2.3 Generierung und Bewertung von Zufallszahlen

Warum brauchen wir den Zufall in der Simulation?
Unsere Wahrnehmung der Realität ist i.d.R. nicht deterministisch
Beispiele:

- Ausfallzeit und Reparaturzeit einer Maschine
- Ankunftszeiten von Kunden
- Einschlagstellen von Blitzen
- ...

Beobachtung dieser Phänomene liefert zufälliges Muster, auch wenn auf einem anderen Abstraktionsniveau Erklärungen möglich sind!
Einsatz von Zufall, wenn

- Komplexität vermieden werden soll
- Details nicht bekannt sind
- Zusammenhänge nicht bekannt sind
- zufällige Prozesse in der Natur auftreten

Mathematisches Modell zur Behandlung von zufälligen Ereignissen: **Wahrscheinlichkeitsrechnung**

Kurze und informelle Einführung der benötigten Konzepte

Zufallsexperiment ist ein Prozess, dessen Ausgang wir nicht mit Gewissheit vorhersagen können

- Die Menge der möglichen einander ausschließenden Ausgänge S heißt die Menge der Elementarereignisse
- E ist eine Ereignismenge, falls gilt
  - $\emptyset \in E$ und $S \in E$
  - $A \in E \Rightarrow S \backslash A \in E$
  - $A_i \in E \Rightarrow \cup_i A_i \in E$ und $\cap_i A_i \in E$
- Wahrscheinlichkeitsmaß P[A] bildet $A \in E$ auf reelle Zahlen ab, so dass
  - $0 \leq P[A] \leq 1$, $P[S]=1$ und $P[\emptyset]=0$
  - falls $A \cap B = \emptyset \Rightarrow P[A \cup B] = P[A]+P[B]$

A|B beschreibt, dass A eintritt unter der Bedingung, dass B eintritt/eingetreten ist

Es gilt dann $P[A|B] = P[A \cap B]/P[B]$

Sei $S=A_1 \cup A_2 \cup \ldots \cup A_K$ und alle A_k seien disjunkt, dann

$P[B] = \Sigma_{k=1,\ldots,K} P[B|A_k] \cdot P[A_k]$

Satz von der totalen Wahrscheinlichkeit

Satz von Bayes

$P[A|B] = P[B|A] \cdot P[A] / P[B]$

Zufallsvariable (ZV) ist eine Variable, deren Wert durch den Ausgang eines Zufallsexperiments bestimmt ist und eine reelle Zahl ist.
Bezeichnung für ZVs: X, Y, Z, ...
Bezeichnung für den Wert einer ZVs: x, y, z, ...

Unterscheidung in
diskrete Zufallsvariablen mit endlichem oder abzählbarem Wertebereich
Beispiele: Münzwurf, Würfeln, Anzahl eingehende Telefonanrufe, ...
kontinuierliche Zufallsvariablen mit überabzählbarem Wertebereich
Beispiele: Zwischenankunftszeit, Bedienzeit, Regenmenge, ...

Charakterisierung von Zufallsvariablen

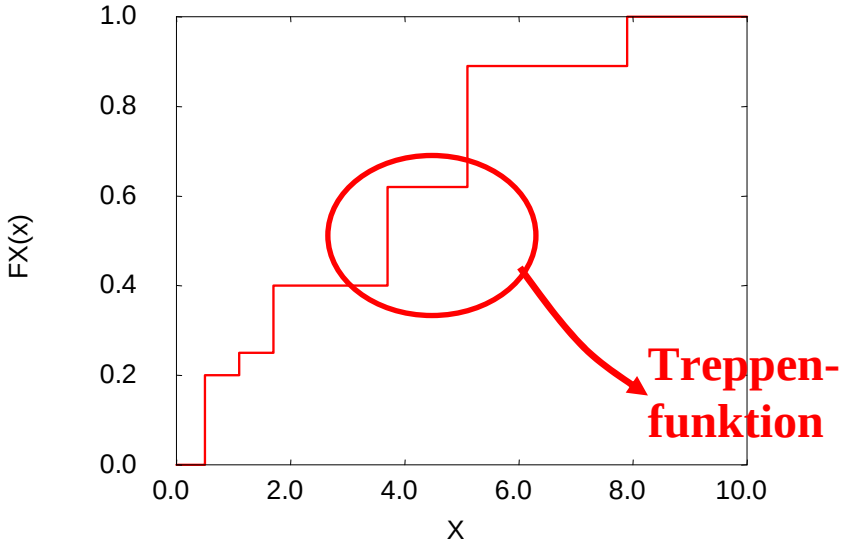
Verteilungsfunktion (Vfkt) $F(x)=P[X \leq x]$ für $-\infty \leq x \leq \infty$

Es gilt:

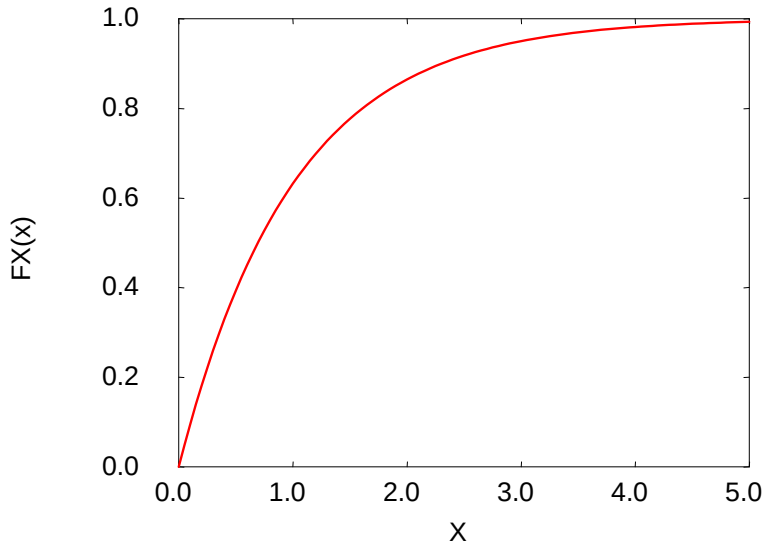
- $0 \leq F(x) \leq 1$
- $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$
- $\lim_{x \to \infty} F(x) = 1$ und $\lim_{x \to -\infty} F(x) = 0$

