1.7 Unabhängige und identische Wiederholungen, Grenzwertsätze und Approximation

Fahrmeir et al., 2004, Kap 7.1 f
teilweise auch Jann, 2002, Kap. 5.4

Gerade in Soziologie häufig *große* Stichprobenumfänge

- Was ist das Besondere daran?
- Insbesondere: Vereinfacht sich etwas? Was?
- Kann man ‚Wahrscheinlichkeitsgesetzmäßigkeiten' durch Betrachten vielfacher Wiederholungen erkennen? (ähnlich Naturgesetze)

1.7.1 Das i.i.d.-Modell

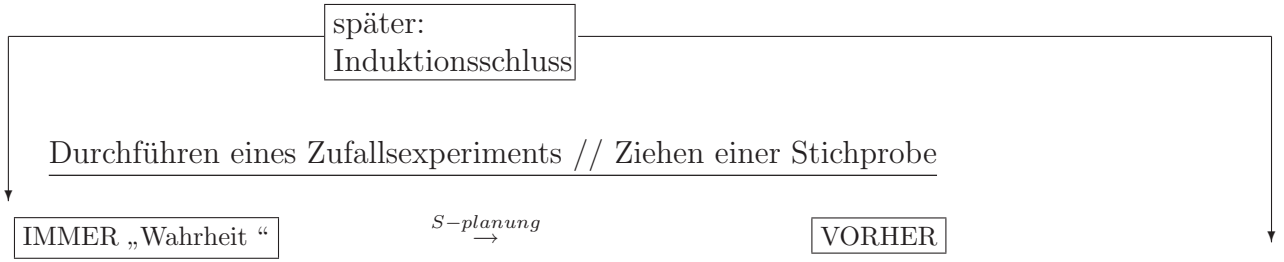
Betrachtet werden diskrete oder stetige Zufallsvariablen $X_1, \ldots, X_n$, die *i.i.d.* (independently, identically distributed) sind, d.h., die

1) unabhängig sind und

2) die gleiche Verteilung besitzen.

Ferner existiere der Erwartungswert μ und die Varianz σ^2; die Verteilungsfunktion werde mit F bezeichnet.
Denke insbesondere an die Situation von Bem. und Bsp. 1.25: $X_1, \ldots, X_n$ i.i.d. Stichprobe eines Merkmals $\widetilde{X}$ bei reiner Zufallsauswahl,

Jede (stetige) Funktion von $X_1, \ldots, X_n$ ist wieder eine Zufallsvariable, z.B $\sum_{i=1}^{n} X_i$. Insbesondere ist damit auch die geplante Auswertung der Realisationen $x_1, \ldots, x_n$ durch Zufallsvariablen beschreibbar, soll heißen:


später:
Induktionsschluss
Durchführen eines Zufallsexperiments // Ziehen einer Stichprobe
IMMER „Wahrheit “
S−planung
VORHER

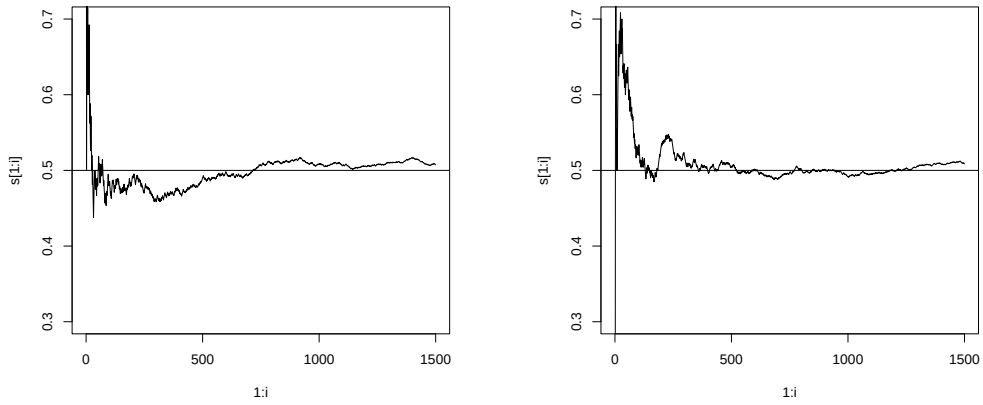
Vor dem Ziehen der Stichprobe: Wahrscheinlichkeitsaussagen möglich $\Longrightarrow$ Wahrscheinlichkeitsrechnung anwenden

