korrelation**

Kovarianz

- Wir wollen nun Abhängigkeiten zwischen Zufallsvariablen besser beschreiben können.
- Hierfür können wir die Kovarianz nutzen. Sie beschreibt, ob zwischen zwei Zufallsvariablen ein linearer Zusammenhang besteht.

- **Definition (Kovarianz)**. Gegeben seien zwei reelle Zufallsvariablen X und Y mit jeweils existierendem Erwartungswert und Varianz. Die Kovarianz $\mathrm{Cov}[X,Y]$ zwischen X und Y ist dann gegeben als

$$\mathrm{Cov}[X,Y] = E[(X - E[X]) \cdot (Y - E[Y])]$$

- **Bemerkungen (Kovarianz)**
  - Intuitiv prüft die Kovarianz, wie stark X und Y von ihren Erwartungswerten abweichen, und ob die Abweichung vermehrt das gleiche oder entgegengesetzte Vorzeichen hat.

Beispiel Kovarianz

- **Beispiel (Kovarianz).** Gegeben zwei reelle Zufallsvariablen X und Y mit ihrer gemeinsamen Zähldichte als

	$X = 5$	$X = 7$	$P(Y = j)$
$Y = 8$	0.4	0.1	$\mathbf{0.5}$
$Y = 9$	0.3	0.2	$\mathbf{0.5}$
$P(X = i)$	$\mathbf{0.7}$	$\mathbf{0.3}$	

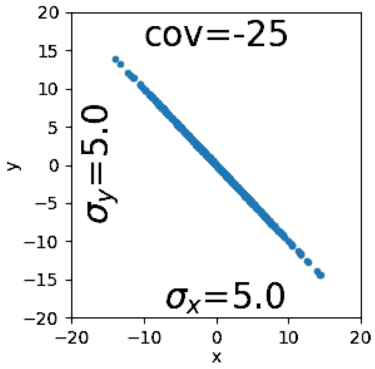
$$E[X] = 0.7 \cdot 5 + 0.3 \cdot 7 = 5.6$$

$$E[Y] = 0.5 \cdot 8 + 0.5 \cdot 9 = 8.5$$

$$\begin{aligned} \mathrm{Cov}[X, Y] &= E[(X - E[X]) \cdot (Y - E[Y])] \\ &= 0.4 \cdot (-0.6) \cdot (-0.5) + 0.3 \cdot (-0.6) \cdot 0.5 + 0.1 \cdot 1.4 \cdot (-0.5) + 0.2 \cdot 1.4 \cdot 0.5 \\ &= 0.12 - 0.09 - 0.07 + 0.14 = 0.1 \end{aligned}$$

Kovarianz

- Die Kovarianz $\mathrm{Cov}(X,Y)$ zwischen X und Y ist dann gegeben als
$$\mathrm{Cov}[X,Y] = E[(X - E[X]) \cdot (Y - E[Y])]$$
- Intuitiv prüft die Kovarianz, wie stark X und Y von ihren Erwartungswerten abweichen, und ob die Abweichung vermehrt das gleiche oder entgegengesetzte Vorzeichen hat.



- **Satz (Interpretation der Kovarianz).** Ist das Vorzeichen der Kovarianz zweier Zufallsvariablen X und Y
  - ... positiv, so treten mit höheren Werten von X tendenziell auch höhere Werte für Y auf.
  - ... negativ, so treten mit höheren Werten von X tendenziell niedrigere Werte für Y auf.
  - ... gleich bzw. fast 0, so besteht keine erkennbare Tendenz für einen linearen Zusammenhang.

