# Stochastik Praktikum

Sebastian Holtz
Humboldt-Universität zu Berlin

Wintersemester 2018/19

# Organisatorisches

## Materialien & Kontakt

- Kurswebsite: unter www.math.hu-berlin.de/~holtz
- Sprechzeiten: nach der Übung & auf Anfrage

## Aufbau

- 16 Termine (10 vor Weihnachten)
- 4-5 Themenblöcke
- Theorie (Vorlesungen) und Anwendung (Programmierübungen)
- bewertete Abgaben zu jedem Block
- Abschlussnote: Abgaben und Diskussion

# Organisatorisches II

## Abgaben

- Bearbeitung und Abgabe in festen Teams
- Betreuung während der Praktikumstermine
- Abgabe von (kommentiertem) Code und Dokumentation

## Themen

- Zufallszahlen & Monte-Carlo-Simulationen
- Lineares Modell & Klassifikation
- Zeitreihen
- Simulation stochastischer Prozesse
- Markov-Chain-Monte-Carlo-Methoden

Eventuell: Nichtparametrische Verfahren, PCA, mehr Datenanalyse

# Literatur


- *Methoden der Statistik*, Buchprojekt von M. Jirak, K. Krenz, M. Reiß, M. Trabs, verfügbar auf der Teaching-Seite von M. Reiß
- *An Introduction to Statistical Learning: With Applications in R*, G. James, D. Witten, T. Hastie, R. Tibshirani
- *Handbook of Monte Carlo Methods*, D. Kroese, T. Taimre, Z.I. Botev
- *Introducing Monte Carlo Methods with R*, C. Robert, G. Casella
- *Monte Carlo Statistical Methods*, C. Robert, G. Casella

# Programmiersprachen

- S: kommerziell, Fokus auf Statistik
- R: kostenlos, Fokus auf Statistik, viele Bibliotheken
- Python: kostenlos, viele Anwendungen
- Julia: kostenlos, relativ neu (v1.0 im August 2018)

Viele weitere Sprachen und Anwendungen (Matlab, GAUSS, Stata,...)

Hier im Kurs: Benutzung von R

# Erzeugung von Zufallszahlen

# Zufallsgeneratoren

Ziel: Erzeuge für eine beliebige (bekannte) Verteilung $P^X$ Zahlen $X_1, \ldots, X_n \sim P^X$ i.i.d.

Herangehensweisen:

- physikalische Methoden (Atomzerfall, Hintergrundstrahlung)
- am Computer mittels 'Zufallsgeneratoren'

Wünschenswerte Eigenschaften von Zufallsgeneratoren:

- besteht viele statistische Tests (z.B. auf Verteilung),
- reproduzierbar,
- schnell, 'billig',
- lange Periode.

# Zufallsgeneratoren für $U(0,1)$

# Zufallsgeneratoren für $U(0,1)$

## Definition

Ein *Generator* (uniformer) *Pseudozufallszahlen* ist ein Algorithmus, der von einem Startwert $u_0$ (Seed) und einer Transformation $T$ ausgehend, eine rekursive deterministische Zahlenfolge $u_n = T^{\circ n}(u_0)$ ($[0,1]$-wertiger) Folgeglieder erzeugt, die sich wie eine zufällige i.i.d. Folge von echten (uniformen) Zufallszahlen verhalten soll.

Zwei Ansätze:

- Pseudo-Zufallszahlen: Eine Zahlenfolge, deren Bildungsgesetz möglichst schwer zu 'erraten' ist.
- Quasi-Zufallszahlen: Eine Zahlenfolge, deren Häufigkeitsverteilung gemäß eines vorgegebenen Abstandsbegriffs einer vorgegebenen Wahrscheinlichkeitsverteilung (i.d.R. $U(0,1)$) möglichst nahe kommt.

# Pseudo-Zufallszahlen

Viele Ansätze (s. Übersicht) mit unterschiedlichen Stärken und Schwächen.

