Zusatzmaterial zur Vorlesung Statistik II

Dr. Steffi Höse
Professurvertretung für Ökonometrie und Statistik, KIT

Wintersemester 2011/2012

(Fassung vom 15.11.2011,
DVI- und PDF-Datei erzeugt am 15. November 2011)

1 Beispiel zur bivariaten Bernoulli-Verteilung (Kap. 10 WT)

Es wird ein 2-dimensionaler Zufallsvektor (X, Y) betrachtet:

- X repräsentiert das Ausfallverhalten Portugals.
- Y repräsentiert das Ausfallverhalten Italiens.
- Die Realisation 1 beschreibt jeweils den Ausfall, die Realisation 0 jeweils den Nichtausfall.

Die gemeinsame Verteilung von X und Y liegt durch die **gemeinsame Verteilungsfunktion**

$$F_{X,Y}(x, y) \stackrel{\text{def}}{=} P(X \leq x, Y \leq y), \qquad \forall x, y \in \mathbb{R}$$

fest.

Sind X und Y **diskrete Zufallsvariablen**, dann gibt es endlich oder abzählbar unendlich viele Werte $x_1, x_2, \ldots$ und $y_1, y_2, \ldots$, so dass

$$\sum_i \sum_j P(X = x_i, Y = y_j) = 1 \tag{1}$$

gilt. Die gemeinsame Verteilung von X und Y kann dann auch durch die **gemeinsame Wahrscheinlichkeitsfunktion** $P(X = x_i, Y = y_j)$, $\forall i, j$ charakterisiert werden.

Beispielsweise sei für obigen Sachverhalt die gemeinsame Wahrscheinlichkeitsfunktion durch[a]

$$P(X = 0, Y = 0) = 0.6 \qquad \text{(gemeinsamer Nichtausfall)}$$
$$P(X = 0, Y = 1) = 0.1 \qquad \text{(nur Italien fällt aus)}$$
$$P(X = 1, Y = 0) = 0.15 \qquad \text{(nur Portugal fällt aus)}$$
$$P(X = 1, Y = 1) = 0.15 \qquad \text{(gemeinsamer Ausfall)}$$

gegeben, womit Bedingung (1) erfüllt ist. Zweidimensionale Wahrscheinlichkeitsverteilungen lassen sich als Stabdiagramm oder in Tabellenform darstellen:

$P(X = x_i, Y = y_j)$	$y_1 = 0$	$y_2 = 1$	$P(X = x_i)$
$x_1 = 0$	0.6	0.1	0.7
$x_2 = 1$	0.15	0.15	0.3
$P(Y = y_j)$	0.75	0.25	1

[a] Die Zahlenwerte stellen ein Beispiel dar und wurden in der Vorlesung gemeinsam mit den Studenten willkürlich gesetzt.

Die Verteilung einer Komponente eines Zufallsvektors wird **Randverteilung** genannt. Die Wahrscheinlichkeitsfunktion der **diskreten Zufallsvariablen** X wird **Randwahrscheinlichkeitsfunktion von** X genannt und ergibt sich aus der gemeinsamen Wahrscheinlichkeitsfunktion als

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j), \qquad i = 1, 2, \ldots.$$

Analog ergibt sich die **Randwahrscheinlichkeitsfunktion von** Y als

$$P(Y = y_j) = \sum_i P(X = x_i, Y = y_j), \qquad j = 1, 2, \ldots.$$

Beispielsweise erhält man für obigen Sachverhalt:

$$\begin{aligned}
P(X = 0) &= P(X = 0, Y = 0) + P(X = 0, Y = 1) = 0.6 + 0.1 = 0.7 \\
P(X = 1) &= P(X = 1, Y = 0) + P(X = 1, Y = 1) = 0.15 + 0.15 = 0.3 \\
P(Y = 0) &= P(X = 0, Y = 0) + P(X = 1, Y = 0) = 0.6 + 0.15 = 0.75 \\
P(Y = 1) &= P(X = 0, Y = 1) + P(X = 1, Y = 1) = 0.1 + 0.15 = 0.25.
\end{aligned}$$

Damit sind X und Y Bernoulli-verteilt, $X \sim Ber(0.3)$, $Y \sim Ber(0.25)$, wobei $P(X = 1)$ als Ausfallwahrscheinlichkeit Portugals und $P(Y = 1)$ als Ausfallwahrscheinlichkeit Italiens betrachtet werden kann.

Die Verteilungsfunktion von X (Y) wird **Randverteilungsfunktion von** X (Y) genannt und ergibt sich aus der Randwahrscheinlichkeitsfunktion von X (Y), siehe oben, oder aus der gemeinsamen Verteilungsfunktion von X und Y.