Kovarianzzerlegung

- **Satz (Kovarianzzerlegung)**. Für zwei reelle Zufallsvariablen X und Y gilt
$$\mathrm{Cov}[X,Y] = E[X \cdot Y] - E[X] \cdot E[Y]$$
- **Beweis**. Die Aussage folgt direkt aus den Eigenschaften des Erwartungswertes
$$\begin{aligned}\mathrm{Cov}[X,Y] &= E[(X - E[X]) \cdot (Y - E[Y])] \\ &= E\big[X \cdot Y - X \cdot E[Y] - Y \cdot E[X] + E[X] \cdot E[Y]\big] \\ &= E[X \cdot Y] - E[X] \cdot E[Y] - E[Y] \cdot E[X] + E[X] \cdot E[Y] \\ &= E[X \cdot Y] - E[X] \cdot E[Y]\end{aligned}$$
- **Bemerkungen (Kovarianzzerlegung)**
  - Ähnlich zum Varianzzerlegungssatz, erlaubt dieser Satz, die Kovarianz einer empirischen Verteilung („Daten“) online („*streaming*“) zu schätzen!
  - Die Kovarianzzerlegung zeigt auch, dass die Kovarianz symmetrisch ist
$$\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X]$$
und auch, dass
$$\mathrm{Cov}[X,X] = \mathrm{Var}[X]$$

Korrelation

- **Frage**: Bis auf ihr Vorzeichen ist die Kovarianz nur eingeschränkt interpretierbar. Warum?
- **Antwort**: Ihr Wertebereich ist abhängig von X und Y und erstreckt sich auf

$$\left[-\sqrt{V[X]\cdot V[Y]}, +\sqrt{V[X]\cdot V[Y]}\right]$$

- **Beweis**: Wir definieren die Zufallsvariable $Z = (X - E[X]) \cdot (Y - E[Y])$. Dann folgt aus der Positivität von $V[Z]$ dass

$$0 \leq V[Z] = E[Z^2] - (E[Z])^2 \Leftrightarrow (E[Z])^2 \leq E[Z^2]$$

Nun ist aber

$$(E[Z])^2 = (E[(X - E[X]) \cdot (Y - E[Y])])^2 = (\mathrm{Cov}[X, Y])^2$$

und

$$E[Z^2] = E[(X - E[X])^2 \cdot (Y - E[Y])^2] = V[X] \cdot V[Y]$$

Daher gilt also

$$(\mathrm{Cov}[X, Y])^2 \leq V[X] \cdot V[Y]$$

Mathe III

Unit 4b – Mehrdimensionale Zufallsvariablen

Korrelation

- **Idee**: Um einschätzen zu können, wie stark die Tendenz zu einem linearen Zusammenhang ist, normieren wir die Kovarianz auf $[-1, +1]$.
- **Definition (Korrelation)**. Gegeben zwei reelle Zufallsvariablen X und Y mit Erwartungswerten $E[X]$ und $E[Y]$ sowie Varianzen $V[X]$ und $V[Y]$ ist die Korrelation $\mathrm{Corr}[X, Y]$ zwischen X und Y gegeben als

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{V[X] \cdot V[Y]}}$$

- **Satz (Interpretation der Korrelation)**. Ist die Korrelation zweier Zufallsvariablen X und Y näher an
  - … 1, so besteht ein stärkerer positiver linearer Zusammenhang zwischen X und Y.
  - … -1, so besteht ein stärkerer negativer linearer Zusammenhang zwischen X und Y.
  - … 0, so besteht kein erkennbar linearer Zusammenhang zwischen X und Y.

Beispiel Korrelation

- **Beispiel (Korrelation).** Gegeben zwei reelle Zufallsvariablen X und Y mit ihrer gemeinsamen Zähldichte als

