Grundlagen zur Wahrscheinlichkeitsrechnung und Pseudo-Zufallszsahlenerzeugung

Alexander Asteroth

Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Fachbereich Informatik
Department of Computer Science

Grundbegriffe/Notationen (Glossar)

Schriftarten

Skalar	Ein Skalar $x \in \mathbb{R}$ wird mit einem Kleinbuchstaben geschrieben.
Vektor	Ein Vektor $\boldsymbol{x} \in \mathbb{R}^n$ wird mit einem fetten Kleinbuchstaben geschrieben.
Menge	Mengen (z.B. $\mathcal{A} \subseteq \mathbb{R}$) werden mit kaligraphischen Buchstaben geschrieben.
Funktion	Funktionsnamen (z.B. Zufallsvariable $X : \mathcal{S} \to \mathbb{R}$) beginnen mit einem Großbuchstaben (oder einem griechischen Kleinbuchstaben). Falls sie vektor-wertig sind werden sie fett geschrieben, ansonsten nicht.
Matrix	Für Matrizen $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ verwenden wir, wie für Vektoren, fette Schrift. Allerdings werden für Matrizen Großbuchstaben verwendet, für Vektoren Kleinbuchstaben.

- Der Datentyp ergibt sich so aus der Schriftart.
- **Ausnahme:** Vektorwertige Funktionen und Matrizen werden beide mit fetten Großbuchstaben geschrieben. Wenn möglich werden für Matrizen daher immer die Buchstaben $\boldsymbol{A} - \boldsymbol{D}$ verwendet.

Wahrscheinlichkeiten

Symbol	Begriff	Definition/Signatur/Beispiel
$\mathcal{S}$	Menge von Stichproben (Stichprobenraum/engl. sample space)	
$\mathcal{E}$	Menge von Ereignissen	$\mathcal{E} \subseteq \mathscr{P}(\mathcal{S})$
P	Wahrscheinlichkeitsmaß	$\mathrm{P} : \mathcal{E} \to [0,1]$
$(\mathcal{S}, \mathcal{E}, \mathrm{P})$	Wahrscheinlichkeitsraum	
	sicheres Ereignis	$\mathrm{P}(\mathcal{S}) = 1$
	unmögliches Ereignis	$\mathrm{P}(\emptyset) = 0$
	$\mathcal{A}$ und $\mathcal{B}$ unabhängig	$\mathrm{P}(\mathcal{A} \cap \mathcal{B}) = \mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B})$
$\mathrm{P}(\mathcal{A} \mid \mathcal{B})$	bedingte Wahrscheinlichkeit	$\mathrm{P}(\mathcal{A} \mid \mathcal{B}) = \frac{\mathrm{P}(\mathcal{A} \cap \mathcal{B})}{\mathrm{P}(\mathcal{B})}$

Zufallsvariablen

X	reelwertige Zufallsvariable	$X : \mathcal{S} \to \mathbb{R}$
F_X	Wahrscheinlichkeitsverteilung	$\mathrm{F}_X(\mathcal{O}) = \mathrm{P}(X^{-1}(\mathcal{O}))$
		$= \mathrm{P}(\{a \in \mathcal{O} \mid X(a) \in O\})$
f_X	Wahrscheinlichkeitsdichte	$\int_{\mathcal{S}} \mathrm{f}_X(x)\ dx = \mathrm{F}_X(\mathcal{S})$
	... und falls F diff.bar	$\mathrm{f}_X(x) = \mathrm{F}'_X(x)$
$X \sim \mathrm{f}$	Kurzschreibweise für	$\mathrm{f} = \mathrm{f}_X$
$X \sim Y$	Kurzschreibweise für	$\mathrm{f}_X = \mathrm{f}_Y$
f_{XY}	gemeinsame Dichte	$\int_{\mathcal{A}} \int_{\mathcal{B}} \mathrm{f}_{XY}(x,y)\ dxdy = \mathrm{P}(X^{-1}(\mathcal{A}) \cap Y^{-1}(\mathcal{B}))$
E, μ	Erwartungswert	$\mathrm{E}(X) = \int_{\mathbb{R}} x\, \mathrm{f}(x)\ dx$
		$\mathrm{E}(X) = \mu_X = \overline{x}$
Var, σ^2	Varianz	$\mathrm{Var}(X) = \sigma_X^2 = \mathrm{E}((X - \overline{x})^2)$
Cov	Kovarianz	$\mathrm{Cov}(X,Y) = \mathrm{E}((X - \overline{x})(Y - \overline{y}))$
Corr	Korrelation	$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$

Zufallsvariablen

	Randdichte (marginal density)	$\mathrm{f}(x) = \int_{\mathbb{R}} \mathrm{f}(x, y)\ dy$
	bedingte Dichte (conditional density)	$\mathrm{f}(x \mid y) = \frac{\mathrm{f}(x,y)}{\mathrm{f}(y)}$
	X, Y bedingt unabhängig geg. Z	$\mathrm{f}(x, y \mid z) = \mathrm{f}(x \mid z)\,\mathrm{f}(y \mid z)$
	Satz von Bayes	$\mathrm{f}(y \mid x) = \frac{\mathrm{f}(x \mid y)\,\mathrm{f}(y)}{\mathrm{f}(x)}$

Noch einmal . . . langsam — an einem Beispiel

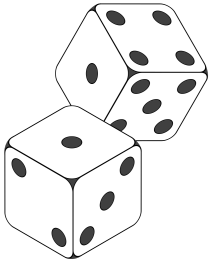
Würfeln

Alea iacta est! (sorry für das langweilige Beispiel)

Würfeln: Wahrscheinlichkeitsmaß

Ausgegangen wird von einem fairen Würfel. Wenn wir als Stichproben eines Zufallsexperiments die Anzahl der Punkte oben auf dem Würfel betrachten erhalten wir als Menge von Stichproben:

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$



Wenn wir zwei Würfel unabhäängig voneinander werfen, besteht der zugehörige Stichprobenraum aus allen Paaren (x, y) wo $x, y \in \mathcal{S}$. Daraus ergibt sich folgender Stichprobenraum:

$$\mathcal{S} \times \mathcal{S}$$

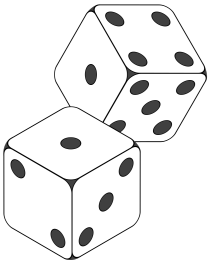
Würfeln: Ereignisraum

Ein Ereignis ist eine Untermenge von $\mathcal{S}$. Der Ereignisraum $\mathcal{E}$ ist eine Menge von Ereignissen