Beispiel: Lineare Kongruenzgeneratoren

- Wähle $a, b, m \in \mathbb{N}$ fest sowie einen Seed $X_0 \in \{0, \dots, m-1\}$.
- Bildungsregel: $X_n = (aX_{n-1} + b) \bmod m$ sowie $U_n := \frac{X_n}{m} \in [0, 1]$.
- Periode $\leq m$, wobei Seed und $b$ nicht so wichtig sind.
- Können extrem schlechte Eigenschaften haben, z.B. (Randu) für $a = 65539 = 2^{16} + 3$ und $m = 2^{31}$. Dann ist (mit mod-Rechnung)

$$\begin{aligned} X_{n+2} &= (2^{16} + 3)X_{n+1} = (2^{16} + 3)^2 X_n \\ &= (2^{32} + 6 \cdot 2^{16} + 9)X_n = (6 \cdot 2^{16} + 9)X_n \\ &= (6 \cdot (2^{16} + 3) - 9)X_n = 6X_{n+1} - 9X_n. \end{aligned}$$

Aus dem Satz von Marsaglia folgt, dass die Tripel $(X_n, X_{n+1}, X_{n+2})$ stets auf einer von 15 Hyperebenen liegen.

# Pseudo-Zufallszahlen II

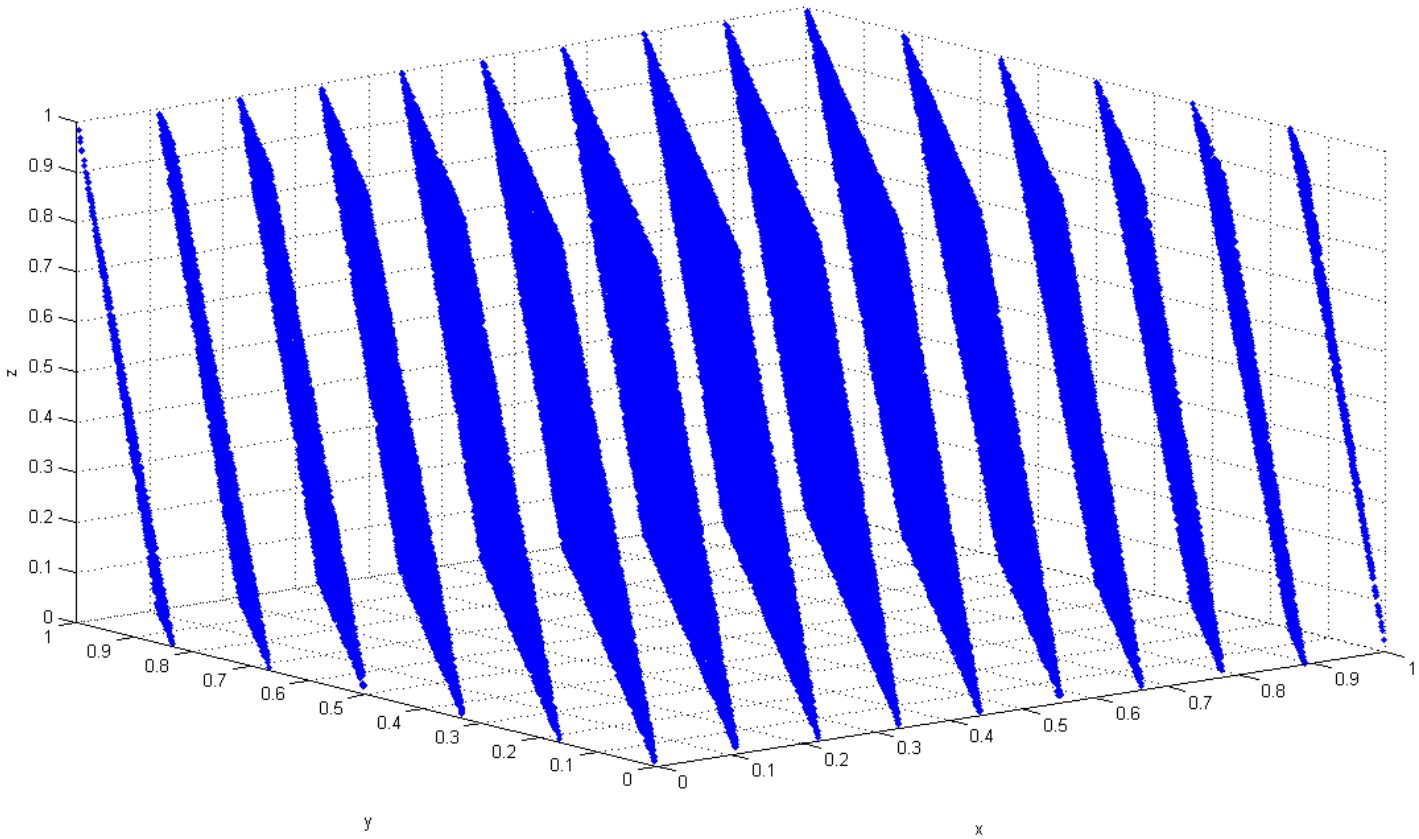


Figure: 100.000 mittels Randu generierte Zahlentripel (von Luis Sanchez mittels MATLAB, CC BY-SA 3.0, Link)

# Pseudo-Zufallszahlen III

Weite Anwendung findet der sogenannte Mersenne-Twister, da