Graphische Repräsentation

diskrete ZV



kontinuierliche ZV

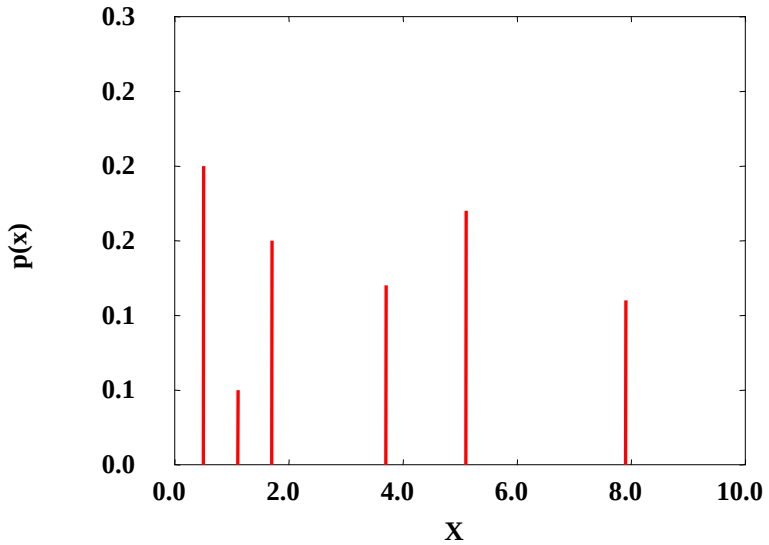


Diskrete ZV X mit Wertebereich W_X

Wahrscheinlichkeit p(x)=P[X=x], es gilt

- $p(x)=0$ für $x \notin W_X$
- $0 \leq p(x) \leq 1$ für $x \in W_X$
- $\sum_{x \in W_X} p(x) = 1.0$
- $\sum_{x \in W_X \wedge x \leq y} p(x) = F(y)$

Graphische Repräsentation



Momente der Verteilung

- $E(X^i)= \sum_{x \in W_X} p(x) \cdot x^i$ i-tes Moment
- $E(X)= E(X^1)$ erstes Moment oder Erwartungswert
- $\sigma^2(X)=E(X^2)-E(X)^2$ Varianz und $\sigma(X)$ Standardabweichung
- $VK(X)=\sigma(X)/E(X)$ Variationskoeffizient
- $C(X,Y)=E((X-E(X)) \cdot (Y-E(Y)))=E(X \cdot Y)-E(X) \cdot E(Y)$ Kovarianz

Typische/Wichtige diskrete Verteilungen

Bernoulli-Verteilung
Parameter $p \in [0,1]$:

$$p(x) = \begin{cases} p & \text{falls } x = 1 \\ 1-p & \text{falls } x = 0 \\ 0 & \text{sonst} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{falls } x < 0 \\ 1-p & \text{falls } 0 \le x < 1 \\ 1 & \text{falls } 1 \le x \end{cases}$$

$E(X)=p$

$\sigma^2(X)=p\cdot(1-p)$

Anwendung:
binäre Entscheidungen

Geometrische-Verteilung
Parameter $p \in (0,1]$: :

$$p(x) = \begin{cases} p \cdot (1-p)^x & \text{falls } x \in \{0,1,2,\ldots\} \\ 0 & \text{sonst} \end{cases}$$

$$F(x) = \begin{cases} 1-(1-p)^{\lfloor x \rfloor + 1} & \text{falls } x \ge 0 \\ 0 & \text{sonst} \end{cases}$$

$E(X)=(1-p)/p$

$\sigma^2(X)=(1-p)/p^2$

Anwendung:
Anzahl erfolgloser Versuche bis zum Erfolg bei Erfolgswahrscheinlichkeit p

Poisson-Prozess (Parameter λ>0)

$$p(x) = \begin{cases} \dfrac{e^{-\lambda} \cdot \lambda^x}{x!} & \text{falls } x \in \{0,1,2,\ldots\} \\ 0 & \text{sonst} \end{cases}$$

$$F(x) = \begin{cases} e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \dfrac{\lambda^i}{i!} & \text{falls } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

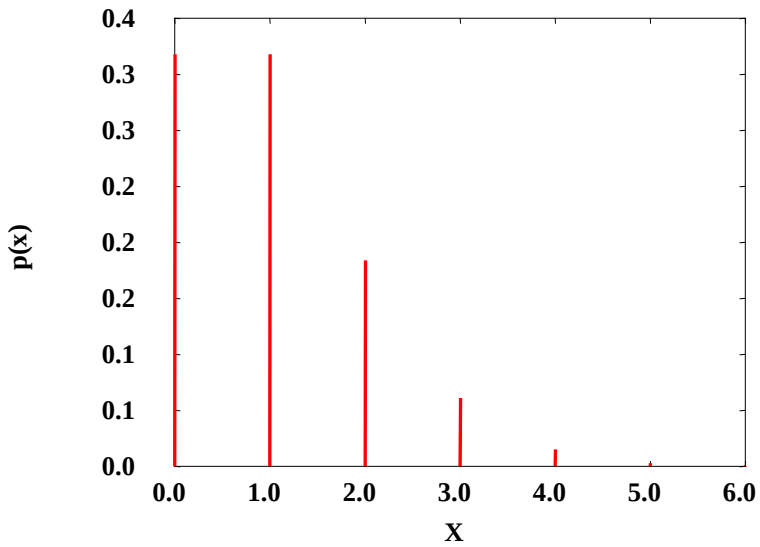
$E(X)=\lambda$

$\sigma^2(X)=\lambda$

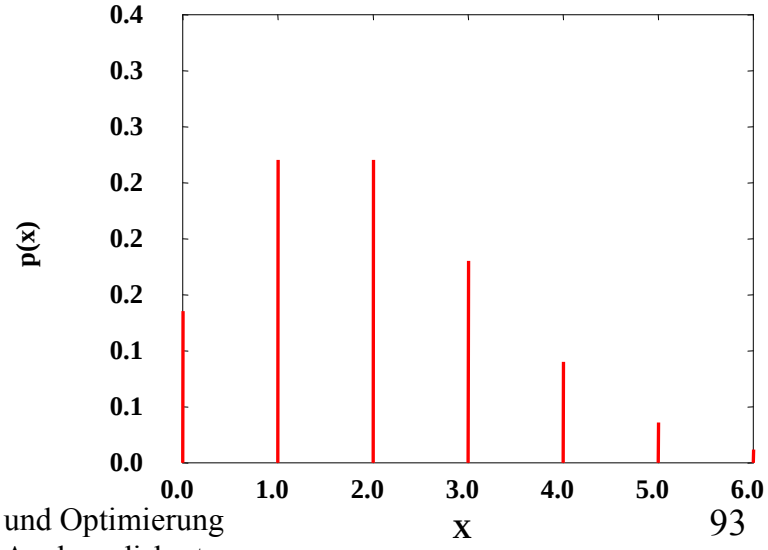
Anwendung:

- Anzahl Ereignisse in einem Zeitintervall bei konstanter Eintrittsrate
- viele praktische Anwendungen

λ=1.0



λ=2.0

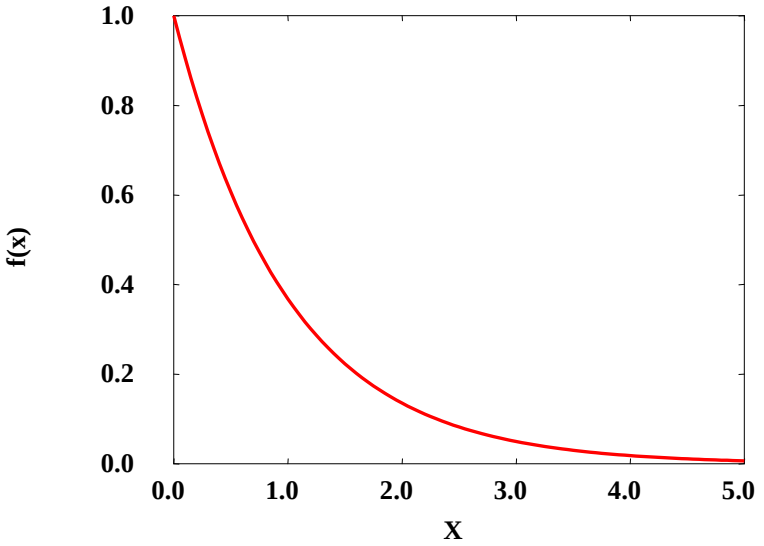


Kontinuierliche ZV X mit Wertebereich W_X

Dichtefunktion (Dfkt) f(x), es gilt

- $f(x)=0$ für $x \notin W_X$
- $0 \leq f(x)$ für $x \in W_X$
- $\int_{x \in W_X} f(x)dx = 1.0$
- $\int_{x \in W_X \wedge x \leq y} f(x)dx = F(y)$
- $\int_y^z f(x)dx = F(z)-F(y)$ falls $z \geq y$

Graphische Repräsentation



Aber $p(x)=P[X \in [x,x]]= \int_x^x f(y)dy=0$ kann gelten

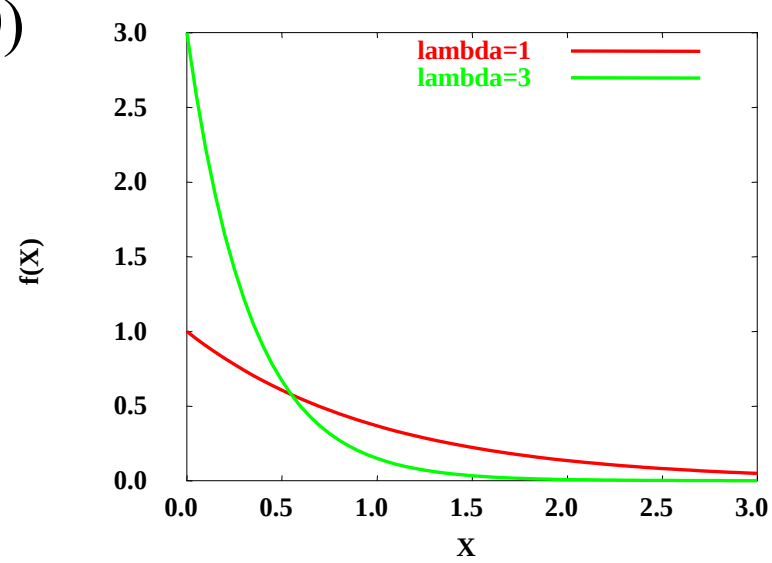
Momente $E(X^i)= \int_{x \in W_X} f(x) \cdot x^i \, dx$

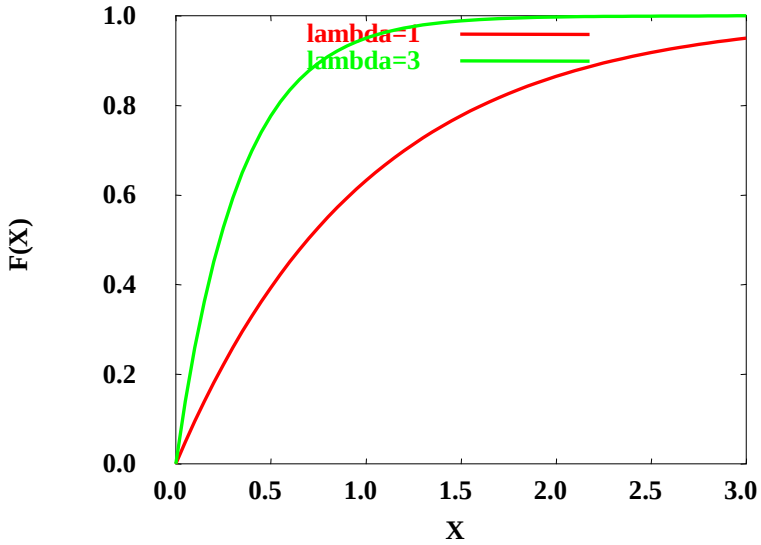
Wichtige kontinuierliche Verteilungen

Exponentialverteilung (Parameter $\lambda>0$)

$$f(x)=\begin{cases}\lambda\cdot e^{-\lambda x} & \text{falls } x\geq 0\\ 0 & \text{sonst}\end{cases}$$

$$F(x)=\begin{cases}1-e^{-\lambda x} & \text{falls } x\geq 0\\ 0 & \text{sonst}\end{cases}$$