(i) $\mathcal{E}$ beinhaltet $\mathcal{S}$ $(\mathcal{S} \in \mathcal{E})$

(ii) ist unter dem Komplement abgeschlossen $(\mathcal{A} \in \mathcal{E} \implies \overline{\mathcal{A}} \in \mathcal{E})$

(iii) ist unter Vereinigung abgeschlossen $(\mathcal{A}, \mathcal{B} \in \mathcal{E} \implies (\mathcal{A} \cup \mathcal{B}) \in \mathcal{E})$



Wenn die Menge die Elementarereignisse $\{o_i\}$ für alle $o_i \in \mathcal{S}$ beinhaltet, dann folgt durch (i)-(iii) dass

$$\mathcal{E} = \mathscr{P}(\mathcal{S})$$

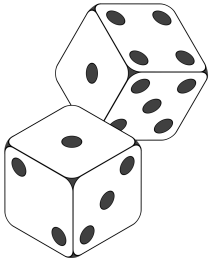
Im Fall von zwei Würfeln:

$$\mathcal{E} = \mathscr{P}(\mathcal{S}), \qquad \text{wobei } \mathcal{S} = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$$

Würfeln: Wahrscheinlichkeitsmaß

Ein Wahrscheinlichkeitsmaß ist eine Funktion

$$\mathrm{P} : \mathcal{E} \to [0, 1]$$



sodass für $\mathcal{A}, \mathcal{B} \in \mathcal{E}$

(i) $\mathrm{P}(\mathcal{S}) = 1$

(ii) $\mathrm{P}(\overline{\mathcal{A}}) = 1 - \mathrm{P}(\mathcal{A})$

(iii) wenn $\mathcal{A} \cap \mathcal{B} = \emptyset$ dann $\mathrm{P}(\mathcal{A} \cup \mathcal{B}) = \mathrm{P}(\mathcal{A}) + \mathrm{P}(\mathcal{B})$

Da von einem fairen Würfel ausgegangen wird, hat jedes Elementarereignis die selbe Wahrscheinlichkeit:

$$\mathrm{P}(\{1\}) = \cdots = \mathrm{P}(\{6\}) = \frac{1}{6}$$

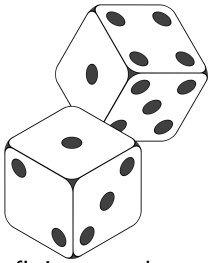
Wahrscheinlichkeiten von zusammengesetzten Ereignissen wie $\{2, 3\}$ (zwei oder drei Punkte) können, da diese disjunkt sind, von Elementarereignissen abgeleitet werden (iii):

$$\mathrm{P}(\{2, 3\}) = \mathrm{P}(\{2\}) + \mathrm{P}(\{3\}) = \frac{1}{3}$$

Würfeln: Zufallsvariablen und Verteilungen

Eine reellwertige Zufallsvariable X ist eine Funktion

$$X : \mathcal{S} \to \mathbb{R}$$



sodass das Urbild von Untermengen von $\mathbb{R}$ den Ereignissen entspricht:

$$\mathcal{A} \subseteq \mathbb{R} \implies X^{-1}(\mathcal{A}) \in \mathcal{E}$$

Dadurch kann die Wahrscheinlichkeitsverteilung F_X von X über $\mathbb{R}$ definiert werden durch

$$\mathrm{F}_X(\mathcal{A}) = \mathrm{P}(X^{-1}(\mathcal{A}))$$

Wenn das Bild $\mathrm{Im}(X)$ endlich/diskret ist nennen wir X eine endliche/diskrete Zufallsvariable (ZV).
Im Fall vom Werfen zweier Würfel können wir eine ZV X definieren indem jedes Paar auf die Summe der Punkte abgebildet wird.

$$X : \mathcal{S} \to \mathbb{R}, \quad \mathrm{Im}(X) = \{2, \ldots, 12\}, \quad \mathcal{S} = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$$

Würfeln: Zufallsvariablen, Verteilungen, Notation

Betrachten wir wieder den Wurf von zwei Würfeln, so definieren wir eine ZV X durch Abbilden jedes Paares auf die Summe der Punkte.

$$X : \mathcal{S} \to \mathbb{R}, \quad \mathrm{Im}(X) = \{2, \ldots, 12\}, \quad \mathcal{S} = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$$

Die Verteilung $\mathrm{F}_X(\mathcal{A})$ misst die Wahrscheinlichkeit von X durch das Generieren eines Wertes aus $\mathcal{A}$. Das wird häufig beschrieben durch:

$\mathrm{P}(X \in \mathcal{A})$ z.B. $\mathrm{P}(X \in [5, 8])$

Wenn $\mathcal{A}$ ein Intervall $\mathcal{A} = [a, b]$ ist, dann wird statt $\mathrm{F}_X(\mathcal{A})$ oft die Notation

$\mathrm{P}(a \leq X \leq b)$ z.B. $\mathrm{P}(5 \leq X \leq 8)$

verwendet. Wenn $\mathcal{A} = \{a\}$ eine einelementige Menge ist, dann wird $\mathrm{F}_X(\mathcal{A})$ beschrieben als:

$\mathrm{P}(X = a)$ z.B. $\mathrm{P}(X = 8)$

Würfeln: Zufallsvariablen und Verteilungsfunktion

Oft ist die Verteilung einer Zufallsvariablen nicht für beliebige Intervalle, sondern für Intervalle der Form definert:

$$\mathcal{A} = (-\infty, x]$$

In diesem Fall definiert die rechte Grenze von $\mathcal{A}$ die gesamte Menge $\mathcal{A}$ und wir können F_X definieren durch

$$\mathrm{F}_X(x) = \mathrm{F}_X((-\infty, x]) = \mathrm{P}(X^{-1}((-\infty, x]))$$

Wenn F so definiert ist wird die Verteilung als *kumulative Verteilungsfunktion* (CDF) von X bezeichnet.
In unserem Beispiel:

$$\mathrm{F}_X(8) \approx 0.72$$

gibt die Wahrscheinlichkeit an, dass die Summe von zwei Würfeln einen Wert kleiner oder gleich 8 annimmt.