Die **gemeinsame Verteilungsfunktion von** X **und** Y ergibt sich aus der gemeinsamen Wahrscheinlichkeitsfunktion der **diskreten Zufallsvariablen** X und Y als

$$F_{X,Y}(x,y) \stackrel{\text{def}}{=} P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} P(X = x_i, Y = y_j), \qquad \forall x, y \in \mathbb{R}.$$

Beispielsweise erhält man für obigen Sachverhalt:

$$F_{X,Y}(x,y) = \begin{cases} 0 & & x < 0 \text{ oder } y < 0 \\ P(X=0, Y=0) = 0.6 & & x \in [0,1[\text{ und } y \in [0,1[\\ P(X=0, Y \in \{0,1\}) = P(X=0) = 0.7 & \text{für} & x \in [0,1[\text{ und } y \in [1,\infty[\\ P(X \in \{0,1\}, Y=0) = P(Y=0) = 0.75 & & x \in [1,\infty[\text{ und } y \in [0,1[\\ P(X \in \{0,1\}, Y \in \{0,1\}) = 1 & & x \in [1,\infty[\text{ und } y \in [1,\infty[. \end{cases}$$

Die **Randverteilungsfunktion von** X ergibt sich aus der gemeinsamen Verteilungsfunktion von X und Y als

$$\begin{aligned} F_X(x) &\overset{\text{def}}{=} P(X \leq x) \\ &= P(X \leq x, Y \text{ beliebig}) \\ &= P(X \leq x, Y \leq \infty) \\ &= \lim_{y\to\infty} P(X \leq x, Y \leq y) \\ &= \lim_{y\to\infty} F_{X,Y}(x,y), \qquad \forall x \in \mathbb{R}. \end{aligned}$$

Analog ergibt sich die **Randverteilungsfunktion von** Y als

$$F_Y(y) \overset{\text{def}}{=} P(Y \leq y) = \lim_{x\to\infty} F_{X,Y}(x,y), \qquad \forall y \in \mathbb{R}.$$

Beispielsweise erhält man für obigen Sachverhalt:

$$\begin{aligned} F_X(0) &= \lim_{y\to\infty} F_{X,Y}(0,y) = P(X=0) = 0.7 \\ F_X(1) &= \lim_{y\to\infty} F_{X,Y}(1,y) = 1 \\ F_Y(0.3) &= \lim_{x\to\infty} F_{X,Y}(x,0.3) = P(Y=0) = 0.75 \\ F_Y(5) &= \lim_{x\to\infty} F_{X,Y}(x,5) = 1. \end{aligned}$$

Wie ändert sich das Ausfallverhalten Italiens, wenn Portugal schon ausgefallen wäre? Wie ändert sich die Ausfallwahrscheinlichkeit Portugals unter der Bedingung, dass Italien nicht ausfällt? Die Antworten können mit Hilfe bedingter Wahrscheinlichkeiten gegeben werden.

Für die **diskreten Zufallsvariablen** X und Y gilt:
Falls $P(X = x_i) > 0$ gilt, ist die **bedingte Wahrscheinlichkeitsverteilung von** Y **gegeben** $X = x_i$ definiert als

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}, \qquad j = 1, 2, \ldots.$$

Analog ist für $P(Y = y_j) > 0$ die **bedingte Wahrscheinlichkeitsverteilung von** X **gegeben** $Y = y_j$ definiert als

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}, \qquad i = 1, 2, \ldots.$$

Beispielsweise erhält man für obigen Sachverhalt vier verschiedene bedingte Verteilungen:

$$P(Y=0|X=0)=\frac{P(X=0,Y=0)}{P(X=0)}=\frac{0.6}{0.7}\approx 0.86,\quad P(Y=1|X=0)=\frac{P(X=0,Y=1)}{P(X=0)}=\frac{0.1}{0.7}\approx 0.14$$
$$P(Y=0|X=1)=\frac{P(X=1,Y=0)}{P(X=1)}=\frac{0.15}{0.3}=0.5,\quad P(Y=1|X=1)=\frac{P(X=1,Y=1)}{P(X=1)}=\frac{0.15}{0.3}=0.5$$
$$P(X=0|Y=0)=\frac{P(X=0,Y=0)}{P(Y=0)}=\frac{0.6}{0.75}=0.8,\quad P(X=1|Y=0)=\frac{P(X=1,Y=0)}{P(Y=0)}=\frac{0.15}{0.75}=0.2$$
$$P(X=0|Y=1)=\frac{P(X=0,Y=1)}{P(Y=1)}=\frac{0.1}{0.25}=0.4,\quad P(X=1|Y=1)=\frac{P(X=1,Y=1)}{P(Y=1)}=\frac{0.15}{0.25}=0.6$$

Die auf den Ausfall Portugals bedingte Ausfallwahrscheinlichkeit Italiens $P(Y=1|X=1)=0.5$ muss mit der unbedingten Ausfallwahrscheinlichkeit Italiens $P(Y=1)=0.25$ verglichen werden.

