

Vorlesung

Statistische Methoden der Datenanalyse

Prof. Dr. Dr. Wolfgang Rhode

Generation von Zufallszahlen

Überblick

- Erzeugung von gleichverteilten Zufallszahlen
 - Linear kongruente Generatoren
 - Multiplikativ kongruente Generatoren
 - XOR-Shift
 - Mersenne-Twister
- Spektraltest
- Weitere Tests
- Erzeugung beliebig verteilter Zufallszahlen
 - Transformation der Gleichverteilung
 - Neumann'sches Rückweisungsverfahren
 - Erzeugung bestimmter Verteilungen

Erzeugung von gleichverteilten Zufallszahlen

- Warum keine echten Generatoren?
(Atmosphärenrauschen, CCD-Sensorrauschen, ...)
- Reproduzierbarkeit
- Fehlersuche
- Geschwindigkeit



$$x_{j+1} = f(x_1, \dots, x_j)$$

Alle generierten Zufallszahlen sind allerdings vollständig deterministisch!

Linear kongruente Generatoren (LCG)

$$x_0 \in \mathbb{N}_+ \quad (\text{genannt: Seed})$$

$$x_{j+1} = ((a \cdot x_j + c) \mod m) \Rightarrow u_j = x_j/m$$

- Beispiel: $c = 3, a = 5, m = 16, x_0 = 0$

Linear kongruente Generatoren (LCG)

$$x_0 \in \mathbb{N}_+ \quad (\text{genannt: Seed})$$

$$x_{j+1} = ((a \cdot x_j + c) \mod m) \Rightarrow u_j = x_j/m$$

- Beispiel: $c = 3, a = 5, m = 16, x_0 = 0$

 0, 3, 2, 13, 4, 7, 6, 1, 8, 11, 10, 5, 12, 15, 14, 9, 0...

- Länge der sich wiederholenden Zahlenfolge wird Periodenlänge genannt
 - LCG: Periodenlänge abhängig von a, c und m ; mit m als Obergrenze für die Periodenlänge

Linear kongruente Generatoren (LCG)

$$x_0 \in \mathbb{N}_+ \quad (\text{genannt: Seed})$$

$$x_{j+1} = ((a \cdot x_j + c) \mod m) \Rightarrow u_j = x_j/m$$

- Wie müssen a, c und m gewählt werden, um die maximale Periodenlänge zu erreichen?
 1. $c \neq 0$
 2. c und m teilerfremd
 3. Jeder Primfaktor von m teilt $(a-1)$
 4. Wenn m durch 4 teilbar ist, dann auch $(a-1)$

XOR-Shift Generator

$$x_0 \in \mathbb{N}_+$$

$$t_1 = ((x_j \ll a) \oplus x_j)$$

$$t_2 = ((t_1 \gg b) \oplus t_1)$$

$$x_{j+1} = ((t_2 \ll c) \oplus t_2)$$

x_j	=	11011100
x_j >> 3	=	11100110
(x_j >> 3) XOR x_j	=	00111010
XOR(1,1)	=	0
XOR(0,0)	=	0
XOR(1,0)	=	1 = XOR(0,1)

- Periodenlänge hängt von der Anzahl k Bits ab die zur Darstellung der Zahlen genutzt wird
 - Periodenlänge: $2^k - 1$
- Welche Anforderungen werden an a, b und c gestellt?
 - Wahl nicht trivial
 - Wenn (a,b,c) maximale Periodenlänge ergibt, dann auch alle Permutationen der Zahlen

Mersenne-Twister (MT19937)

- Aktuell der wohl meist eingesetzte Zufallszahlengenerator
- Benötigt 624 Variablen, um seinen Zustand zu speichern
- Auch müssen 624 Startwerte festgelegt werden
- Verwendet u.A. XOR-Shift, um bitweise zufällige Zahlen zu erzeugen.
- Erzeugt 624 Zufallszahlen gleichzeitig