- $E(X)=1/\lambda$ $\Rightarrow VK(X)=1$
- $\sigma^2(X)=1/\lambda^2$

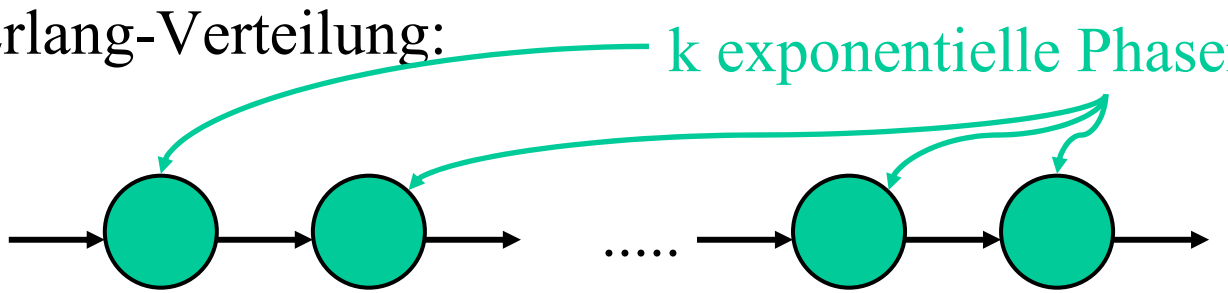
Anwendung:

Addition vieler seltener Ereignisse durch „Gedächtnislosigkeit“ analytisch handhabbar

Erweiterung der Modellierungsmächtigkeit der Exponentialverteilung:

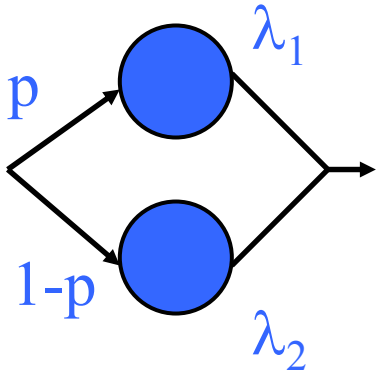
- Kombination von exponentiellen Phasen

Erlang-Verteilung:



$E(X)=k/\lambda$, $\sigma^2(X)=k/\lambda^2$ und $VK(X)=1/k^{1/2}$
(weniger variabel als Exp.-Verteilung)

Hyperexponential-Verteilung:

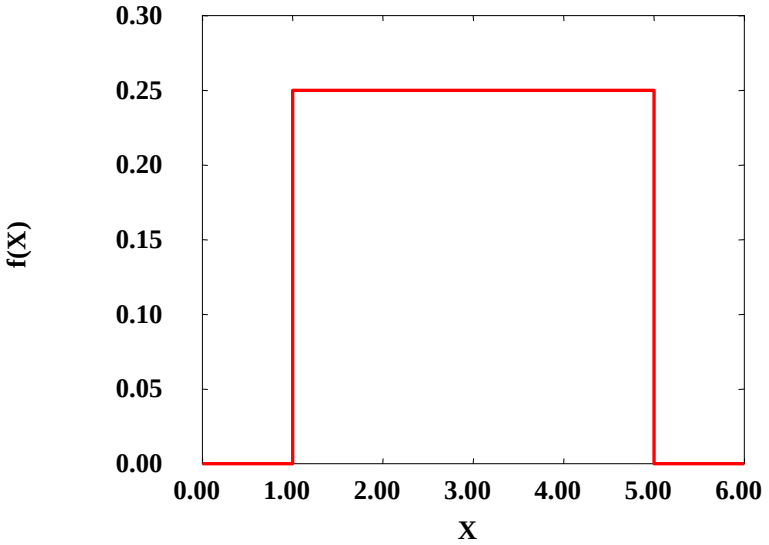


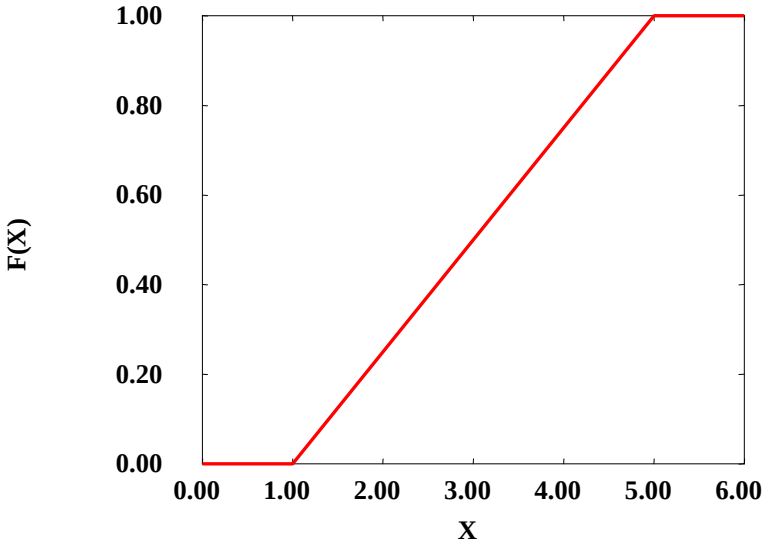
- $E(X)=p/\lambda_1+(1-p)/\lambda_2$,
- $\sigma^2(X)=p(2-p)/\lambda_1^2 + (1-p^2)\ )/\lambda_2^2 - 2p(1-p)/(\lambda_1\lambda_2)$
  (mindestens so variabel wie Exp.-Verteilung, beliebe $VK \geq 1$ realisierbar)

Gleichverteilung (Parameter a,b)

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{falls } a \le x \le b \\ 0 & \text{sonst} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{falls } x < a \\ \dfrac{x-a}{b-a} & \text{falls a} \le \text{x} < \text{b} \\ 1 & \text{falls b} \le \text{x} \end{cases}$$