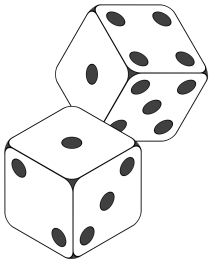
Würfeln: Zufallsvariablen und Dichte

Die Dichte einer Zufallsvariablen X ist eine Funktion f_X sodass

$$\mathrm{F}_X(\mathcal{A}) = \int_{\mathcal{A}} \mathrm{f}_X(x)\ dx \tag{1}$$

oder äquivalent

$$\mathrm{F}_X(x) = \int_{-\infty}^{x} \mathrm{f}_X(x')\ dx' \tag{2}$$



Anmerkung

Da das Integral immer zu einer stetigen Funktion führt, existiert f_X **nicht** wenn F_X nicht stetig ist.

Würfeln: Zufallsvariablen und Dichte

Es gilt wieder $X : \mathcal{S} \to \mathbb{R}$ die Summe der Augenzahl von zwei Würfeln

$$\mathcal{S} = \{1, \dots, 6\}^2$$

es gibt 36 mögliche Ausgänge und für alle Stichproben $(x, y) \in \mathcal{O}$

$$\mathrm{P}((x, y)) = \frac{1}{36}$$

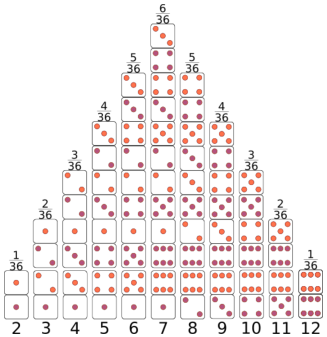
Würfeln: Zufallsvariablen und Dichte

Es gilt wieder $X : \mathcal{S} \to \mathbb{R}$ die Summe der Augenzahl von zwei Würfeln

$$\mathcal{S} = \{1, \dots, 6\}^2$$

es gibt 36 mögliche Ausgänge und für alle Stichproben $(x, y) \in \mathcal{O}$

$$\mathrm{P}((x, y)) = \frac{1}{36}$$



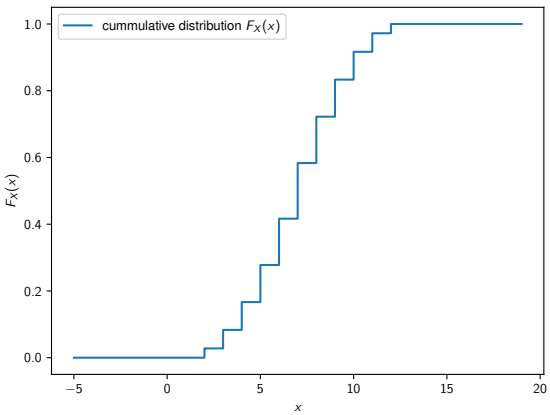
Würfeln: Zufallsvariablen und Dichte

Es gilt wieder $X : \mathcal{S} \to \mathbb{R}$ die Summe der Augenzahl von zwei Würfeln

$$\mathcal{S} = \{1, \dots, 6\}^2$$

es gibt 36 mögliche Ausgänge und für alle Stichproben $(x, y) \in \mathcal{O}$

$$\mathrm{P}((x, y)) = \frac{1}{36}$$



Die kumulative Verteilungsfunktion (CDF) $\mathrm{F}_X(x)$ ist rechts abgebildet. **Es handelt sich nicht um eine kontinuierliche Funktion, daher hat X keine Dichte f_X.**

Würfeln: Wahrscheinlichkeitsvektor

X hat keine Dichte, die Wahrscheinlichkeiten sind verteilt über eine endliche Menge von diskreten Werten:

$$E = \mathrm{Im}(X) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Jeder der Werte entspricht einem Elementarergebnis, dessen Wahrscheinlichkeit gemessen werden kann:

$$p_i = \mathrm{F}_X(\{e_i\}) = \mathrm{P}(X^{-1}(\{e_i\})), \qquad e_i \in E$$

(hier haben wir die erste Version der Verteilung F angewendet, nicht die CDF)
Da $E = \mathrm{Im}(X)$ eine endliche Menge ist, weisen wir X den Wahrscheinlichkeitsvektor zu

$$\boldsymbol{p}_X = (p_1, \ldots, p_n)^T$$

In unserem Beispiel

$$\begin{aligned}\boldsymbol{p}_X &= \frac{1}{36}(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1)^T \\ &\approx (0.028, 0.056, 0.083, 0.111, 0.139, 0.167, 0.139, 0.111, 0.083, 0.056, 0.028)^T\end{aligned}$$

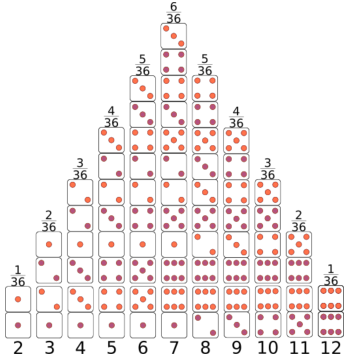
Würfeln: Erwartungswert und Varianz

Der Erwartungswert ist definiert als:

$$\mathrm{E}(X) = \int_{\mathbb{R}} x\, \mathrm{f}(x)\; dx$$

im Fall von diskreten Variablen wird aus dem Integral eine Summe, so ergibt sich für unser Beispiel:

$$\mu = \mathrm{E}(X) = \sum_{j=1}^{11} e_j p_j$$



Umsortieren der Terme

$$= (2+12)\frac{1}{36} + (3+11)\frac{2}{36} + \cdots$$
$$= 7 \sum p_j = 7$$

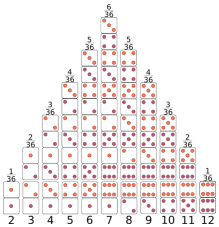
Würfeln: Erwartungswert und Varianz

Varianz ist definiert als:

$$\mathrm{Var}(X) = \mathrm{E}((X - \mu)^2)$$

in unserem Beispiel:

$$\sigma^2 = \mathrm{Var}(X) = \sum_{j=1}^{11} p_j (e_j - 7)^2$$



Umsortieren der Terme

$$= 2 \cdot 5^2 \frac{1}{36} + 2 \cdot 4^2 \frac{2}{36} + 2 \cdot 3^2 \frac{3}{36} + 2 \cdot 2^2 \frac{4}{36} + 2 \cdot 1^2 \frac{5}{36} + 2 \cdot 0^2 \frac{6}{36}$$