- zuverlässig,
- besteht viele statistische Tests,
- Periode $2^{19937} - 1$,
- Standard in R, Python, Matlab, Excel, etc.

Weitere Ansätze: Kombination von Zufallsgeneratoren, andere moderne (robuste) Verfahren...

# Quasi-Zufallszahlen

Idee: $X_n$ basierend auf $X_1, \ldots, X_{n-1}$ minimiert die Diskrepanz

$$D_n(X_1, \ldots, X_n) := \sup_{u \in [0,1]} \left| \frac{|\{X_i : X_i \in [0, u)\}|}{n} - u \right|.$$

Beispiel: Halton-Folge

- Stelle $m \in \mathbb{N}$ durch fixe Primzahl $p$ (Basis) dar: $m = \sum_{j=0}^{k} a_j p^j$.
- Bilde die Halton-Zahlen $h_{m,p} := \sum_{j=0}^{k} a_j p^{-j-1}, \ m \geq 1$.
- Für $[0,1]^2$: Generiere für prime $p_1, p_2$ Tupel $(h_{m,p_1}, h_{m,p_2})_{m \geq 1}$.

# Vergleich von Quasi- & Pseudo-Zufall

- Pseudo-Zufallszahlen (Mersenne): $D_n \approx C' n^{-1/2}$.
- Quasi-Zufallszahlen (Halton): $D_n \leq C(\log(n)/n)$.

D.h. kleinere Fehler bei *Quasi*-MC-Integration. I.A. haben Pseudo-Zufallszahlen dennoch bessere Eigenschaften.