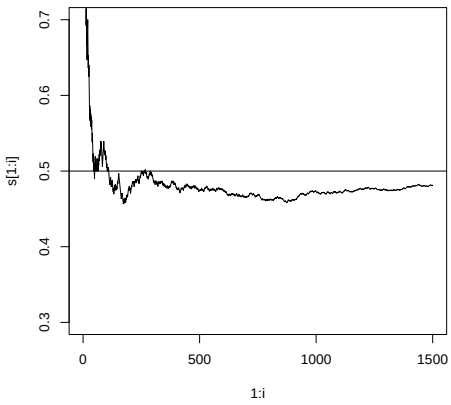
- Gerade bei diesen Zufallsgrößen ist die Abhängigkeit von n oft wichtig, man schreibt dann $\overline{X}_n$, $\widetilde{S}_n^2$
- Sind oben $X_1, \ldots, X_n$ jeweils $\{0,1\}$-Variablen, so ist $\bar{X}_n$ gerade die empirische *relative Häufigkeit* von Einsen in der Stichprobe vom Umfang n. Notation: H_n

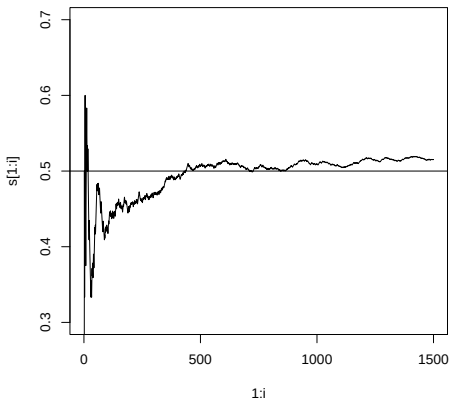
1.7.2 Das schwache Gesetz der großen Zahlen

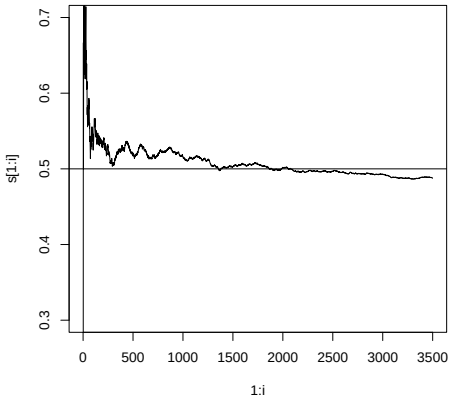
Betrachte für sukzessiv wachsendes n:

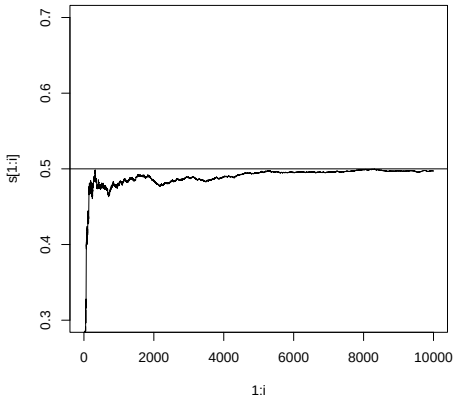
- $X_1, \ldots, X_n$ *i.i.d.*
- $X_i \in \{0,1\}$ Variablen mit $\pi = P(X_i = 1)$
- H_n, die relative Häufigkeit der Einsen in den ersten n Versuchen











Beobachtung:

Satz 1.51 *Theorem von Bernoulli*
Seien $X_1, \ldots, X_n,$ i.i.d. mit $X_i \in \{0, 1\}$ und $P(X_i = 1) = \pi$. Dann gilt für $H_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ (relative Häufigkeit der "Einsen") und beliebig kleines $\epsilon > 0$

$$\lim_{n \to \infty} P(|H_n - \pi| \leq \epsilon) = 1 \qquad (2.1.1)$$

Anschaulich:

Zwei wichtige Konsequenzen

1) Häufigkeitsinterpretation:

2) induktiv:

Satz 1.52 *Schwaches Gesetz der großen Zahl*
Gegeben seien $X_1, \ldots, X_n,$ i.i.d. Zufallsvariablen mit (existierendem) Erwartungswert μ und (existierender) Varianz σ^2. Dann gilt für $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ und beliebiges $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \leq \epsilon) = 1 \qquad (2.2.2)$$