$$= \frac{50 + 64 + 54 + 32 + 10}{36} = \frac{210}{36} \approx 5.8$$

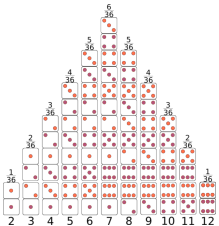
Würfeln: Varianz und Standardabweichung

i

$$\sigma^2 = \frac{210}{36} \approx 5.8$$

Standardabweichung:

$$\sigma = \sqrt{210/36} \approx 2.4$$



Generell gilt 68% der Strichproben X werden in das Intervall $[\mu - \sigma, \mu + \sigma] = [4.6, 9.4]$ fallen.
Wir können von der Abbildung rechts ableiten, dass $24/36 = 2/3 \approx 0.67$ tatsächlich in das oben angegebene Intervall fallen.
Darüberhinaus fallen 95% der Stichproben von X in ein 2σ Intervall und 99.7% in ein 3σ Intervall.

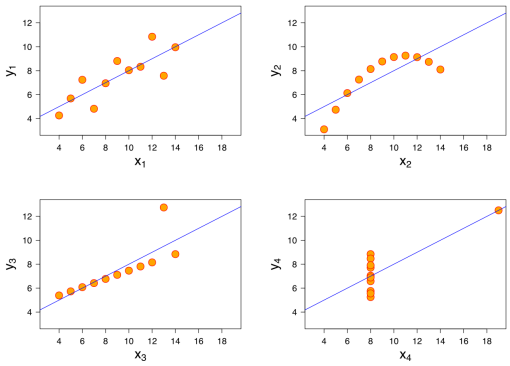
Was machen wir nun mit diesen Konzepten?

1. Gegeben Daten: bestimme passende Verteilung (z.B. maximum likelihood/max. aposteriori)
2. Gegeben einen randomisierten Algorithmus: bestimme Verteilung der Ergebnisse (probabilistische Analyse)
3. Gegeben eine Verteilung: bestimme deren Eigenschaften (z.B. Mittelwert Varianz, ...)
4. Gegeben eine Verteilung: ziehe Stichproben

Warum sollte man Stichproben ziehen?

Um einen randomisierten Algorithmus zufällige Entscheidungen treffen zu lassen.

Warum sollte man Stichproben ziehen?



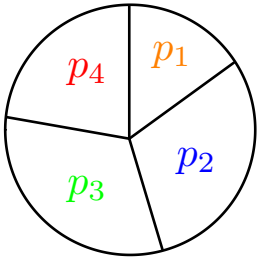
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	± 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3 + 0.5x$	to 2 decimal places
R^2	0.67	to 2 decimal places

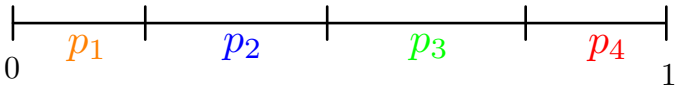
(info about Anscombe's quartet from Wikipedia)

(Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.)

Würfeln: Stichproben (roulette wheel selection)

Angenommen eine RV X und ihr Wahrscheinlichkeitsvektor $\boldsymbol{p}_X = (p_1, \ldots, p_n)^T$ sind gegeben und wir wollen eine Stichprobe x von X ziehen.





Die Gleichverteilung $\mathrm{U}(a, b)$

Wenne eine Zufallsvariable X auf ein Intervall $[a, b]$ mit konstanter Dichte auf dem ganzen Intervall abbildet (und Null außerhalb des Intervalls), nennen wir X gleichverteilt auf $[a, b]$ und schreiben dies als:

$$X \sim \mathrm{U}(a, b)$$

Example 1.

Oft gilt $a = 0, b = 1$. In diesem Fall gilt

$$\mathrm{f}_X(x) = 1 \text{ für alle } x \in [0, 1] \text{ und } \mathrm{f}_X(x) = 0 \text{ sonst}$$

Die Verteilung einer solchen Zufallsvariablen auf dem Intervall $[0, a]$ für ein $a \in [0, 1]$ ist linear in $[0, 1]$

$$\mathrm{F}_X([0, a]) = a$$

Allgemein gilt, dass die Dichte von $X \sim \mathrm{U}(a, b)$ konstant auf $[a, b]$ und die Verteilung linear auf $[a, b]$ ist.

Die Gleichverteilung $\mathrm{U}(a, b)$

Formaler:

$$f(x) = \mathrm{U}(a, b)(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{sonst} \end{cases}$$

Würfeln: Ziehen von Stichproben (roulette wheel selection)

Angenommen, eine Zufallsvariable X und der passende Wahrscheinlichkeitsvektor $\boldsymbol{p}_X = (p_1, \ldots, p_n)^T$ seien gegeben, und wir wollen eine Stichprobe x aus X ziehen. Ein Ansatz das zu tun ist:

1. teile $[0, 1]$ in n Unterintervalle I_j der Länge p_j
2. ziehe eine Stichprobe $T \sim \mathrm{U}(0, 1)$ aus der Gleichverteilung[1] auf $[0, 1]$, Bezeichne die Stichprobe als t
3. finde den Index j sodass $t \in I_j$
4. setze $x = e_j$

Die Aufteilung von $[0, 1]$ in Unterintervalle und das anschließende gleichmäßige Ziehen aus $[0, 1]$, sind änlich wie das Konzept eines Roulette-Rades. Diese Methode ist daher auch bekannt als *roulette wheel selection*.