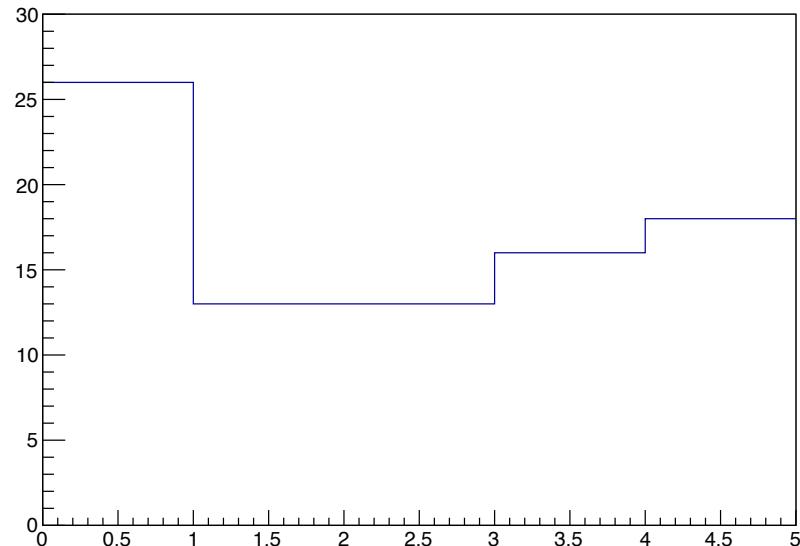
- Hat eine Periodendauer von

$$2^{19937} - 1$$

Spektraltest

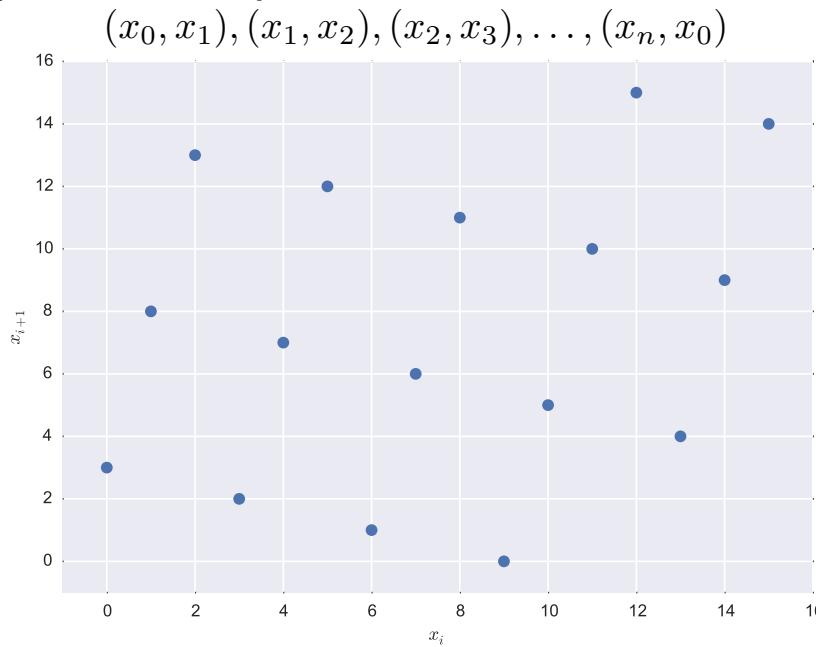
- 1-dim: Wie häufig sind die erzeugten Zahlen (z.B. 0-5)?

- Problematisch:
 - Reihenfolge der Zahlen wird vernachlässigt
 - z.B. 0, 1, 2, 3, 5, 0, 1, 2, 3, 4, 0, 1, 2, 4, 5...



Spektraltest

- 2-dim: Wie häufig sind Wertepaare?



- Beispiel: LCG ($a=5$, $c=3$, $m=16$, $x_0=1$)



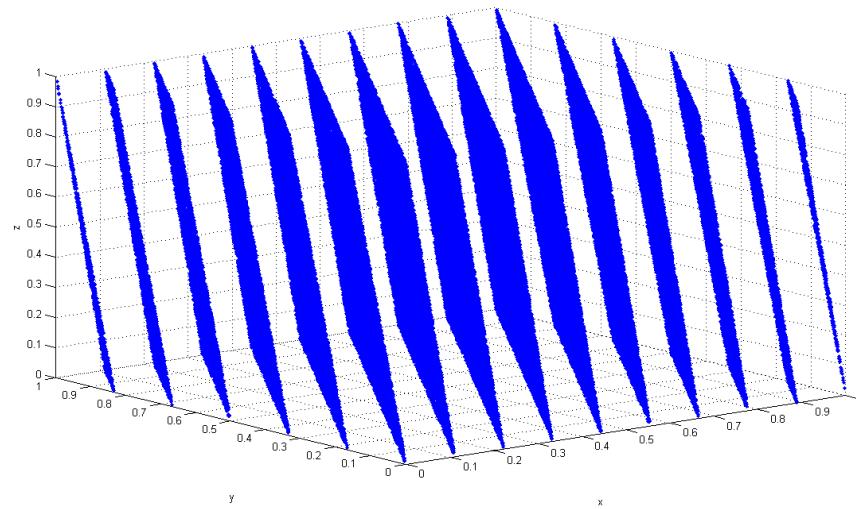
0, 3, 2, 13, 4, 7, 6, 1, 8, 11, 10, 5, 12, 15, 14, 9, 0...