- E(X)=(a+b)/2 $\Rightarrow$VK(X)≈1.73
- $\sigma^2(X)=(b-a)^2/12$

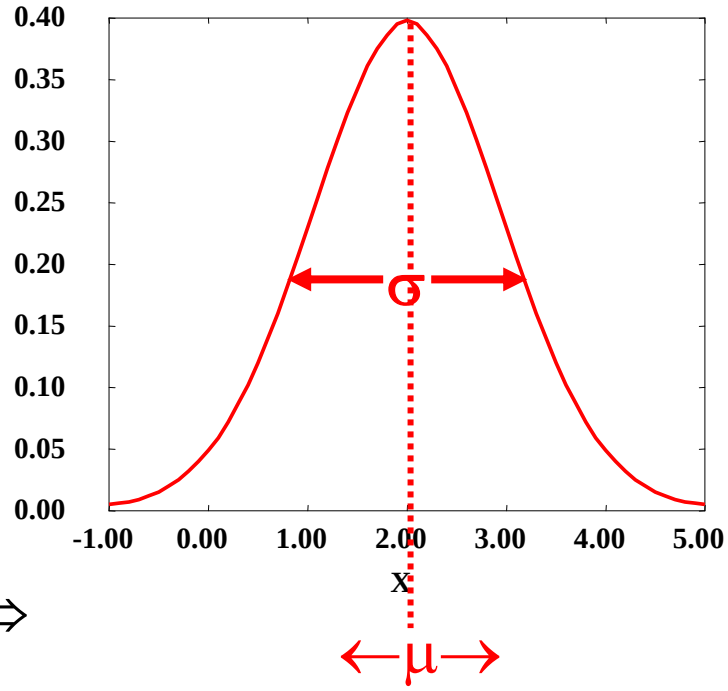
Anwendung:

[0,1]-Gleichverteilung ist die Basis zur Generierung allg. Vert.

Normalverteilung (Parameter μ,σ)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Keine geschlossene Form für F(X)



- E(X)=μ
- $\sigma^2(X)=\sigma^2$

Normalverteilt mit Parametern μ, σ ⇒
Schreibweise $X \sim N(\mu, \sigma)$

Anwendung:

Zentraler Grenzwertsatz: Summe unabhängiger ZVs konvergiert gegen eine Normalverteilung (unter relativ allg. Bedingungen)

⇒ Grundlage vieler statistischer Testverfahren
u.a. der Bestimmung von Konfidenzintervallen (siehe 2.1.5)

Stichproben und Schätzer

Sei X eine ZV mit unbekannter Verteilung
$x_1,..., x_n$ sei eine Menge von Beobachtungen von X (eine Stichprobe)
alle Beobachtungen seien unabhängig und aus der selben Verteilung

Oft sollen Aussagen über Parameter Θ der Verteilung von X mit Hilfe von $x_1,..., x_n$ gewonnen werden

Ein Schätzer $\tilde{\Theta}$ für einen Parameter Θ der Verteilung einer ZV X auf Basis einer Stichprobe $x_1,..., x_n$ ist eine Funktion

$g(x_1,..., x_n) \rightarrow \tilde{\Theta}$ (g ist die Schätzfunktion).

$\tilde{\Theta}$ heißt

- erwartungstreu, wenn $E(\tilde{\Theta}) = \Theta$
- asymptotisch erwartungstreu, wenn $\lim_{n\rightarrow\infty} E(\tilde{\Theta}) = \Theta$
- konsistent, wenn $\lim_{n\rightarrow\infty} P[|\tilde{\Theta} - \Theta| > \varepsilon] = 0$ für jedes $\varepsilon > 0$

Erzeugung von Zufallszahlen

Ziel: Realisierungen einer ZV X sollen generiert werden
„Ziehen von Zufallszahlen (ZZ)“

- Gesucht Methode $zz(X) \rightarrow x$, so dass
  für so erzeugte x, $P[x<y]=P[X<y]$ für alle y
  ($\Rightarrow E(x^i)=E(X^i)$ für alle i)
- Eine Sequenz $x_1, x_2, \ldots$ liefere unabhängig ZZs

Schritte der Erzeugung von ZZs:

1. Erzeugung von gleichverteilten ganzzahlige ZZs im Intervall [0,m)
2. Transformation in (approximativ) [0,1)-gleichverteilte ZZs
3. Transformation in ZZs der gewünschten Verteilung

Basis aller ZZ-Generatoren: Erzeugung von [0,1)-gleichverteilten ZZs

Echte Zufallszahlen:

- Münzwurf (Z=1, K=0)
- Würfeln (Zahlen 1,..6)
- Physikalische Messungen (z.B. radioaktiver Zerfall)

echte ZZs sind nicht reproduzierbar!

Pseudozufallszahlen:
Sequenz von ZZs, die für einen Beobachter zufällig ausssieht

- Tafeln mit ZZs (z.B. Tippett 1927, 41600 gleichvert. Zahlen aus Daten der Finanzverwaltung)
- Generierungsalgorithmus der Form $x_i=g(s_i)$ und $s_{i+1}=f(s_i)$

Pseudo-ZZs sind reproduzierbar!

In der Simulation werden fast nur Pseudozufallszahlen eingesetzt (Reproduzierbarkeit von Programmläufen, ...)

Jeder Generierungsalgorithmus erzeugt eine endliche Sequenz von ZZs

Anforderungen

- Generierten ZZs müssen gleichverteilt sein
- Generierte ZZs müssen unabhängig sein
  (d.h. Kenntnis von n ZZs darf keine zusätzlich Information über die n+1te ZZ liefern, ohne dass der Generierungsalgorithmus bekannt ist)
- die Sequenzlänge bis zur Wiederholung muss groß sein
- die Erzeugung muss effizient sein
- der Algorithmus muss portabel sein

Unterschiedliche Klassen von Generatoren existieren, wir betrachten nur die am weitesten verbreitete Klasse!