Schreibweise: $\bar{X}_n \xrightarrow{P} \mu$ („Stochastische Konvergenz“, „X_n konvergiert in Wahrscheinlichkeit gegen μ.)
Konsequenz:

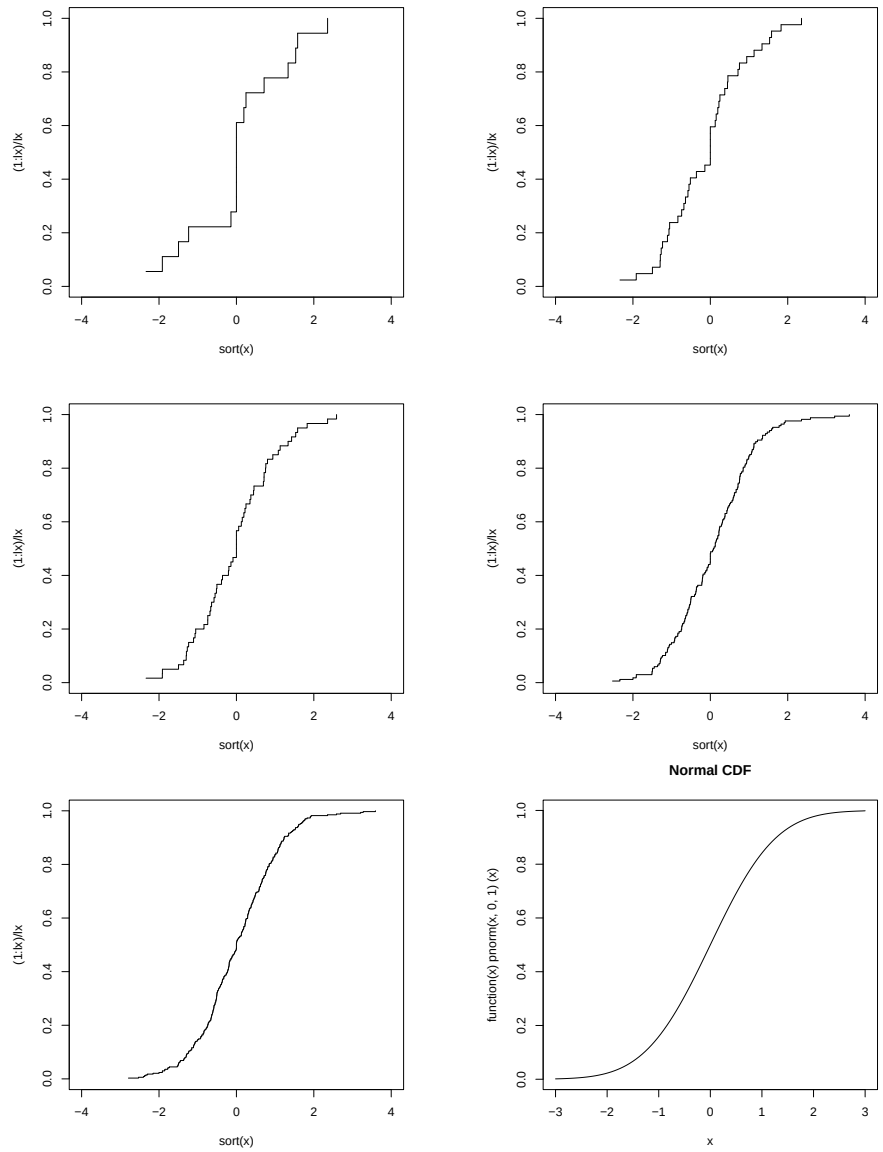
1.7.3 Der Hauptsatz der Statistik

Satz 1.53 *Hauptsatz der Statistik*:
$X_1, \ldots, X_n$ i.i.d. mit Verteilungsfunktion F
Sei $F_n^{X_1,\ldots,X_n}(x)$ die empirische Verteilungsfunktion der ersten n Beobachtungen und

$$D_n := \sup_x |F_n^{X_1,\ldots X_n}(x) - F(x)|, \qquad (2.2.4)$$

so gilt für jedes $c > 0$

$$\lim_{n\to\infty} P(D_n > c) = 0 \; . \qquad (2.2.5)$$


(1:k)/k
sort(x)
Normal CDF
function(x) pnorm(x, 0, 1) (x)
x

1.7.4 Der zentrale Grenzwertsatz

- Gibt es für große Stichprobenumfänge Regelmäßigkeiten im Verteilungstyp?
- Gibt es eine Standardverteilung, mit der man oft bei großen empirischen Untersuchungen rechnen kann?