Spektraltest

1. Im Wertebereich $1 \leq x \leq n$ existieren n^2 mögliche Wertepaare.
2. Nur **n Wertepaare** sind realisiert.
3. Normiert man die ganzen Zahlen, so ergibt sich ein Gitterabstand von $1/m$ und die Kantenlänge
4. Durch die besetzten Punkte lassen sich endlich viele Familien von Geraden legen.
5. Betrachte den Abstand von benachbarten Linien einer Familie (die Steigungen dieser Geraden sind gleich)
6. Ist das Gitter **gleichbesetzt** ist der Abstand der Linienpaare der minimale realisierte Abstand $d_2 = m^{-1/2}$.
7. Ist das Gitter **ungleichmäßig** besetzt, dann ist der Abstand $d_2 \gg m^{-1/2}$

Spektraltest

- 3-dim: Wie häufig sind Wertepaare?
- Beispiel: MLCG ($a=65539, m=2^{31}, x_0=1$)



Spektraltest

- 4-dim? n -dim? Schlecht darstellbar!
- $(n-1)$ -dimensionale Hyperebenen
- Beim n -dimensionalen Fall ergibt sich:
- Ist das Gitter **gleichbesetzt** ist der Abstand der Linienpaare der minimale realisierte Abstand $d_n \approx m^{-1/n}$.
- Ist das Gitter **ungleichmäßig** besetzt, dann ist der Abstand $d_n \gg m^{-1/n}$

Weitere Tests

- Birthday Spacing Test:
 - m Geburtstage in einem Jahr mit n Tagen
 - Verteilung der Abstände aller Geburtstage zueinander sollte Poissonverteilt sein
 - Beim tatsächlichen Test: $n = 2^{24}$, $m = 2^{10}$
- Runs Test:
 - Zähle Anzahl n von aufeinander folgenden 0 oder 1 bei den generierten Zahlen
 - Anzahl n sollte Binomialverteilt sein mit $B(n, 0.5)$
- Testbibliotheken:
 - Diehard-Testsuite
 - TestU01 (aktuell)

Hinweise zum praktischen Einsatz

- **Portabilität:** PRNG sollten auf allen Systemen die gleichen Zahlenfolge erzeugen (Meist nicht der Fall!)
- **Seed:** Falsch gewählte Startparameter können die Periodendauer stark verkürzen.
- **"Anlaufzeit":** Bei manchen Generatoren müssen einige der ersten erzeugten Zufallszahlen verworfen werden, wenn die Startparameter schlecht gewählt sind (MT19937)

- **Kombinieren:** Sollte die Periodendauer eines verwendeten PRNG zu kurz sein, so können mehrere Generatoren kombiniert werden.

Erzeugung beliebig verteilter Zufallszahlen

- Transformation der Gleichverteilung

Transformation der Gleichverteilung

- Gesucht: y Zufallsvariable mit der Wahrscheinlichkeitsdichte

$$g(y) \quad y \in [y_{\min}, y_{\max}]$$

- Gegeben: u gleichverteilte Zufallsvariable mit der Wahrscheinlichkeitsdichte

$$f(u) = U(0, 1) = \begin{cases} 1, & 0 \leq u < 1 \\ 0, & \text{sonst} \end{cases}$$

- Zusammenhang: $g(y) = \left| \frac{du}{dy} \right| \cdot f(u)$

Transformation der Gleichverteilung

$$f(u) = U(0, 1) \Rightarrow g(y)dy = U(0, 1)du$$

$$g(y) = \frac{dG(y)}{dy} \Rightarrow dG(y) = g(y)dy = U(0, 1)du$$

Integration liefert:

$$u = \int_{u_{\min}=0}^u U(0, 1)du = G(y) = \int_{y_{\min}}^y g(y')dy'$$
$$y = G^{-1}(u)$$

- Anwendung von G^{-1} auf gleichverteilte Zufallsvariable liefert Zufallsvariable mit gewünschter Verteilung!