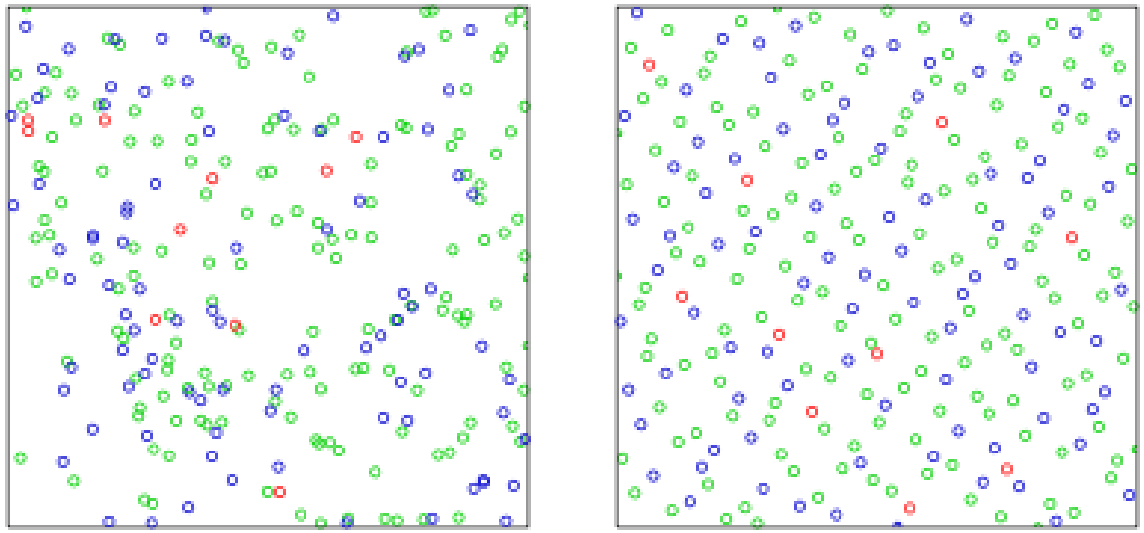

Figure: Pseudo-Zufallsfolge (links) und Halton-Folge in $[0,1]^2$ (von Jheald - Own work, CC BY-SA 3.0, Link1, Link2)

# Erzeugung von Zufallszahlen mittels $U(0,1)$

# 1. Inversionsmethode

## Definition

Sei $F : [0,1] \to \mathbb{R}$ eine Verteilungsfunktion. Die Funktion

$$F^{-1}(u) := \begin{cases} \inf\{x : F(x) \geq u\}, & u \in (0,1] \\ \sup\{x \in \mathbb{R} : F(x) = 0\}, & u = 0 \end{cases}$$

heißt Quantilfunktion oder verallgemeinerte Inverse von $F$.

Sei $X$ reellwertige Zufallsvariable mit Verteilungsfunktion $F$. Es gilt:

- Aus $U \sim U(0,1)$ folgt $F^{-1}(U) \sim X$,
- $F(X) \sim U(0,1)$.

Beachte: $F^{-1}$ ist i.A. nicht (explizit) verfügbar!

# 2. Diskrete Verteilungen auf $\mathbb{N}_0$

Bei diskreten Verteilungen sind $F$ und $F^{-1}$ i.A. bekannt.

Erzeugung von Zufallszahlen:

- Erzeuge $U \sim U(0,1)$,
- Setze $X := \sum_{k \geq 0} k \mathbb{1}\{P(X \leq k) < U < P(X \leq k+1)\}$.

Mitunter ist es ineffizient immer wieder bei $k = 0$ zu starten, z.B. für $X \sim \text{Poiss}(100)$ liegen die Werte fast ausschließlich in $[70, 130]$.
Mögliche Lösung: Ignoriere Bereiche mit kleiner Wahrscheinlichkeit.

# 3. Accept-Reject (Verwerfunsgmethode)

Ziel: Erzeuge $X \sim P_f$ zu gegebener Dichte $f$.

Idee: Wähle *Kandidatendichte* $g$, von der Zufallszahlen 'leicht' erzeugt werden können und mit $f(x) \leq\leq Mg(x),\ \forall x$, für eine Konstante $M > 0$.

Ablauf:

(1) Erzeuge $U \sim U(0,1)$ und $Y \sim P_g$ mit $U \perp Y$.

(2) Gilt $U \leq f(Y)/(Mg(Y))$ so setze $X := Y$, sonst zurück zu (1)

Beispiel: Erzeuge $X \sim \text{Beta}(4,3)$, d.h. $f(x) = 60x^3(1-x)^2$, mittels $Y \sim U(0,1)$, d.h. $g(y) = 1$. Die Konstante $M \approx 2.2$ erhält man via:

$$f(x) \leq Mg(x) \Longleftrightarrow 60x^3(1-x)^2 \leq M, \quad \forall x \in [0,1].$$