Ein erster Versuch: Midsquare Methode (von Neumann 1940)

1. Wähle eine vierstellige Zahl
2. Quadriere diese
   ($\Rightarrow$ achtstellige Zahl, u.U. von Links mit 0 auffüllen)
3. Wähle die mittleren vier Stellen als Nachkommastellen einer ZZ und fahre mit der Zahl bei 1. fort

Beispiel:

7182	$\rightarrow$	51581124	ZZ: 0.5811
5811	$\rightarrow$	33767721	ZZ: 0.7677
7677	$\rightarrow$	58936329	ZZ: 0.9363
9363	$\rightarrow$	87665769	ZZ: 0.6657
6657	$\rightarrow$	44315649	ZZ: 0.3156

usw.

Generierte Zufallszahlen sind sehr schlecht!

Lineare Kongruenzgeneratoren (Lehmer 1951)

Eine Funktion $x_i = \left(\sum_{j=1}^{r} a_j \cdot x_{i-j} + c \right) (\mathrm{mod}\, m)$

mit $x_0, x_1, ..., x_{r-1}, a_1, ..., a_r, c \in$ ☠
heißt linearer Kongruenzgenerator (LCG)

Heute meist verwendet $x_i = (a \cdot x_{i-1} + c)(\mathrm{mod}\ m)$

Falls c=0 multiplikativer Generator, sonst gemischter Generator!
x_0 ist die Saat des Generators

Eigenschaften:

- $x_i \in [0,m)$ falls $c>0$ und $x_i \in (0,m)$ falls $c=0$ (notwendigerweise)
- $x_i = x_j \Rightarrow x_{i+k} = x_{j+k}$ für alle $k \geq 0$ (Generator beginnt von vorn)
- $\kappa = \min_{|i-j|} (x_i = x_j)$ heißt die Periode des LCGs

Ein einfaches Beispiel:
a=5, c=0 und m=17 also $x_{i+1} = (5x_i)(\text{mod } 17)$ mit Saat $x_0=5$

i	0	1	2	3	4	5	6	7	8
$5x_i$	25	40	30	65	70	10	50	80	60
x_{i+1}	8	6	13	14	2	10	16	12	9
i	9	10	11	12	13	14	15	16	
$5x_i$	45	55	20	15	75	35	5	25	
x_{i+1}	11	4	3	15	7	1	5	8	

Maximale Periode 1,...,16 wird erreicht!

Beispiele für (einige ältere) Generatoren:

- Unix $m=2^{32}$, $a=1103515245$, $c=12345$
- RANDU $m=2^{31}$, $a=65539$
- Simula/Univac $m=2^{31}$, $a=5^{13}$
- SIMPL-I (IBM) $m=2^{31}-1$, $a=48271$
- SIMSCRIPT II.5 $m=2^{31}-1$, $a=630360016$
- L'Ecuyer $m= 2^{63}-25$, $a=4645906587823291368$

schlechte LCGs!

Anforderungen an einen guten LCG:

1. große Periode (großes m)
2. effiziente Berechnung
3. generierte ZZs bestehen statistische Tests

Große Periode eines LCG $\Rightarrow$ möglichst m oder m-1 Werte!

Ein gemischter LCG hat genau dann volle Periodenlänge, wenn:

1. ggT(m,c) = 1 (c und m sind relativ prim)
2. a mod p_i = 1 für alle Primfaktoren p_i von m
3. a mod 4 = 1 falls 4 Faktor von m ist

Ein multiplikativer LCG hat genau dann volle Periodenlänge, wenn

- Falls $m = 2^b$, dann ist der maximale Wert für $\kappa = m/4 = 2^{b-2}$, wird erreicht für ungerade Saaten und a = 3 + 8k oder a = 5 + 8k mit k = 0,1 , …
- Falls m eine Primzahl ist, dann wird $\kappa = m-1$ erreicht, falls die kleinste Zahl k, für die a^k-1 durch m ganzzahlig dividiert werden kann, gleich m-1 ist (d.h. a ist Primitivwurzel von m)

Wie groß sollte/muss m heute sein?

Periode vieler heutiger Generatoren $\kappa \approx 2^{31}\text{-}2^{32}$

Vergrößerung der Periode z.B. durch Kombination von Generatoren:

- k Generatoren mit Periode und Modul m_j für den j-ten Generator
- sei $x_{i,j}$ i-ter Wert des j-ten Generators
- $$x_i = \left(\sum_{j=1}^{k} (-1)^{j-1} x_{i,j} \right) (\mathrm{mod}\, m_1 - 1)$$

  ist $[0, m_1\text{-}2]$-gleichverteilt
- maximal erreichbare Periode

$$\left(\prod_{i=1}^{k} (m_i - 1) \right) \Big/ 2^{k-1}$$

Effizienz der Generierung

Division ist aufwändig ⇒ Effizienz erfordert wenige/keine Divisionen!

Falls $m=2^e$, kann mod durch „Weglassen“ von Stellen realisiert werden! (shiften auf Bitebene)

Beispiel: $10111011 \bmod 2^6 = 00111011$

in Dezimaldarstellung $187 \bmod 64 = 59$

Aber $m=2^e$ bedingt Zyklen der niederwertigen Bits (schlechte ZZs!)

Beispiel: $x_{i+1} = (5x_i + 1) \bmod 16$

i	0	1	2	3	4	5	6	7
x_i	1	6	15	12	13	2	11	8
b_0	1	0	1	0	1	0	1	0
b_1b_0	01	10	11	00	01	10	11	00
$b_2b_1b_0$	001	110	111	100	101	010	011	000

Alternative: Wähle $m = 2^e - 1$

Effiziente Berechnung ohne Division:

Gesucht $z = (a \cdot x) \pmod{2^e - 1}$

Sei $y = (a \cdot x) \pmod{2^e}$ (effizient berechenbar)