Transformation der Gleichverteilung

- Vorteile:
 - Sehr effizient
 - Kein Verwerfen nötig
 - Keine Verschwendungen von Rechenzeit
- Nachteile:
 - Nur anwendbar für integrierbare Zufallsvariablen
 - Umkehrfunktion muss existieren

Erzeugung beliebig verteilter Zufallszahlen

- Transformation der Gleichverteilung
 - Effizient
 - Bedingung:
 - Verteilungsfunktion muss definiert sein
→ Wahrscheinlichkeitsdichte muss integrierbar sein
 - Verteilungsfunktion muss invertierbar sein

Transformation der Gleichverteilung

- Beispiel: Generation von Zufallszahlen im Bereich von 0 bis π , die der Funktion $g(x)=\sin(x)$ folgen
 - Funktion in eine Wahrscheinlichkeitsdichte verwandeln
→ gewünschten Bereich normieren

Normierung (Fläche unter kompletter Kurve): $A = \int_0^\pi \sin(x)dx = 2$

- Integrieren und Invertieren

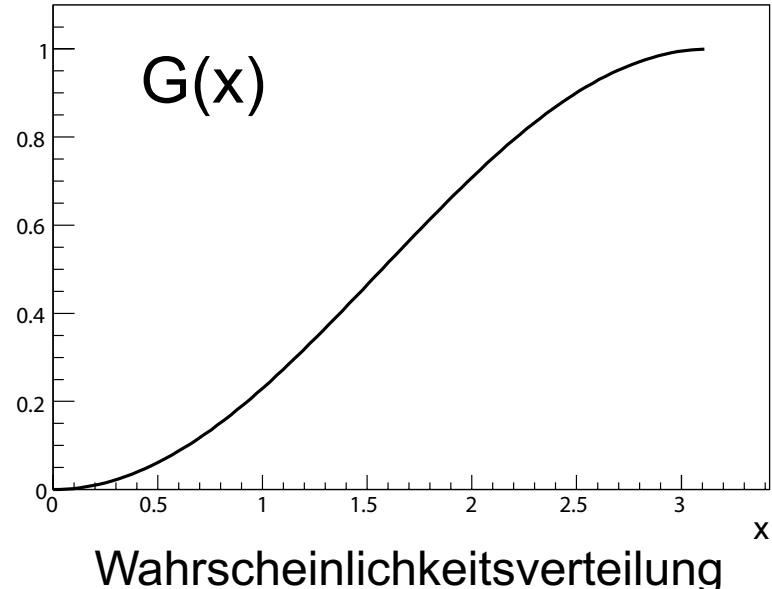
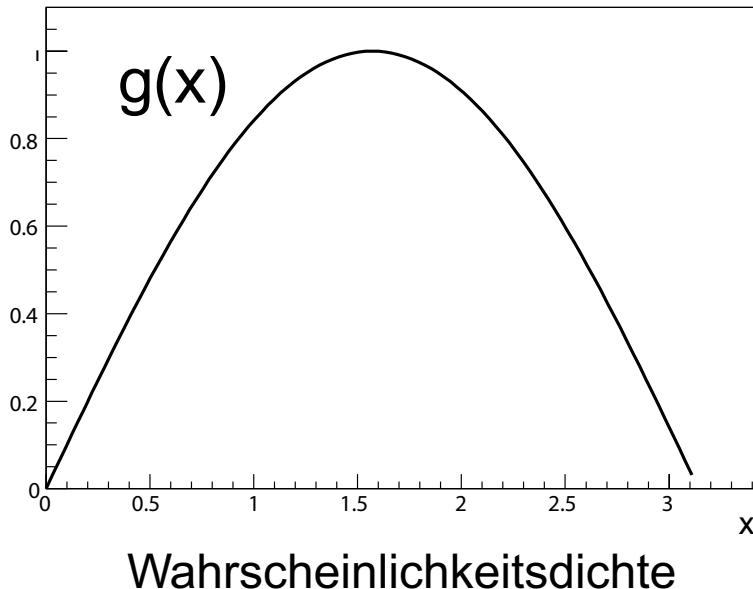
Fläche bis zur Zufallsvariablen x : $A(x) = \int_0^x \sin(t)dt = 1 - \cos(x)$