Die auf den Nichtausfall Italiens bedingte Ausfallwahrscheinlichkeit Portugals $P(X=1|Y=0)=0.2$ muss mit der unbedingten Ausfallwahrscheinlichkeit Portugals $P(X=1)=0.3$ verglichen werden.

Liegt ein Zusammenhang zwischen dem Ausfallverhalten Portugals und dem Italiens vor? Eine Antwort gibt das Konzept der stochastischen Unabhängigkeit.

Zwei Zufallsvariablen X und Y bzw. deren Verteilungen heißen **stochastisch unabhängig**, wenn

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y), \qquad \forall x,y \in \mathbb{R}.$$

Sind zwei **diskrete Zufallsvariablen** X und Y bzw. deren Verteilungen stochastisch unabhängig, dann gilt

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y), \qquad \forall x,y \in \mathbb{R}.$$

Falls $P(Y = y) > 0$, dann ist $P(X = x|Y = y)$ definiert und es gilt

$$P(X = x|Y = y) = P(X = x), \qquad \forall x \in \mathbb{R}.$$

Falls $P(X = x) > 0$, dann ist $P(Y = y|X = x)$ definiert und es gilt

$$P(Y = y|X = x) = P(Y = y), \qquad \forall y \in \mathbb{R}.$$

Beispielsweise gilt für obigen Sachverhalt:

$$P(X = 1, Y = 1) = 0.15 \neq 0.075 = 0.3 \cdot 0.25 = P(X = 1) \cdot P(Y = 1)$$

Damit sind im Beispiel die Zufallsvariablen X und Y nicht stochastisch unabhängig.

2 Rechenregeln für Erwartungswert, Varianz, Kovarianz und Korrelation (Kap. 12 und 13 WT)

2.1 Erwartungswert

- Der Erwartungswert der degenerierten Zufallsvariablen X_c mit $P(X_c = c) = 1$ bzw. der Konstanten $c \in \mathbb{R}$ ist
$$E(X_c) = E(c) = c.$$
- Häufig interessieren Erwartungswerte von Funktionen $g(X)$ einer Zufallsvariablen X. Mit X ist auch $Y = g(X)$ eine Zufallsvariable, wenn $g : \mathbb{R} \to \mathbb{R}$ eine messbare Funktion ist. Der Erwartungswert von Y kann direkt aus der Verteilung von X berechnet werden:
$$E(Y) = E(g(X)) = \begin{cases} \int\limits_{-\infty}^{\infty} g(x) f_X(x) dx, & \text{falls } X \text{ stetige Zufallsvariable mit Dichtefunktion } f_X \text{ ist,} \\ \sum\limits_i g(x_i) P(X = x_i), & \text{falls } X \text{ diskrete Zufallsvariable mit Realisationen } x_i \text{ ist.} \end{cases} \tag{2}$$
- Beispiel: Falls $E(X)$ existiert, existiert für $a, b \in \mathbb{R}$ auch $E(a + bX)$ und es gilt
$$E(a + bX) = a + bE(X).$$

- Es sei $g : \mathbb{R}^n \to \mathbb{R}$ eine messbare Funktion und $(X_1, \ldots, X_n)$ ein Zufallsvektor. Dann ist auch $g(X_1, \ldots, X_n)$ eine Zufallsvariable und der Erwartungswert von $g(X_1, \ldots, X_n)$ kann direkt aus der Verteilung des Zufallsvektors $(X_1, \ldots, X_n)$ berechnet werden.
  - Ist der Zufallsvektor $(X_1, \ldots, X_n)$ stetig mit Dichtefunktion $f_{(X_1,\ldots,X_n)}$, dann gilt

$$E(g(X_1, \ldots, X_n)) = \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) dx_1 \cdots dx_n. \quad (3)$$

  - Ist der Zufallsvektor $(X_1, \ldots, X_n)$ diskret mit Realisationen $\left(x_1^{(r)}, \ldots, x_n^{(r)}\right)$, dann gilt

$$E(g(X_1, \ldots, X_n)) = \sum_r g\left(x_1^{(r)}, \ldots, x_n^{(r)}\right) P\left((X_1, \ldots, X_n) = \left(x_1^{(r)}, \ldots, x_n^{(r)}\right)\right). \quad (4)$$

- Beispiele: $E(X_1), \ldots, E(X_n)$ seien endlich.
  - Es gelten

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2), \qquad \forall a, b \in \mathbb{R}$$

    und

$$E\left(a_0 + \sum_{i=1}^{n} a_i X_i\right) = a_0 + \sum_{i=1}^{n} a_i E(X_i), \qquad \forall a_0, \ldots, a_n \in \mathbb{R}.$$

– Für das Stichprobenmittel $\bar{X} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$ gilt