	$X = 5$	$X = 7$	$P(Y = j)$
$Y = 8$	0.4	0.1	$\mathbf{0.5}$
$Y = 9$	0.3	0.2	$\mathbf{0.5}$
$P(X = i)$	$\mathbf{0.7}$	$\mathbf{0.3}$	

$$E[X] = 5.6 \qquad E[Y] = 8.5 \qquad \mathrm{Cov}[X, Y] = 0.1$$

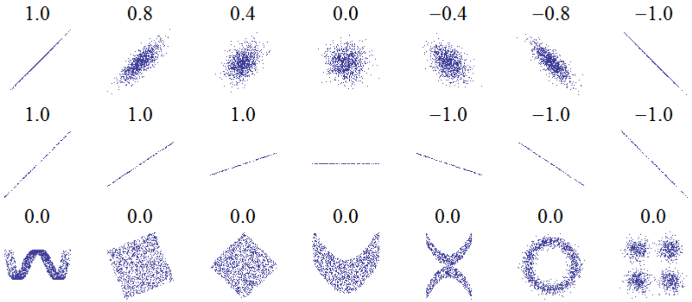
$$V[X] = 0.7 \cdot (-0.6)^2 + 0.3 \cdot 1.4^2 = 0.84$$

$$V[Y] = 0.5 \cdot (-0.5)^2 + 0.5 \cdot 0.5^2 = 0.25$$

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{V[X] \cdot V[Y]}} = \frac{0.1}{\sqrt{0.84 \cdot 0.25}} \approx 0.22$$

Unabhängigkeit und Kovarianz

- Wie wir schon gezeigt haben, gilt für zwei unabhängige Zufallsvariablen X und Y mit Erwartungswert, dass $E[X \cdot Y] = E[X] \cdot E[Y]$.
- **Satz (Unabhängigkeit und Kovarianz)**. Sind zwei reelle Zufallszahlen X und Y mit Erwartungswert und Varianz unabhängig, so gilt für die Kovarianz
$$\operatorname{Cov}[X, Y] = 0$$
- **Beweis**: Folgt direkt aus $\operatorname{Cov}[X, Y] = E[X \cdot Y] - E[X] \cdot E[Y]$.
- **Bemerkungen (Unabhängigkeit und Kovarianz)**.
  - Die Rückrichtung dieses Satzes gilt ausdrücklich nicht!
  - Kovarianz und Korrelation messen nur die Stärke des **linearen** Zusammen-hangs!
  - Gleiches gilt auch für die Kovarianz.



Rechenregeln für Kovarianz

- **Satz (Rechenregel für Kovarianz).** Seien X und Y zwei reelle Zufallsvariablen mit Erwartungswert und Varianz sowie $a, b, c, d \in \mathbb{R}$. Dann gilt

$$\mathrm{Cov}[a \cdot X + b, c \cdot Y + d] = a \cdot c \cdot \mathrm{Cov}[X, Y]$$

- **Beweis**: Folgt aus der Linearität des Erwartungswertes

$$\mathrm{Cov}[a \cdot X + b, c \cdot Y + d] = E[(a \cdot X + b) \cdot (c \cdot Y + d)] - E[a \cdot X + b] \cdot E[c \cdot Y + d]$$

$$E[(a \cdot X + b) \cdot (c \cdot Y + d)] = a \cdot c \cdot E[X \cdot Y] + a \cdot d \cdot E[X] + b \cdot c \cdot E[Y] + b \cdot d$$

$$-E[a \cdot X + b] \cdot E[c \cdot Y + d] = -a \cdot c \cdot E[X] \cdot E[Y] - a \cdot d \cdot E[X] - b \cdot c \cdot E[Y] - b \cdot d$$

$$\mathrm{Cov}[a \cdot X + b, c \cdot Y + d] = a \cdot c \cdot (E[X \cdot Y] - E[X] \cdot E[Y])$$

$$= a \cdot c \cdot \mathrm{Cov}[X, Y]$$

Rechenregeln für Kovarianz: Linearität

- **Satz (Kovarianz einer Summe)**. Seien X_1, X_2 und Y drei reelle Zufallsvariablen mit Erwartungswert und Varianz. Dann gilt
$$\mathrm{Cov}[X_1 + X_2, Y] = \mathrm{Cov}[X_1, Y] + \mathrm{Cov}[X_2, Y]$$
- **Beweis**: Folgt aus der Linearität des Erwartungswertes