Normierte relative Fläche: $r(x) = \frac{A(x)}{A} = \frac{1 - \cos(x)}{2}$

Invertierung: $x(r) = \arccos(1 - 2r)$

Transformation der Gleichverteilung

- Beispiel: Generation von Zufallszahlen im Bereich von 0 bis π , die der Funktion $g(x)=\sin(x)$ folgen



Erzeugung beliebig verteilter Zufallszahlen

- Transformation der Gleichverteilung
 - Effizient
 - Bedingung:
 - Verteilungsfunktion muss definiert sein
→ Wahrscheinlichkeitsdichte muss integrierbar sein
 - Verteilungsfunktion muss invertierbar sein
- Neumann'sches Rückweisungsverfahren

Neumann'sches Rückweisungsverfahren

- Gesucht: y Zufallsvariable mit der Wahrscheinlichkeitsdichte

$$g(y) \quad y \in [y_{\min}, y_{\max}]$$

Wahrscheinlichkeitsdichte nicht integrierbar oder Verteilungsfunktion nicht invertierbar

- Gegeben: (u_1, u_2) gleichverteilte Zufallszahlen der Wahrscheinlichkeitsdichten

$$f(u_1) = U(y_{\min}, y_{\max}) = \begin{cases} \frac{1}{y_{\min} - y_{\max}}, & y_{\min} \leq x < y_{\max} \\ 0, & \text{sonst} \end{cases}$$

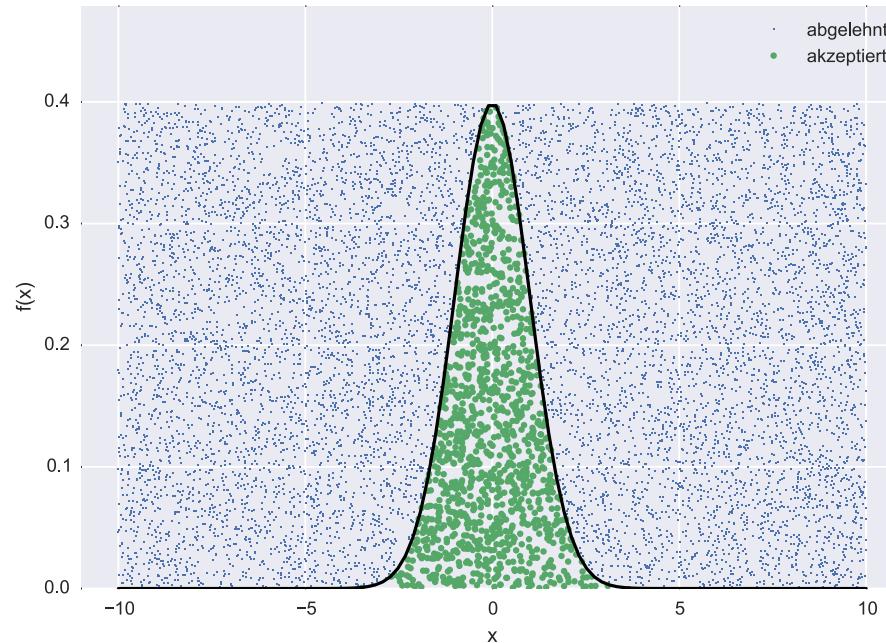
$$f(u_2) = U(0, g_{\max}) = \begin{cases} \frac{1}{g_{\max}}, & g_{\max} = \max(g(y)) \\ 0, & \text{sonst} \end{cases}$$

Neumann'sches Rückweisungsverfahren

- Vorgehen:
 - Wenn $g(u_1) \leq u_2$ wird u_1 verworfen
 - Wenn $g(u_1) > u_2$ wird u_1 als Zufallszahl akzeptiert

Neumann'sches Rückweisungsverfahren

- Vorgehen: Wenn $g(u_1) \leq u_2$ wird u_1 verworfen
Wenn $g(u_1) > u_2$ wird u_1 als Zufallszahl akzeptiert
- Beispiel: Normalverteilung zwischen -10 und 10



Generation von Zufallszahlen

Neumann'sches Rückweisungsverfahren

- Für jede potentielle Zufallszahl müssen zwei gleichverteilte Zufallszahlen erzeugt werden
- Verwerfen vieler Zufallszahlpaare
→ Ineffizient