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i).$$

Haben die Zufallsvariablen zusätzlich alle **denselben** Erwartungswert $E(X_i) = \mu$ für $i = 1, \ldots, n$, erhält man

$$E(\bar{X}) = \mu.$$

2.2 Varianz

- Die Varianz der degenerierten Zufallsvariablen X_c mit $P(X_c = c) = 1$ bzw. der Konstanten $c \in \mathbb{R}$ ist
$$Var(X_c) = Var(c) = 0.$$
- Sei $g : \mathbb{R} \to \mathbb{R}$ eine messbare Funktion und X eine Zufallsvariable. Dann ist auch $Y = g(X)$ eine Zufallsvariable und für endlichen Erwartungswert $E(g(X))$ erhält man
$$Var(Y) = Var(g(X)) = E(g(X)^2) - E(g(X))^2.$$
Beide Erwartungswerte können mit Hilfe von (2) berechnet werden.
- Beispiel: Falls $Var(X)$ existiert, existiert für $a, b \in \mathbb{R}$ auch $Var(a + bX)$ und es gilt
$$Var(a + bX) = b^2 Var(X).$$
- Es sei $g : \mathbb{R}^n \to \mathbb{R}$ eine messbare Funktion und $(X_1, \ldots, X_n)$ ein Zufallsvektor. Dann ist auch $g(X_1, \ldots, X_n)$ eine Zufallsvariable und für endlichen Erwartungswert $E(g(X))$ gilt
$$Var(g(X_1, \ldots, X_n)) = E(g(X_1, \ldots, X_n)^2) - E(g(X_1, \ldots, X_n))^2.$$
Beide Erwartungswerte können im Fall eines stetigen Zufallsvektors mit (3) und im Fall eines diskreten Zufallsvektors mit (4) berechnet werden.

- Beispiele: $Var(X_1), \dots, Var(X_n)$ seien endlich.
  - Es gelten

$$Var(aX_1 + bX_2) = a^2 Var(X_1) + 2ab Cov(X_1, X_2) + b^2 Var(X_2), \qquad \forall a, b \in \mathbb{R}$$

  und

$$Var\left(a_0 + \sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2 \sum_{\substack{i=1 \\ i<j}}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j), \qquad \forall a_0, \dots, a_n \in \mathbb{R}.$$

  - Sind alle Zufallsvariablen **unkorreliert**, dann sind alle Kovarianzen Null und es gilt

$$Var\left(a_0 + \sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

  - Im Falle unkorrelierter Zufallsvariablen, hat $\bar{X} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i$ die Varianz

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i).$$

  Haben die Zufallsvariablen zusätzlich alle **dieselbe** Varianz $Var(X_i) = \sigma^2$ für $i = 1, \dots, n$, dann gilt

$$Var(\bar{X}) = \sigma^2 / n.$$

2.3 Kovarianz

Im Folgenden seien alle jeweils notwendigen Kovarianzen endlich. Dann gelten für die Kovarianz folgende Rechenregeln:

$$
\begin{aligned}
Cov(X,X) &= Var(X) \\
Cov(X,Y) &= Cov(Y,X) \\
Cov(X,Y) &= E(XY) - E(X)E(Y) \\
Cov(a+bX,Y) &= bCov(X,Y), \qquad \forall a,b \in \mathbb{R} \\
Cov(aX_1+bX_2,Y) &= aCov(X_1,Y) + bCov(X_2,Y), \qquad \forall a,b \in \mathbb{R} \\
Cov\left(\sum_{i=1}^{n} X_i, Y\right) &= \sum_{i=1}^{n} Cov(X_i,Y) \\
Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i,Y_j) \\
Cov\left(a_0 + \sum_{i=1}^{n} a_i X_i, b_0 + \sum_{j=1}^{m} b_j Y_j\right) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i,Y_j), \qquad \forall a_0,\dots,a_n,b_0,\dots,b_m \in \mathbb{R}
\end{aligned}
$$

2.4 Korrelation

Sind X und Y Zufallsvariablen für die $0 < Var(X) < \infty$ und $0 < Var(Y) < \infty$ gilt, dann ist $Corr(X,Y)$ definiert und es gilt

$$
\begin{aligned}
Corr(X,X) &= 1 \\
Corr(X,Y) &= Corr(Y,X) \\
Corr(a+bX, c+dY) &= \frac{bd}{|bd|} Corr(X,Y), \qquad \forall a,c \in \mathbb{R},\ b,d \in \mathbb{R}\setminus\{0\}, \\
&= \begin{cases} Corr(X,Y), & \text{falls } b>0 \text{ und } d>0 \text{ oder } b<0 \text{ und } d<0 \\ -Corr(X,Y), & \text{falls } b<0 \text{ und } d>0 \text{ oder } b>0 \text{ und } d<0. \end{cases}
\end{aligned}
$$