[1] Pseudo-Zufallszahlengeneratoren für diese Verteilung gibt es in allen Programmiersprachen

Würfeln: Ziehen von Stichproben (roulette wheel selection)

In unserem fortlaufenden Beispiel:

$$E = \mathrm{Im}(X) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$
$$\boldsymbol{p}_X = \frac{1}{36}(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1)^T$$

1. teile $[0, 1]$ in Unterintervalle

$$\begin{array}{rcccccccccc} [0,1] & = & I_1 & \cup & I_2 & \cup & I_3 & \cup & \cdots & \cup & I_{11} \\ & = & [0, \frac{1}{36}) & \cup & [\frac{1}{36}, \frac{3}{36}) & \cup & [\frac{3}{36}, \frac{6}{36}) & \cup & \cdots & \cup & [\frac{35}{36}, 1] \end{array}$$

2. ziehe eine zufällige Zahl aus $T \sim \mathrm{U}(0, 1) \qquad \longrightarrow \qquad t = 0.3145$
3. finde den Index j sodass $t \in I_j$:

$$0.3145 \in \left[\frac{6}{36}, \frac{10}{36}\right) = I_4 \qquad \Longrightarrow \qquad j = 4$$

4. setze

$$x = e_j = e_4 = 5$$

Jetzt wieder zurück zu kontinuierlichen Verteilungen, für die Dichten existieren.

Ziehen von Stichproben bei nicht-diskreten ZV

Angenommen eine Zufallsvariable X mit Dichte $\mathrm{f}(x)$ ist gegeben und wir wollen eine Stichprobe daraus ziehen. Beginnen wir mit folgender Beobachtung – führe eine neue ZV T ein

$$T = \mathrm{F}(x)$$

wie ist T auf seinem Bild $\mathrm{Im}(T) = [0, 1]$ verteilt[2]? Sei $a \in [0, 1]$:

$$\begin{aligned} \mathrm{F}_T(a) = \mathrm{P}(T \leq a) &= \mathrm{P}(\mathrm{F}(x) \leq a) \\ &= \mathrm{P}(X \leq \mathrm{F}^{-1}(a)) = \mathrm{F}(\mathrm{F}^{-1}(a)) = a \end{aligned}$$

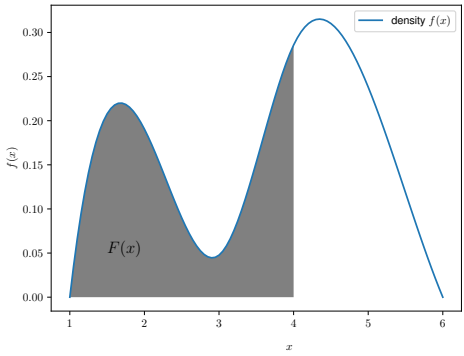
Daher ist T (die Verteilung von X) gleichverteilt in $[0, 1]$.

[2] unter der Annahme, dass F eine streng ansteigende Funktion in $[0, 1]$ so ist F^{-1} wohl definiert auf $[0, 1]$ (andernfalls wird der Beweis nur viel komplizierter)

Ziehen von Stichproben bei nicht-diskreten ZV

Dies führt zu der folgenden Methode:

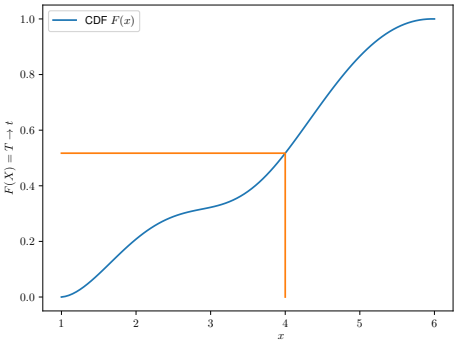
- berechne aus der Dichte $\mathrm{f}(x)$ die Verteilung $\mathrm{F}(x)$

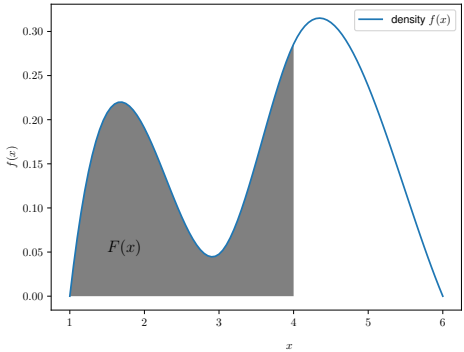


Ziehen von Stichproben bei nicht-diskreten ZV

Dies führt zu der folgenden Methode:

- berechne aus der Dichte $\mathrm{f}(x)$ die Verteilung $\mathrm{F}(x)$





- ziehe eine Stichprobe aus $\mathrm{U}(0, 1)$, nenne sie t
- finde x sodass $t = \mathrm{F}(x)$

How to sample $T \sim \mathrm{U}(0,1)$? — PRNG's

Let's start with an example of a *pseudo random number generators*[3], (*PRNG*):
Ansi C function `rand()` has the following implementation

$$
\begin{aligned}
s_0 &= 12345 \\
s_{i+1} &= 1103515245 s_i + 12345 \mod 2^{31}, \quad i = 0, 1, \ldots
\end{aligned}
$$

such a PRNG is calld a *linear congruential generator* (LCG).

[3] It is also possible to create *real* random numbers (TRNG). (see `/dev/random`, `/dev/urandom`).

Linear congruential generators

LCG's in general have the form

$$s_{i+1} = as_i + b \mod c$$

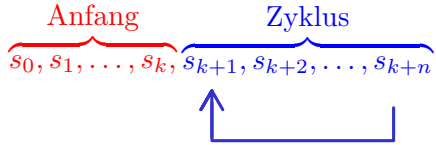
For efficiency reasons c is often chosen as a power of 2.
Let's consider a minimalisitic example:

$$s_0 = 3$$
$$s_{i+1} = as_i \mod 8$$

and investigate the influence of a on the generated sequence:

$$a = 2 \qquad (s_i) = 3, 6, 4, 0, 0, \ldots$$
$$a = 3 \qquad (s_i) = 3, 1, 3, 1, 3, \ldots$$
$$a = 4 \qquad (s_i) = 3, 4, 0, 0, 0, \ldots$$
$$a = 5 \qquad (s_i) = 3, 7, 3, 7, 3, \ldots$$
$$a = 6 \qquad (s_i) = 3, 2, 4, 0, 0, \ldots$$
$$a = 7 \qquad (s_i) = 3, 5, 3, 5, 3, \ldots$$

In general the generated sequence of random numbers has the following structure:



(Anfang = beginning, Zyklus=cycle)
The quality of a PRNG depends of the length and structure of the cycle. In particular:

a) length of the cycle
b) distrubution of numbers in cycle
c) autocorrelation of the sequence
d) cryptographically secure RNG's (will not be considered here)

While a) and b) should be more or less obvious, c) needs some explanation. Assume a LCG generates integer valued random numbers from $[0, c-1]$ ($\mod c$).Then the longest possible cycle is c elements long. But if this cycle is

$$0, 1, 2, 3, \ldots, c-1$$

we would not consider it very random. Therefore the autocorrelation of the sequence can help to find out about this kind of structure in the sequence.
If noncyclic seqences are of interest, approximations to irrational numbers are an option.