Satz 1.54 *Zentraler Grenzwertsatz*
Seien $X_1, \ldots, X_n$ i.i.d. mit $\mathbb{E}(X_i) = \mu$ und $Var(X_i) = \sigma^2 > 0$
und

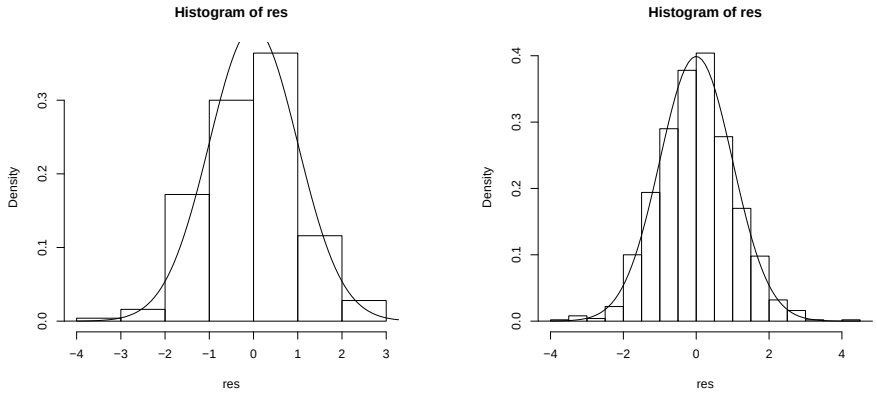
$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right) \qquad (2.3.1)$$

Dann gilt:
Z_n ist *asymptotisch standardnormalverteilt*, in Zeichen: $Z_n \overset{asym}{\sim} N(0;1)$, d.h. es gilt für jedes z

$$\lim_{n \to \infty} P(Z_n \leq z) = \Phi(z). \qquad (2.3.2)$$

Zurück zu der Eingangsfrage:

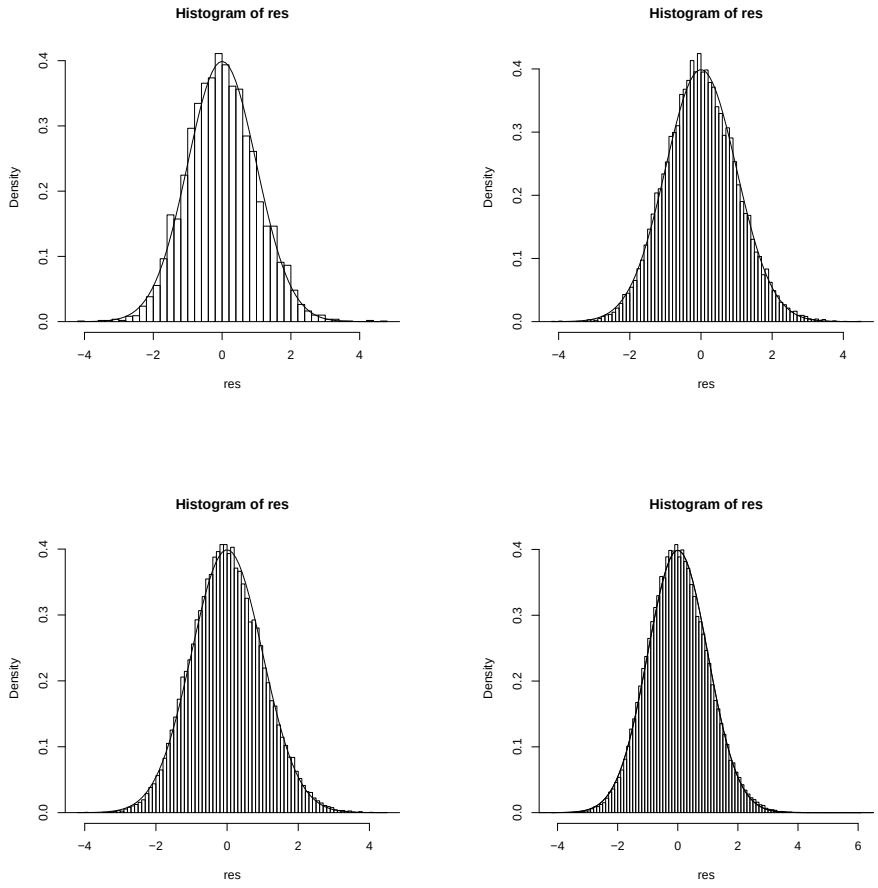
Ja, wenn man die Variablen geeignet mittelt und standardisiert, dann kann man bei großem n näherungsweise mit der Normalverteilung rechnen. Dabei ist für festes n die Approximation umso besser, je „symmetrischer“ die ursprüngliche Verteilung ist.