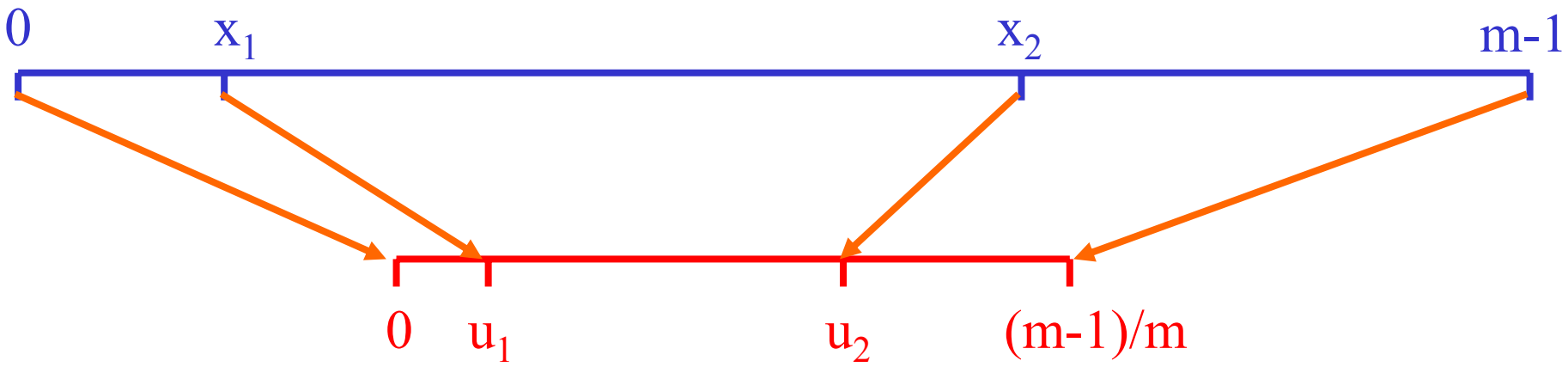
Es gilt:

$$z = \begin{cases} y + k & \text{falls } y + k < 2^e - 1 \\ y + k - (2^e - 1) & \text{sonst} \end{cases}$$

mit $k = \lfloor a \cdot x / 2^e \rfloor$

Ähnlich effizient, aber mit besseren Eigenschaften als 2^e!

Generierung von [0,1)-gleichverteilten reellwertigen ZZs aus [0,m)-gleichverteilten ganzzahligen ZZs



Streng genommen werden nur diskrete Werte i/m aus [0,1) erzeugt

- aber die Werte liegen sehr dicht
  (bei Beachtung der Maschinendarstellung sogar optimal dicht)
- d.h. bei voller Periode sind alle im Intervall darstellbaren Zahlen enthalten

Testverfahren für Zufallszahlen

[0,1)-gleichverteilte ZZs sind die Basis für weitere (oft exakte) Transformationen $\Rightarrow$ gute [0,1)-Gleichverteilung ist notwendig

Anforderungen an generierte ZZs:

- Sie müssen gleichverteilt in [0,1) sein
- Sie müssen unabhängig sein

Testverfahren dienen zur Bewertung dieser Eigenschaften

Man unterscheidet:

- Empirische Testverfahren
  Bewertung der ZZ auf Basis einer Stichprobe
- Theoretische Testverfahren
  Bewertung aller generierten ZZ (oft schwieriger)

Testen von Zufallszahlengeneratoren:

Grundsätzliches Vorgehen beim Testen:

Aufstellen einer Hypothese H_0

- Mit Generator G erzeugte ZZs sind unabhängig, identisch [0,1)-gleichverteilt oder
- Mit Generator G erzeugte ZZs sind nicht unabhängig, identisch [0,1)-gleichverteilt

H_0 heißt **Nullhypothese**, $H_1 = \neg H_0$ heißt **Alternativhypothese**

Testverfahren dienen dazu herauszufinden, ob H_0 gilt

Auf Grund statistischer Schwankungen von Stichproben, können falsche Folgerungen gezogen werden (Tests sind keine Beweise!)

Mögliche Fehler

- (statistischer) Fehler der 1. Art (α Fehler): H_1 wird angenommen, obwohl H_0 gegeben (fälschliches Verwerfen der Nullhypothese)
- (statistischer) Fehler der 2. Art (β Fehler): H_0 wird angenommen, obwohl H_1 gegeben (fälschliches Annehmen der Nullhypothese)

Impliziert ein bestimmter Test mit Wahrscheinlichkeit $\leq \alpha$ Fehler der 1. Art, so heißt er „Test zum (Signifikanz-)Niveau α“

unabhängig (!) von der Wahrscheinlichkeit eines Fehlers der 2. Art!

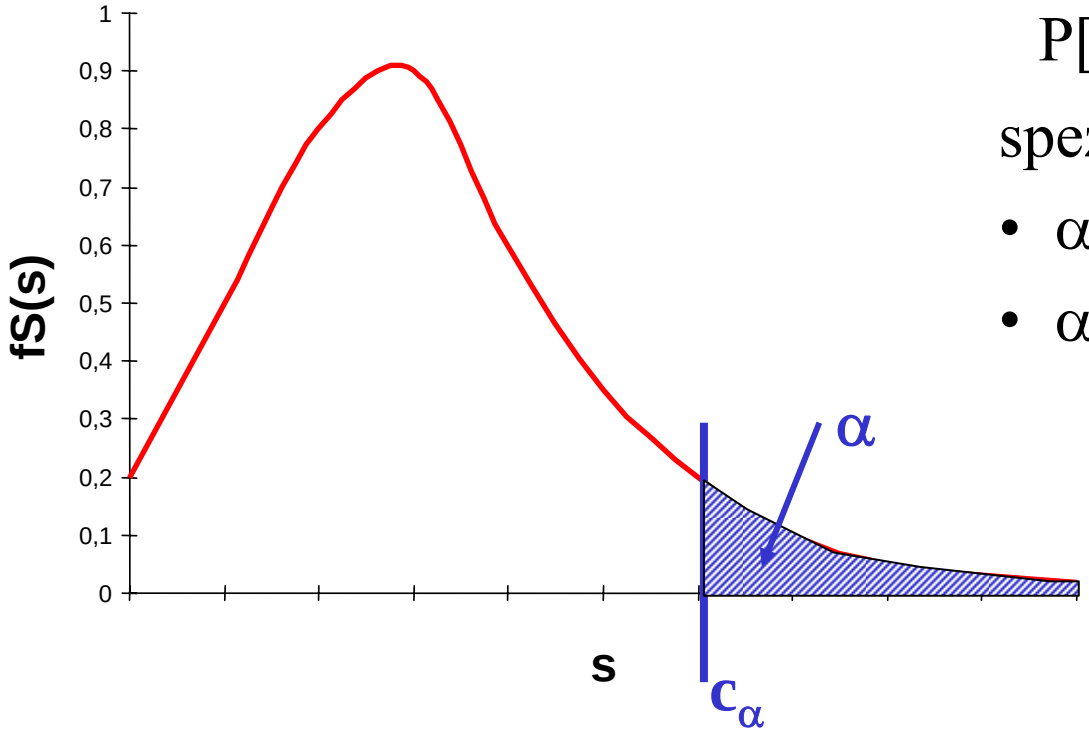
Vorgehen beim Testen auf Basis einer Stichprobe (d.h. einer Realisierung der ZVs $Y_1, ..., Y_n$):