Pseudo Random Number Generators

. . . have the structure:

$$s_0 = \text{seed}, \quad s_{i+1} = f(s_i, s_{i-1}, \ldots)$$

Linearly Congruential Generators (LCG's)

An LCG generates the next element of the sequence s_{i+1} solely based on the previous s_i:

$$s_{i+1} = as_i + b \mod c$$

often $c = 2^n$

For LCG's it holds that:

Lemma 2.

An LCG has a maximal cycle lenght of c, if

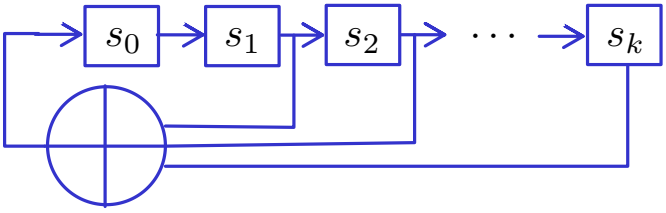
- c *and* b *are relatively prime*
- *every prime factor of* c *is also a prime factor of* $a - 1$

If $c = 2^n, a = 4k + 1$ for $k \in \mathbb{N}$ and b odd, above mentioned requirements are fulfilled Let

$$s_{i+1} = 13s_i + 5 \mod 2^{10}$$

then (s_i) becomes cyclic after exactly 1024 steps.

Linear Recursive Shift Register



Definition 3 (Linear Recursive Shift Register (LRSR)).

A LRSR (linear recursive shift register) with feedback weights $\boldsymbol{b} = (b_0, \ldots, b_k)$ and seed $\boldsymbol{s}_0 = (s_0^0, s_1^0, \ldots, s_k^0)$ is defined by it's register assignements at time $t+1$ according to:

$$s_{i+1}^{t+1} = s_i^t, \qquad i = 0 \ldots k-1, \qquad\qquad s_0^{t+1} = \bigoplus_{i=0}^{k} b_i s_i^t$$

The register can be considered either as a PRNG of numbers in the range $[0, 2^{k+1})$ or as a generator of a bit sequence (using an arbitrary bit).

Remark 1.

Often global behaviour of shift registers is good while local behaviour is "bad". Autocorrelation is usually nearly 0 but it is assumed that higher order autocorrelation can be bad.

Exercise. . .

Kurzschreibweise

Als Variable für die Werte der Zufallsvariablen X, Y, Z verwenden wir die Buchstaben x, y, z. Dies ermöglicht es uns, einige Indizees weg fallen zu lassen:

$$\mathrm{f}(x) = \mathrm{f}_X(x), \qquad \mathrm{F}(x) = \mathrm{F}_X(x), \qquad \mathrm{f}(y) = \mathrm{f}_Y(y), \qquad \dots$$

Unabhängig und identisch verteilte Zufallsvariablen (iid)

Oft beobachten wir das Ergebnis ein und der selben Zufallsvariable mehrere male. $x_i, \ldots, x_n$. Wenn wir z.B. etwas über die Summe dieser x_i herausfinden möchten, dann ist diese selbst wieder eine Zufallsvariable

$$Y = \sum_{i=1}^{n} X_i$$

Wobei alle Zufallsvariablen X_i hierbei die gleiche Verteilung haben wie die ursprüngliche Zufallsvariable X. Beachte, dass wir so viele Zufallsvariablen einführen mussten, wie wir Beobachtungen haben und nicht einfach $\sum_{i=1}^{n} X$ schreiben können, weil es sonst nicht eindeutig wäre, ob X hier einmal oder mehrere Male ausgewertet wurde.
Wir notieren den Sachverhalt, dass alle ZVn die gleiche Verteilung haben wie X als

$$X_i \sim X$$

und nennen X_i identisch verteilt.
Wenn zusätzlich alle ZVn $\mathrm{X_i}$ unabhängig sind, nennt man sie unbhängig und identisch verteilt (engl. *independent and identically distributed*) und schreibt dies als **iid**.

Theorem 4 (Gesetz der großen Zahlen).

Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt (iid) mit Mittelwert μ. Dann

$$\lim_{n\to\infty} \frac{1}{n}(X_1 + \cdots X_n) = \mu$$

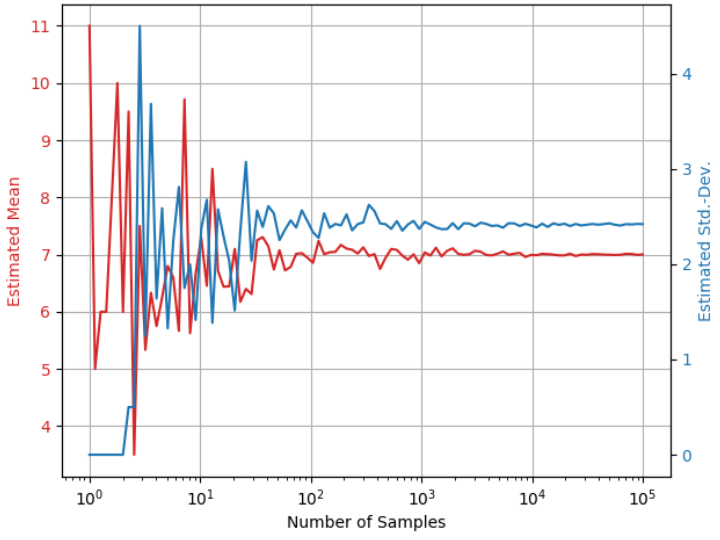
nahezu sicher[4]*.*

Example 5.