# 3. Accept-Reject (Verwerfunsgmethode) II

Mathematische Grundlage des Algorithmus:

$$P(X \leq x) = P\Big(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\Big) = \frac{P(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)})}{P(U \leq \frac{f(Y)}{Mg(Y)})}$$

$$= \frac{\int_{-\infty}^{x} \int_{0}^{f(y)/(Mg(y))} dug(y)dy}{\int_{-\infty}^{\infty} \int_{0}^{f(y)/(Mg(y))} dug(y)dy} = \frac{\frac{1}{M}\int_{-\infty}^{x} f(y)dy}{\frac{1}{M}\int_{-\infty}^{\infty} f(y)dy} = P_f((-\infty, x])$$

- Ist unabhängig von der Dimension, $M$ muss nicht scharf gewählt werden.
- Problem: Finden von $g$ (Intuitiv: Generierung $Y \sim P_g$ schwieriger wenn $M \to 1$).
- Akzeptanzwahrscheinlichkeit: $P(U \leq f(Y)/(Mg(Y))) = 1/M$, d.h. je kleiner $M$, desto weniger Samples werden abgelehnt

# 4. Zufallsvariablen auf Basis anderer Zufallsvariablen

- $U \sim U(0,1),\ a > 0 \Rightarrow aU + b \sim$
- $U_1, \ldots, U_n \sim U(0,1)$ i.i.d. $\Rightarrow X_i := 1(U \leq p) \sim$
- Box-Mueller-Methode: $U_1, U_2 \sim U(0,1)$ i.i.d. $\Rightarrow (X_1, X_2) := (\sqrt{-2\log(U_1)}\cos(2\pi U_2), \sqrt{-2\log(U_1)}\sin(2\pi U_2)) \sim$
- $X \sim \mathcal{N}(0, I_d),\ \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d} \Rightarrow \Sigma X + \mu \sim$
- $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$ i.i.d. $\Rightarrow \sum_{i=1}^n X_i^2 \sim$
- $X_1, \ldots, X_n \sim \mathrm{Exp}(1) \Rightarrow \sum_{i=1}^n X_i \sim$

# Statistische Tests für Zufallszahlen

# Statistische Tests für Zufallszahlen

Für erzeugte Zufallszahlen $X_1, \ldots, X_n$ soll entschieden werden, ob diese der angestrebten Verteilung $P^X$ hinreichend entsprechen.

Idee: Konstruiere einen statistischen Test basierend auf empirischer Verteilungsfunktion $F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i)$ und $F_X(x) := P^X((-\infty, x])$.

Für $D_n := \sup_x |F_n(x) - F(x)|$ gilt:

- $D_n \overset{\text{f.s.}}{\to} 0$ (Glivenko-Cantelli, s. Stochastik I)
- $\sqrt{n} D_n \overset{d}{\to} K$, $K$ folgt Kolmogorov-Verteilung (Stochastik II)

# Statistische Tests für Zufallszahlen II

Der sogenannte Kolmogorov-Smirnov-Test untersucht das Problem

$$\Theta_0 = \{F_X\} \quad \text{vs.} \quad \Theta_1 = \Theta \backslash \{F_X\},$$

wobei $\Theta$ die Menge aller Verteilungsfunktionen beschreibt.

Explizit: $\varphi(X_1, \ldots, X_n) = \mathbb{1}(\sqrt{n} D_n \geq c_\alpha)$, wobei $c_\alpha$ gemäß $P(\sqrt{n} D_n \geq c_\alpha) = \alpha$ für ein Niveau $\alpha \in (0, 1)$ gewählt wird.

In R: Aufruf mit `ks.test(x,y)` (z.B. mit y=pnorm) liefert einen p-Wert ('beobachtetes Signifikanzniveau').

# Monte-Carlo-Methode

# Monte-Carlo-Methode

Ursprung: Fermi, Ulam, von Neumann in den 1930er und 40er Jahren.

Idee: Verwendung vieler (meist identisch verteilter) Zufallsexperimente um mathematische Fragestellungen mittels der Stochastik numerisch zu lösen.

Anwendungen und Problemfelder:

- analytisch aufwendig/schwer lösbare Probleme,
- Verteilungseigenschaften von Zufallsvariablen unbekannten Typs,
- Nachahmung komplexer Systeme aufeinander wirkender Prozesse.

Mathematische Grundlage: Gesetz der großen Zahlen.