- Effizienz:

$$E = \frac{\int_a^b g(y)dy}{(b-a)d}$$

Ist $g(y)$ normiert ($\int_a^b g(y)dy = 1$), gilt

$$E = \frac{1}{(b-a)d}$$

Erzeugung beliebig verteilter Zufallszahlen

- Transformation der Gleichverteilung
 - Effizient
 - Bedingung:
 - Verteilungsfunktion muss definiert sein
→ Wahrscheinlichkeitsdichte muss integrierbar sein
 - Verteilungsfunktion muss invertierbar sein
- Neumann'sches Rückweisungsverfahren
 - Ineffizient
 - Zwei gleichverteilte Zufallszahlen für jede potentielle Zufallszahl
 - Verwerfen vieler Paare
 - Kann für jede Funktion genutzt werden

Erzeugung normalverteilter Zufallszahlen – Box-Müller-Methode

- Problem bei Transformationsverfahren: Normalverteilung ist nur numerisch integrierbar
- Lösung: Integration der Normalverteilung in 2D

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^x \int_{-\infty}^y \exp\left(-\frac{1}{2}(x'^2 + y'^2)\right) dx' dy'$$

- Transformation in Polarkoordinaten: $x = r \cos \varphi, \quad y = r \sin \varphi$

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} d\varphi \int_0^r dr' r' \exp\left(-\frac{1}{2}r'^2\right) = 1 - \exp\left(-\frac{1}{2}r^2\right)$$

- → Inversionsmethode für φ und r

Erzeugung normalverteilter Zufallszahlen – Box-Müller-Methode

- 1. Inversionsmethode für φ :

$$u_1 = F(\varphi) = \frac{1}{2\pi} \int_0^\varphi d\varphi' \int_0^\infty dr' r' \exp\left(-\frac{1}{2}r'^2\right) = \frac{\varphi}{2\pi} \Leftrightarrow \varphi = 2\pi u_1$$

- 2. Inversionsmethode für r :

$$u_2 = F(r) = \frac{1}{2\pi} \int_0^{2\pi} d\varphi' \int_0^r dr' r' \exp\left(-\frac{1}{2}r'^2\right) = 1 - \exp\left(-\frac{1}{2}r^2\right) \Leftrightarrow r = \sqrt{-2 \ln(u_2)}$$

- Nach Rücktransformation erhält man zwei unabhängige Zufallszahlen x, y :

$$x = r \cos \varphi = \sqrt{-2 \ln(u_2)} \cos(2\pi u_1)$$

$$y = r \sin \varphi = \sqrt{-2 \ln(u_2)} \sin(2\pi u_1)$$

Erzeugung normalverteilter Zufallszahlen – Polarmethode

- Box-Müller-Methode → Polarmethode: Ersetze Auswertung trigonometrischer Funktionen durch Rückweisungsverfahren
- Polarmethode:
 - Erzeuge gleichverteilte u_1, u_2
 - Umformung $v_1 = 2u_1 - 1, v_2 = 2u_2 - 1$
 - Berechne $s = v_1^2 + v_2^2$
 - Verwerfe, wenn $s \geq 1$
 - Berechne $x_1 = v_1 \sqrt{-\frac{2}{s} \ln s}, x_2 = v_2 \sqrt{-\frac{2}{s} \ln s}$
- x_1, x_2 sind nun unabhängige, normalverteilte Zufallszahlen

Erzeugung normalverteilter Zufallszahlen

- Begründung:

- Betrachte Polarkoordinaten des Punktes (x_1, x_2)

$$x_1 = \cos \theta \sqrt{-2 \ln s}, \quad x_2 = \sin \theta \sqrt{-2 \ln s}$$

- Verteilungsfunktion für $\sqrt{-2 \ln s} \leq r = \sqrt{s}$:

$$F(r) = P(\sqrt{-2 \ln s} \leq r) = P(-2 \ln s \leq r^2) = P(s \geq e^{-\frac{r^2}{2}})$$

- $s=r^2$ gleichverteilt zwischen 0 und 1 $\Rightarrow F(r) = 1 - e^{-\frac{r^2}{2}}$

- Dazugehörige Wahrscheinlichkeitsdichte:

$$f(r) = \frac{dF(r)}{dr} = r e^{-\frac{r^2}{2}}$$

Erzeugung normalverteilter Zufallszahlen

- Begründung:

- Gemeinsame Verteilungsfunktion:

$$\begin{aligned} F(x_1, x_2) &= P(x_1 \leq k_1, x_2 \leq k_2) \\ &= P(r \cos \theta \leq k_1, r \sin \theta \leq k_2) \\ &= \frac{1}{2\pi} \int_{x_1 < k_1} \int_{x_2 < k_2} r e^{-\frac{r^2}{2}} dr d\varphi \\ &= \frac{1}{2\pi} \int_{x_1 < k_1} \int_{x_2 < k_2} e^{-\frac{x_1^2 + x_2^2}{2}} dx dy \\ &= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{k_1} e^{-\frac{x_1^2}{2}} dx_1 \right) \cdot \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{k_2} e^{-\frac{x_2^2}{2}} dx_2 \right) \end{aligned}$$

→ Produkt von 2 standardisierten Normalverteilungen

Erzeugung Poisson-verteilter Zufallszahlen

- Erinnerung:

$$P(r) = \frac{\mu^r e^{-\mu}}{r!}$$

Erzeugung Poisson-verteilter Zufallszahlen

- 1. Möglichkeit:
 - Erzeuge exponentialverteilte Zufallszahlen u_i
 - Summiere u_i bis Summe größer als Mittelwert μ der Poisson-Verteilung
 - Zufallszahl x um eins kleiner als Anzahl der Summenglieder
- Numerischer Trick: Multiplikation mit Logarithmen
 - Logarithmus exponential-verteilter Zufallszahlen = gleichverteilte Zufallszahlen
 - Vergleich mit $e^{-\mu}$

Erzeugung Poisson-verteilter Zufallszahlen

- Numerischer Trick:

$$\frac{1}{\tau} e^{-t/\tau} \text{ mit } t = -\tau \ln x$$

$$\Rightarrow \sum t_i = -\tau \sum \ln x_i$$

$$\Rightarrow \frac{\sum t_i}{\tau} = - \sum \ln x_i$$

$$\text{Exponentiere: } e^{\frac{\sum t_i}{\tau}} = \prod x_i$$

Erzeugung Poisson-verteilter Zufallszahlen

- 2. Möglichkeit:
 - Für große μ : nähere mit Gauß-Verteilung
 - Groß bedeutet $\mu > 10$
- Für eine normalverteilte Zufallszahl Z :

$$n = \max(0, \text{int}(\mu + Z\sqrt{\mu} + 0.5))$$

Erzeugung χ^2 -verteilter Zufallszahlen

- n gerade:
 - Bilde Produkt von $n/2$ gleichverteilten Zahlen

$$x = -2 \ln \left(\prod_{i=1}^{n/2} u_i \right)$$

→ x sind χ^2 -verteilte Zufallszahlen

Erzeugung χ^2 -verteilter Zufallszahlen

- n ungerade:
 - Addiere zum Produkt das Quadrat einer normalverteilten Zufallszahl

$$x = -2 \ln \left(\prod_{i=1}^{(n-1)/2} u_i \right) + Z^2$$

→ x sind χ^2 -verteilte Zufallszahlen

Erzeugung χ^2 -verteilter Zufallszahlen

- n groß ($n > 30$):

- Nähere mit Gauß-Verteilung
- Zufallsvariable y ist angenähert standardisiert normalverteilt