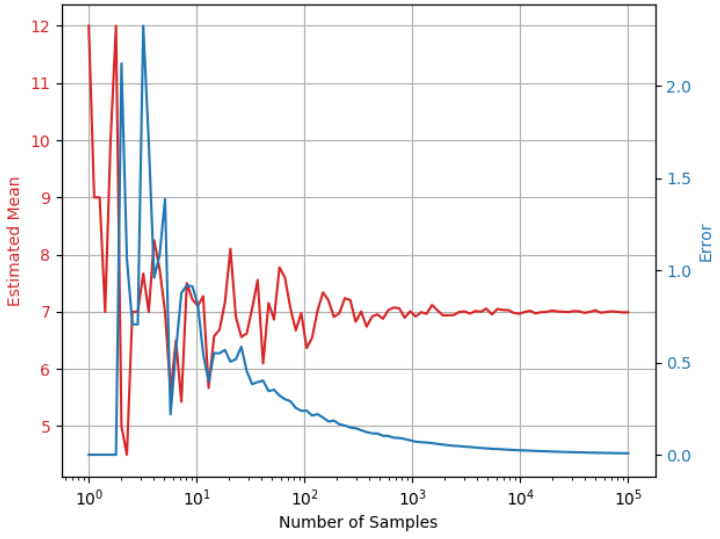
Das Gesetz der großen Zahlen wird angewendet, wenn μ, σ^2 und bekannt sind und nicht analytisch abgeleitet werden können. Stichproben werden von f gezogen und μ wird abgeleitet. Wenn μ bekannt ist, kann σ^2 auf die gleiche Weise abgeleitet werden.

[4] "nahezu sicher" bedeutet mit Wahrscheinlichkeit 1.

Example 6 (Schätzung von μ mit Hilfe des Gesetzes der großen Zahlen).



Example 6 (Schätzung von μ mit Hilfe des Gesetzes der großen Zahlen).



Abstandsmaße für Zufallsvariablen

Definition 7 (Kullback Leibler Divergenz).

Gegeben seien zwei diskrete Zufallsvariablen X und Y, dann ist die Kullback-Leibler Divergenz (d_{KL}) definiert als

$$d_{KL}(X||Y) = \sum_i p_i \log \frac{p_i}{q_i}$$

wobei $\boldsymbol{p}, \boldsymbol{q}$ die Wahrscheinlichkeitsvektoren der ZV X und Y sind und es gilt

$$0 \log \frac{0}{q} = 0, \text{ und wenn } p \neq 0,\ p \log \frac{p}{0} = \infty$$

im kontinuierlichen Fall wird sie definiert als:

$$d_{KL}(X||Y) = \int_{-\infty}^{\infty} \mathrm{f}_X(z) \log \frac{\mathrm{f}_X(z)}{\mathrm{f}_Y(z)}\ dz$$

Abstandsmaße für Zufallsvariablen

Anmerkungen:

- Wenn $p_i, q_i \neq 0$ ist die Berechnung von d_{KL} leicht:

```
kl = (p * numpy.log(p/q)).sum()
```

  Häufig wird ein kleiner Wert $\varepsilon > 0$ zu allen p_i, q_i hinzugefügt um die Behandlung von Spezialfällen zu vermeiden und eine effiziente Berechnung von d_{KL} zu erlauben.

- Beachte, dass d_{KL} keine symmetrische Funktion ist, wie man es von einer richtigen Abstandsmetrik erwarten würde. Daher wird oft die folgende Abstandsmetrik verwendet:

$$d_{2KL}(X||Y) := d_{KL}(X||Y) + d_{KL}(Y||X)$$

- Wenn X, Y kontinuierliche ZV's sind und wenn das Integral nicht gelöst werden kann, wird gewöhnlich $\mathrm{f}_X, \mathrm{f}_Y$ gesampled und d_{KL} genauso bestimmt wie für diskrete ZV's.

Die Standardnormalverteilung

Definition 8 (Verteilung einer Zufallsvariablen).

Wir notieren eine Zufallsvariable mit der Dichtefunktion f durch

$$X \sim \mathrm{f}(x) \qquad (\text{d.h. } \mathrm{f}_X = \mathrm{f})$$

Definition 9 (Normalverteilung).

Eine *normal verteilte Zufallsvariable X hat eine Dichte*

$$\mathrm{f}(x) = \mathrm{N}(x_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{\sigma^2}(x-x_0)^2}$$

also ist X normal verteilt (mit Mittelwert x_0 und Varianz σ^2) genau dann wenn

$$X \sim \mathrm{N}(x_0, \sigma^2)$$

Was ist “normal” an der Normalverteilung

Theorem 10 (Zentraler Grenzwertsatz).

Seien $X_1, \ldots, X_n$ iid Zufallsvariablen mit der gleichen Dichte $X_i \sim \mathrm{f}$, Erwartungswert μ und Varianz σ^2. Die Zufallsvariable

$$Z_n := \frac{1}{\sigma\sqrt{n}} \left(X_1 + \cdots + X_n - n\mu\right)$$

konvergiert gegen eine standard normal verteilte Zufallsvariable:

$$Z = \lim_{n\to\infty} Z_n, \qquad Z \sim \mathrm{N}(0,1)$$

alternativ: der Durchschnittswert von n Erkenntnissen der selben Verteilung $X_i \sim \mathrm{f}$ konvergiert gegen eine normalverteilte ZV

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Zentraler Grenzwertsatz: Sensor Beispiel

Betrachten wir einen Roboter, der die Eingaben, welche durch eine Zufallsvariable Y realisiert werden, von einem verrauschten Sensor liest. Während der Roboter sich bewegt, liest er zu jedem Zeitpunkt t eine Realisierung von y_t von Y.
Es gibt unterschiedliche Arten von Rauschen (additiv, multiplikativ, Gauß'sch, nicht-Gauß'sch, etc.). Meistens wird Rauschen als additiv mit Mittelwert 0 angenommen, was wir vorerst tun werden:

$$Y = x + E$$

wobei x der echte Wert ist und E additives Rauschen unabhängig von x mit Mittelwert 0 und endlicher Varianz σ^2.
Nehmen wir an, wir können jede Sensormessung zum Zeitpunkt t n-mal wiederholen

$$y_t^{(j)} = x_t + e_t^{(j)}, \qquad 1 \leq j \leq n$$

und betrachten den Mittelwert

$$\overline{y_t} = \frac{1}{n}\sum_{j=1}^{n} y_t^{(n)} = \frac{1}{n}\sum_{j=1}^{n} x_t + e_t^{(j)} = x_t + \frac{1}{n}\sum_{j=1}^{n} e_t^{(j)} = x_t + \overline{e}$$