$$\begin{aligned}\mathrm{Cov}[X_1 + X_2, Y] &= E[(X_1 + X_2) \cdot Y] - E[X_1 + X_2] \cdot E[Y] \\ &= E[X_1 \cdot Y] + E[X_2 \cdot Y] - (E[X_1] + E[X_2]) \cdot E[Y] \\ &= E[X_1 \cdot Y] - E[X_1] \cdot E[Y] + E[X_2 \cdot Y] - E[X_2] \cdot E[Y] \\ &= \mathrm{Cov}[X_1, Y] + \mathrm{Cov}[X_2, Y]\end{aligned}$$

Rechenregeln für Kovarianz: Linearität

- **Satz (Linearität der Kovarianz).** Seien $X_1, \dots, X_n$ und $Y_1, \dots, Y_m$ reelle Zufallsvariablen mit Erwartungswert und Varianz. Dann gilt

$$\operatorname{Cov}\left[\sum_{i=1}^{n} X_i \,, \sum_{j=1}^{m} Y_j\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{Cov}[X_i, Y_j]$$

- **Beweis**: Folgt direkt aus dem letzten Satz

$$\begin{aligned}
\operatorname{Cov}\left[\sum_{i=1}^{n} X_i \,, \sum_{j=1}^{m} Y_j\right] &= \sum_{i=1}^{n} \operatorname{Cov}\left[X_i, \sum_{j=1}^{m} Y_j\right] \\
&= \sum_{i=1}^{n} \operatorname{Cov}\left[\sum_{j=1}^{m} Y_j \,, X_i\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{Cov}[Y_j, X_i] = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{Cov}[X_i, Y_j]
\end{aligned}$$

Varianz der Summe

- **Satz (Varianz der Summe).** Gegeben eine Folge von n reellen Zufallsvariablen X_i mit $1 \leq i \leq n$ mit Erwartungswert und Varianz. Dann gilt

$$V\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} V[X_i] + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}\left[X_i, X_j\right]$$

  Im Spezialfall zweiter reeller Zufallsvariablen X und Y gilt

$$V[X+Y] = V[X] + V[Y] + 2 \cdot \mathrm{Cov}[X,Y]$$

- **Bemerkungen (Varianz der Summe)**
  - Falls X und Y unabhängig sind, folgt direkt $V[X+Y] = V[X] + V[Y]$.
  - Falls alle X_i paarweise unabhängig sind, folgt direkt

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

Varianz der Summe: Beweis

- **Beweis**: Wir benutzen dass $V[Z] = \mathrm{Cov}[Z, Z]$

$$
\begin{aligned}
V\left[\sum_{i=1}^{n} X_i\right] &= \mathrm{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}[X_i, X_j] \\
&= \sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} \mathrm{Cov}[X_i, X_j] + \boxed{\mathrm{Cov}[X_i, X_i]}\right) \\
&= \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathrm{Cov}[X_i, X_j] + \sum_{i=1}^{n} \boxed{V[X_i]}
\end{aligned}
$$

Beispiel: Varianz der Summe

- **Beispiel (Varianz der Summe).** Wir werfen einen Würfel 10-mal hintereinander. Welche Varianz hat die Summe der Augenzahlen?
- **Lösung**: Wenn wir die Augenzahl des $i$ten Wurfs mit X_i bezeichnen, dann wissen wir für die Varianz $V[X_i]$ von X_i

$$V[X_i] = \frac{1}{6} \cdot (1 + 4 + 9 + 16 + 25 + 36) - \left(\frac{1 + 2 + 3 + 4 + 5 + 6}{6}\right)^2 = \frac{35}{12}$$

Da die Würfe unabhängig voneinander sind, gilt

$$V\left[\sum_{i=1}^{10} X_i\right] = \sum_{i=1}^{10} V[X_i] = \frac{10 \cdot 35}{12} = \frac{175}{6} \approx 29.2$$

Viel Spaß bis zur nächsten Vorlesung!