Ziel: Für eine Zufallsvariable $X$ mit Verteilung $P^X$ und reellwertige Funktionen $g$ werte Erwartungswerte der Form $\mathbb{E}[g(X)]$ aus.

Grobe Idee:

- Erzeuge i.i.d. Zufallsvariablen $X_1, \ldots, X_n \sim P^X$
- Bilde $S_n := \bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} g(X_i)$ als Schätzer von $\mathbb{E}[g(X)]$
- Nutze das starke Gesetz der großen Zahlen (sofern $g(X) \in L^1$):

$$S_n \overset{\text{f.s.}}{\to} \mathbb{E}[g(X)], \quad \text{für } n \to \infty.$$

Erste Fehlerbetrachtung: $S_n$ ist *erwartungstreu*, *konsistent* und für $\varepsilon > 0$ beträgt die Schätzabweichung (sofern $g(X) \in L^2$)

$$P\left(|S_n - \mathbb{E}[g(X)]| > \frac{\varepsilon}{\sqrt{n}}\right) \leq \frac{\mathrm{Var}(g(X))}{\varepsilon^2}.$$

# Beispiel: Monte-Carlo-Integration

Ziel: Berechne für reelwertige Funktionen $g$ Integrale der Form

$$\int_0^1 g(x)dx.$$

Idee:

- Nutze $\int_0^1 g(x)dx = \mathbb{E}[g(X)]$ für $X \sim U(0,1)$.
- Werte $\mathbb{E}[g(X)]$ mit Monte-Carlo-Simulationen aus.
- Vorteile: maximale Konvergenzrate $n^{-1/2}$ unabhängig von $g$ und der Dimension.
- Nachteil: nutzt eventuelle Glattheit von $g$ nicht aus.

# Fehlerbetrachtung der Monte-Carlo-Methode

Mittels ZGWS

- Seien $X_1, \ldots, X_n \sim P^X$ i.i.d., $\mu := \mathbb{E}[g(X)]$ sowie $\sigma_n^2 := \mathrm{Var}(S_n)$.
- Dann: $\sqrt{n}\left(\frac{S_n - \mu}{\sigma_n}\right) \xrightarrow{d} \mathcal{N}(0,1)$ und $P\left(\sqrt{n}\left|\frac{S_n - \mu}{\sigma_n}\right| \geq 2\right) \approx 0.05$.
- Verwende das Konfidenzband $\mu - 2\frac{\sigma_n}{\sqrt{n}} \leq S_n \leq \mu + 2\frac{\sigma_n}{\sqrt{n}}$.
- Ersetze für festes $n$ mittels empirischen Werten: $\mu \approx S_n$ und $\sigma_n^2 \approx \frac{1}{n}\sum_{i=1}^n (X_i - S_n)^2$.

Mittels Simulation der Varianz

- Führe $m$ MC-Simulationen durch und erhalte $S_n^{(1)}, \ldots, S_n^{(m)}$
- Nutze $\mu \approx \frac{1}{n}\sum_{k=1}^m S_n^{(k)}$ und $\sigma_n^2 \approx \frac{1}{m}\sum_{k=1}^m (S_n^{(k)} - \mu)^2$.

# Monte-Carlo-Bänder

Eindrücke über Abweichungen der empirischen Verteilung von Pseudo-Zufallszahlen zur zu modellierenden Verteilung $P^X$ liefern sogenannte Monte-Carlo-Bänder.

Idee:

- Bilde die empirischen Verteilungsfunktionen $S_n^{(1)}(x), \ldots, S_n^{(m)}(x)$ (d.h. Monte-Carlo-Simulationen mit $g(X) = \mathbb{1}(X \leq x)$).
- Vergleiche $\frac{1}{m} \sum_{k=1}^{m} S_n^{(k)}(x)$ mit $m(x) := \min_{1 \leq k \leq m} S_n^{(k)}(x)$ und $M(x) := \max_{1 \leq k \leq m} S_n^{(k)}(x)$.
- Allgemeiner können Bänder $x \mapsto (m_\alpha(x), M_\alpha(x))$ gemäß eines Konfidenzniveaus $\alpha \in (0, 1)$ erzeugt werden, so dass die relative Häufigkeit (empirische Wahrscheinlichkeit) für Werte außerhalb von $(m_\alpha(x), M_\alpha(x))$ nicht größer als $\alpha$ ist.