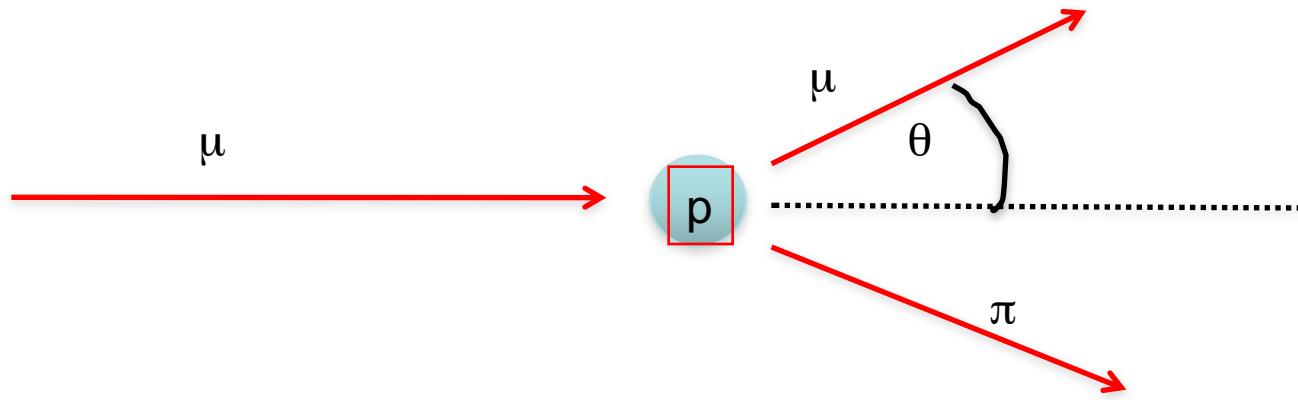
$$y = \sqrt{2\chi^2} - 2\sqrt{2n-1}$$

- Erzeuge Zufallszahl Z der standardisierten Normalverteilung
- Berechne $x = \frac{1}{2}(Z + \sqrt{2n-1})^2$
- Verwerfe, wenn

$$Z < -\sqrt{2n-1}$$

Generation von Wechselwirkungen in der Teilchenphysik

- Beispiel: inelastische Myon-Nukleon-Wechselwirkung, π -Produktion, Laborsystem



- Gegeben: Doppelt differenzieller Wirkungsquerschnitt
 - hängt ab von: (Energie E , Energieübertrag ν , Streuwinkel θ)

$$\frac{d^2\sigma(E,\nu,\theta)}{d\nu d\theta} = f(E,\nu,\theta)$$

Berechnung (1)

- 1. Berechne den totalen Wirkungsquerschnitt , Einheiten: [1/cm²]

$$\int_{\nu_{\min}}^{\nu_{\max}} \int_{\theta_{\min}}^{\theta_{\max}} \frac{d^2\sigma(E, \nu, \theta)}{d\theta d\nu} \cdot d\nu \cdot d\theta = \sigma_{tot}(E)$$

- 2. Berechne die (totale) **Wechselwirkungs**-Wahrscheinlichkeit P_W dafür, dass in einem Medium mit einer Dichte ρ und einem Atomgewicht A auf einem Weg L eine Wechselwirkung stattfindet. N_A sei die Avogadro-Zahl, $f(E)$ sei die Wahrscheinlichkeit dafür, dass das Projektil, die Energie E hat (wann ist das eine Delta-Funktion ;) ?):

$$P_W = \frac{N_A \cdot \rho \cdot L}{A} \cdot \int_{E=0}^{E=\infty} \sigma(E) \cdot f(E) \cdot dE$$

Berechnung (2)

- Beachte: Wird die Schrittweite in dem Medium L sehr groß gewählt, wird auch $P > 1$ (Unsinn!), ist die Schrittweite L sehr klein, müssen sehr viele Operationen ausgeführt werden, bis eine Wechselwirkung stattfindet (Ressourcenverschwendungen, numerische Probleme)
- Die **Ereignis**-Wahrscheinlichkeit P_E , dafür, dass auf der Strecke L in dem Medium n Wechselwirkungen stattfinden, folgt einer Poisson-Verteilung

$$P_E(n, \lambda = P_W) = \frac{\lambda^n}{n!} \cdot e^{-\lambda}$$

Berechnung (3)

- Variante (a): Der (Teil-)Detektor sei relativ dünn (P_W klein). Dann berechnet man die Wahrscheinlichkeit dafür, dass mindestens eine Wechselwirkung auftritt als $1 - P_E(n=0)$.
 - Nun sind Ort(sintervall) und Energie der Wechselwirkung bekannt. Für feste E kann das Verfahren analog zunächst zur Bestimmung des Energieübertrages genutzt werden, dann bei fester Energie und festem Energieübertrag zur Bestimmung des Streuwinkels.
- Variante (b):
 - Berechne die mittlere freie Weglänge

$$l = \frac{\int x \cdot P_{E(n=0)}(x) \cdot dx}{\int P_{E(n=0)}(x) \cdot dx}$$

Berechnung (4)

- Bestimme die Wechselwirkungswahrscheinlichkeit als 1 – der Überlebenswahrscheinlichkeit:

$$P_{\text{int}} = 1 - P(x) = 1 - e^{-\frac{x}{l}}$$

- Berechne den Wechselwirkungspunkt mit der Transformationsmethode
- Weitere Schritte wie in Variante (a)