Bem 1.55:

Interessant ist auch die Anwendung auf $\bar{X}$ als Funktion von n.
Gemäß dem Gesetz der großen Zahlen weiß man:

$$\bar{X}_n \longrightarrow \mu$$



Für die Praxis ist es aber zudem wichtig, die konkreten Abweichungen bei großem aber endlichem n zu quantifizieren, etwa zur Beantwortung folgender Fragen:

i) Gegeben eine Fehlermarge ε und der Stichprobenumfang n: Wie groß ist die Wahrscheinlichkeit, dass $\bar{X}$ höchstens um ε von μ abweicht?

ii) Gegeben eine Fehlermarge ε und eine „Sicherheitswahrscheinlichkeit“ γ: Wie groß muss man n mindestens wählen, dass mit mindestens Wahrscheinlichkeit γ das Stichprobenmittel höchstens um ε von μ abweicht *(Stichprobenplanung)*?

Gemäß Satz 1.54 ist

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right) = \frac{\sum_{i=1}^{n}X_i - n\mu}{\sqrt{n}\cdot\sigma} \overset{\text{asympt.}}{\sim} N(0,1)$$

$$\sum_{i=1}^{n}X_i = n\cdot\bar{X}_n$$

$$\frac{n\bar{X}_n - n\mu}{\sqrt{n}\cdot\sigma} = \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \overset{\text{asympt.}}{\sim} N(0,1)$$

oder auch

$$\bar{X}_n \overset{\text{asympt.}}{\sim} N(\mu,\frac{\sigma^2}{n}). \qquad (2.3.4)$$

Bem 1.56: **Wichtige Anwendung: Approximation der Binomialverteilung**

Aufgabe: Wende den zentralen Grenzwertsatz auf die Binomialverteilung an.

a) $X \sim B(n,\pi)$; kann man die Verteilung von X approximieren?

X ist die Anzahl der Treffer in einer *i.i.d.* Folge $Y_1,\dots Y_n$ von Einzelversuchen, wobei

$$Y_i = \begin{cases} 1 \text{ Treffer} \\ 0 \text{ kein Treffer} \end{cases}, \quad (Y_i \sim Bin(1,\pi))$$

$$X = \sum_{i=1}^{n} Y_i, \quad \mathbb{E}(Y_i) = \pi, \quad Var(Y_i) = \pi\cdot(1-\pi).$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{Y_i-\pi}{\sqrt{\pi(1-\pi)}}\right) \overset{\text{asympt.}}{\sim} N(0,1)$$

$$\frac{1}{\sqrt{n}}\frac{\sum Y_i - n\cdot\pi}{\sqrt{\pi(1-\pi)}} \sim N(0,1)$$

$$\frac{\sum Y_i - n\cdot\pi}{\sqrt{n\cdot\pi(1-\pi)}} \sim N(0,1) \qquad \frac{X-n\cdot\pi}{\sqrt{n\cdot\pi(1-\pi)}} \sim N(0,1)$$

$$P(X\le x) \approx \Phi\left(\frac{x-n\cdot\pi}{\sqrt{n\cdot\pi(1-\pi)}}\right)$$

falls n groß genug.

b) Stetigkeitskorrektur

Durch die Approximation der **diskreten** Binomialverteilung durch die **stetige** Normalverteilung geht der diskrete Charakter verloren; man erhält als Approximation $P(X = x) \approx 0$ für jedes $x \in \mathbb{N}$, was gerade für mittleres n unerwünscht ist.
Deshalb: benutze

$$P(X \leq x) = P(X \leq x + 0.5)$$

bei ganzzahligem x.
Man erhält also als bessere Approximation

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \tag{2.3.5}$$

und

$$P(X = x) \approx \Phi\left(\frac{x + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) - \Phi\left(\frac{x - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \tag{2.3.6}$$