# Monte-Carlo-Bänder II

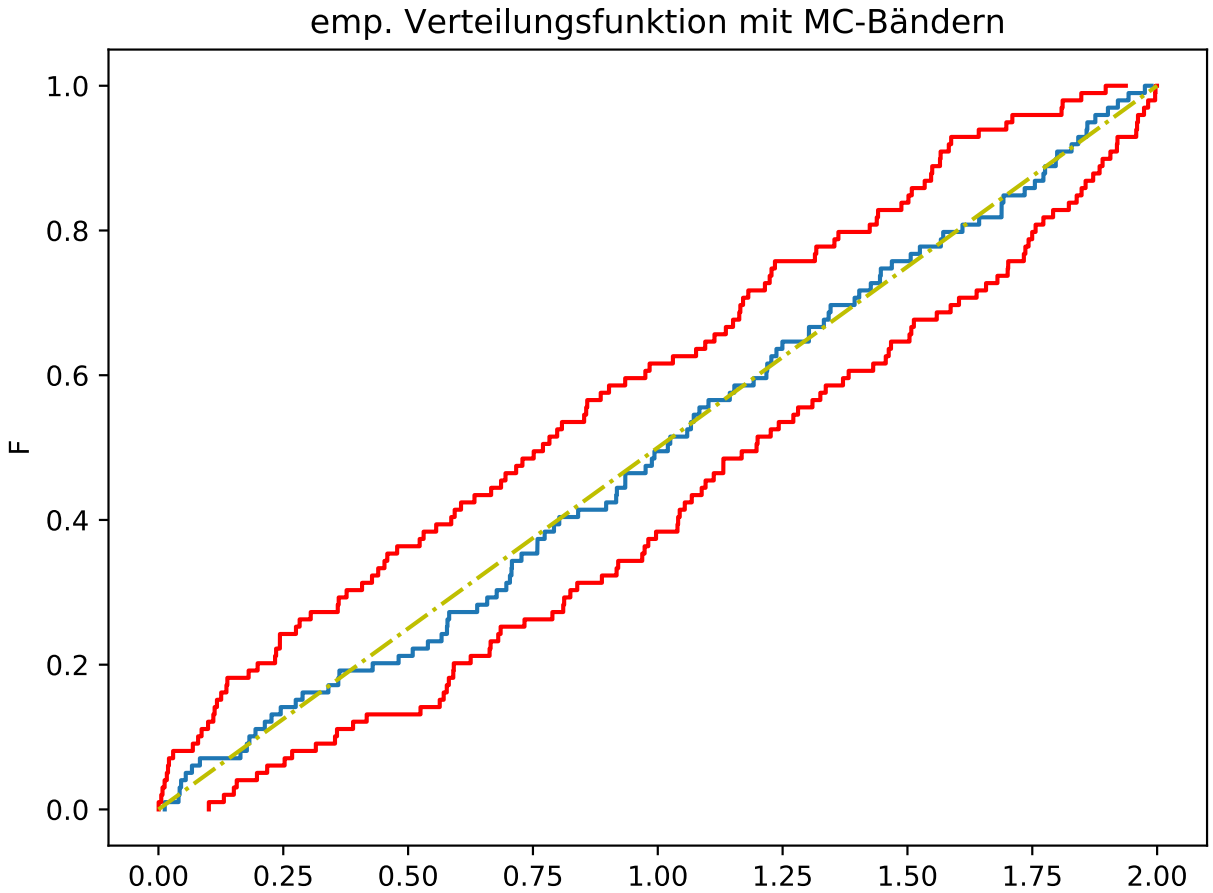


Figure: Monte-Carlo-Bänder für $U(0, 2)$