Zentraler Grenzwertsatz: Robotersensor Beispiel

$$\overline{y_t} = x_t + \overline{e}$$

Das Rauschen $\overline{e}$ ist eine Realisierung der Zufallsvariable

$$E_t = \frac{1}{n}\left(E_t^{(1)} + E_t^{(e)} + \cdots E_t^{(n)}\right)$$

Da $E_t^{(j)} \sim E$ iid mit Mittelwert 0 und Varianz σ^2 sind

$$E_t \sim \mathrm{N}\left(0, \frac{\sigma^2}{n}\right)$$

Daher kann der Rauschlevel σ^2 durch Wiederholung auf $\sqrt{\frac{\sigma^2}{n}}$ reduziert werden, also reduzieren n wiederholte Messungen den Rauschlevel um einen Faktor von $\frac{1}{\sqrt{n}}$.
Um additives zero mean Sensorrauschen auf 10% zu reduzieren brauchen wir 100 unabhängige Messungen. Für eine Reduktion auf 50% werden nur 4 Wiederholungen benötigt.

Definition 11 (geometrische Verteilung).

Ein Zufallsexperiment mit den Ausgängen "*Erfolg*" und "*Misserfolg*" wird so lange durchgeführt, bis "*Erfolg*" eintritt. Es resultiert ein Zufallsexperiment mit dem Stichprobenraum

$$\mathcal{S} = \{m^n e \mid n \in \mathbb{N}_0, m = \text{"Misserfolg"}, e = \text{"Erfolg"}\}$$

Die Wahrscheinlichkeiten für "*Erfolg*" sei konstant p. Definiere eine reellwertige Zufallsvariable $X(m^n e) = n + 1$. Dann gilt für das Bild $\mathrm{Im}(X) = \mathbb{N}$, und es gilt:

$$\mathrm{P}(X = t) = (1 - p)^{t-1} p$$

$\mathrm{P}(X = t)$ ist also eine geometrische Folge $(a_t) = (a_0 q^t)$ mit $a_0 = \frac{p}{1-p}$ und $q = (1 - p)$. Die Verteilung heißt daher *geometrische Verteilung*.

Satz 1.

Sei X eine geometrisch verteilte Zufallsvariable und $p \in (0,1)$, dann gilt:

$$E(X) = \frac{1}{p}$$

Beweis.

Sei $q = 1 - p$

$$
\begin{aligned}
E(X) &= \sum_k k \cdot \mathrm{P}(X = k) = \sum_k k q^{k-1} p \\
&= p \sum_k k q^{k-1} \\
&= p \cdot (q^0 + 2q^1 + 3q^2 + 4q^3 + \cdots) \\
&= p \cdot \quad (q^0 \quad + \quad 2q^1 \quad + \quad 3q^2 \quad + \quad 4q^3 \quad + \quad \cdots \quad) \\
&= p \cdot (\ (q^0 \quad + \quad q^1 \quad + \quad q^2 \quad + \quad q^3 \quad + \quad \cdots \quad) \\
&\qquad\qquad + \quad (q^1 \quad + \quad q^2 \quad + \quad q^3 \quad + \quad \cdots \quad) \\
&\qquad\qquad\qquad + \quad (q^2 \quad + \quad q^3 \quad + \quad \cdots \quad) \\
&\qquad\qquad\qquad\qquad\qquad\quad \cdots \qquad\qquad\quad) \\
&= p \cdot \sum_{1 \le k < \infty} \sum_{i=k}^{\infty} q^{i-1} = p \cdot \sum_{1 \le k < \infty} q^{k-1} \sum_{i=0}^{\infty} q^i \\
&= p \cdot \frac{1}{1-q} \sum_{k=0}^{\infty} q^k = p \cdot \frac{1}{1-q} \cdot \frac{1}{1-q} = \frac{1}{p}
\end{aligned}
$$

Satz 2 (Markoff'sche Ungleichung).

Sei $W = (\mathcal{S}, \mathcal{E}, \mathrm{P}), \mathcal{E} = \{e_1, e_2, \ldots\}$ *und* X *eine reellwertige Zufallsvariable,* $X \geq 0$. *Dann gilt für alle* $t > 0$

$$\mathrm{P}(X \geq t) \leq \mathrm{E}(X)/t$$

Beweis.

Wir definieren eine zweite Zufallsvariable:

$$Y(e_i) = \begin{cases} t & \text{if } X(e_i) \geq t \\ 0 & \text{otherwise} \end{cases}$$

expaction of this variable is:

$$E[Y] = 0 \cdot \mathrm{P}(X < t) + t \cdot \mathrm{P}(X \geq t) = t \cdot \mathrm{P}(X \geq t)$$
$$E(X) \geq E(Y) = t \cdot \mathrm{P}(X \geq t)$$

□

Satz 3 (Chebyscheff'sche Ungleichung).

Für alle $t > 0$ *ist*

$$\mathrm{P}(|X - \mathrm{E}(X)| \geq t) \leq \frac{\sigma_X^2}{t^2}$$

Beweis.

Sei $Y = |X - \mathrm{E}(X)|^2$. Mit der Markow'schen Ungleichung folgt:

$$\mathrm{P}(Y \geq t^2) \leq \frac{\mathrm{E}(Y)}{t^2}$$

Da $t > 0$ gilt:

$$Y = |X - \mathrm{E}(X)|^2 \geq t^2 \quad \Leftrightarrow \quad |X - \mathrm{E}(X)| \geq t \tag{3}$$

somit folgt die Behauptung

$$\mathrm{P}(|X - \mathrm{E}(X)| \geq t) \overset{(3)}{=} \mathrm{P}(Y \geq t^2) \leq \frac{|X - \mathrm{E}(X)|^2}{t^2} = \frac{\sigma_X^2}{t^2}$$

□

Satz 4 (Chernoff'sche Ungleichung).

Sei $0 < p < 1$ *und* $X = X_1 + \cdots + X_n$ *für unabhängige Indikatorvariablen* $X_i \in \mathbb{B}$ *mit* $\mathrm{P}(X_i = 1) = p$.
Dann ist $\mathrm{E}(X) = np$ *und für alle* $\delta \in (0, 1)$ *gilt:*

$$\mathrm{P}(X \leq (1 - \delta)\,\mathrm{E}(X)) \leq e^{-\,\mathrm{E}(X)\delta^2/2}$$

Beweis.

ohne Beweis (siehe Skript). □