Es gibt verschiedene Faustregeln, ab wann diese Approximation gut ist:

$$\begin{array}{ll} \text{z.B.} & n \cdot \pi \geq 5 \text{ und } n \cdot (1-\pi) \geq 5 \\ \text{oder} & n \cdot \pi(1-\pi) \geq 9 \end{array}$$

Wichtig ist: Ob die Approximation hinreichend genau ist,

Bsp 1.57 (fiktiv)

Politiker S ist von einer gewissen umstrittenen Maßnahme überzeugt und überlegt, ob es taktisch geschickt ist, zur Unterstützung der Argumentation eine Mitgliederbefragung zu dem Thema durchzuführen.
Er wählt dazu 200 Mitglieder zufällig aus und beschließt, eine Mitgliederbefragung zu „riskieren“, falls er in der Stichprobe mindestens 52% Zustimmung erhält.

Wie groß ist die Wahrscheinlichkeit, in der Stichprobe mindestens 52% Zustimmung zu erhalten, obwohl der wahre Anteil nur 48% beträgt?

1.8 Mehrdimensionale Zufallselemente

1.8.1 Grundbegriffe

Def. 1.58 Betrachtet werden zwei eindimensionale Zufallselemente X und Y (zu demselben Zufallsexperiment). Die Wahrscheinlichkeit

$$P(X = x_i, Y = y_j) := P(\{X = x_i\} \cap \{Y = y_j\})$$

in Abhängigkeit von x_i und y_j heißt *gemeinsame Verteilung* der mehrdimensionalen Zufallsvariable $\binom{X}{Y}$ und der Variablen X und Y.
Randwahrscheinlichkeiten:

$$p_{i\cdot} = P(X = x_i) = \sum_{j=1}^{m} P(X = x_i, Y = y_j) \tag{3.1.2}$$

$$p_{\cdot j} = P(Y = y_j) = \sum_{i=1}^{k} P(X = x_i, Y = y_j) \tag{3.1.3}$$

bedingte Verteilungen:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \tag{3.1.5}$$

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} \tag{3.1.6}$$

stetiger Fall: Zufallsvariable mit zweidimensionaler Dichtefunktion $f(x, y)$:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \left(\int_c^d f(x, y) dy \right) dx$$

1.8.2 Kovarianz und Korrelation

Def 1.59 *Kovarianz*
X, Y Zufallsvariable. Dann heißt

$$\sigma_{X,Y} := Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \tag{3.2.1}$$

Kovarianz von X und Y
Rechenregeln:

- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$
- $Cov(X, Y) = Cov(Y, X)$ (3.2.4)
- Mit $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ ist
  $Cov(\tilde{X}, \tilde{Y}) = a_X \cdot a_Y \cdot Cov(X, Y)$ (3.2.5)
- $\mathbf{Var(X + Y)} = \mathbf{Var(X)} + \mathbf{Var(Y)} + \mathbf{2 \cdot Cov(X, Y)}$ (3.2.6)

Def. 1.60 Zwei Zufallsgrößen X, Y mit $Cov(X, Y) = 0$ heißen *unkorreliert*.

Bem 1.61 Beachte: Stochastisch unabhängige Zufallsvariablen sind unkorreliert; die Umkehrung gilt jedoch im allgemeinen nicht.

Def. 1.62 *Korrelationskoeffizient*
Gegeben seien zwei Zufallsgrößen X und Y. Dann heißt

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{VarX}\sqrt{VarY}} \qquad (3.2.7)$$

Korrelationskoeffizient von X und Y.

Eigenschaften des Korrelationskoeffizienten

i) Mit $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ ist $|\rho(\tilde{X}, \tilde{Y})| = |\rho(X, Y)|$.

ii) $-1 \leq \rho(X, Y) \leq 1$

iii) $|\rho(X, Y)| = 1 \Longleftrightarrow Y = aX + b$

iv) Ist $Var(X) > 0,\ Var(Y) > 0$, so ist
$\rho(X, Y) = 0$ genau dann, wenn $Cov(X, Y) = 0$.