Teststatistik $S(Y_1, ..., Y_n)$ „bewertet“ die Stichprobe, je größer der Wert, desto unwahrscheinlicher H_0 (und implizit je wahrscheinlicher H_1)

Zur Anwendung erforderlich:

- Bestimme die Verteilung von S (unter der Voraussetzung, dass H_0 gilt)
- Ermittlung der kritischen Werte c_α (oder $c_{1-\alpha}$) ab denen H_0 zum Niveau α verworfen wird

Skizze des Prinzips



c_α so bestimmt, dass

$P[S(Y_1,\ldots,Y_n)>c_\alpha|H_0] = \alpha$

spezielle α-Werte:

- α=0.05 signifikant
- α=0.01 hochsignifikant

Verteilung sagt nichts über den Fehler 2. Art aus

Testverfahren können schlechte Generatoren identifizieren, nicht aber die generelle Güte eines Generators nachweisen!

Runs Test

Paare aufeinander folgender ZZs werden in Klassen unterteilt:

- Ein Paar fällt in Klasse +, wenn der erste Wert kleiner als der zweite ist
- Ansonsten fällt das Paar in Klasse –

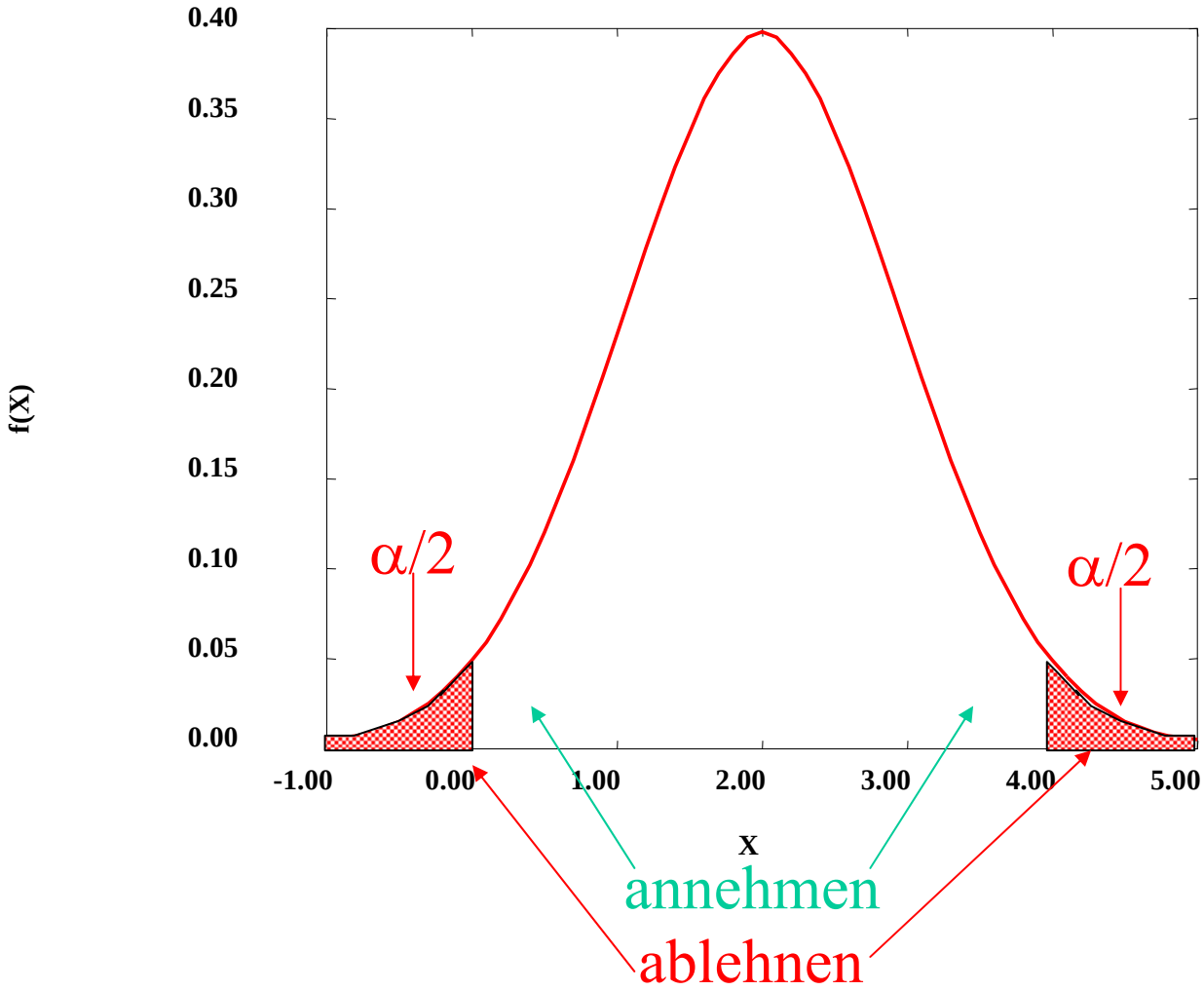
Ein run ist eine Folge identischer Zeichen + oder –

Sei R_n die Anzahl der runs, n die Größe der Stichprobe, dann ist die Anzahl der runs für große n normalverteilt mit

- Erwartungswert $E(R) =(2n - 1)/3$
- Varianz $\sigma^2(R) = (16n – 29)/90$

Auf Basis der kritischen Werte der Normalverteilung kann ein statistischer Test durchgeführt werden und damit die Hypothese „[0,1)-verteilte ZZs“ abgelehnt oder angenommen werden

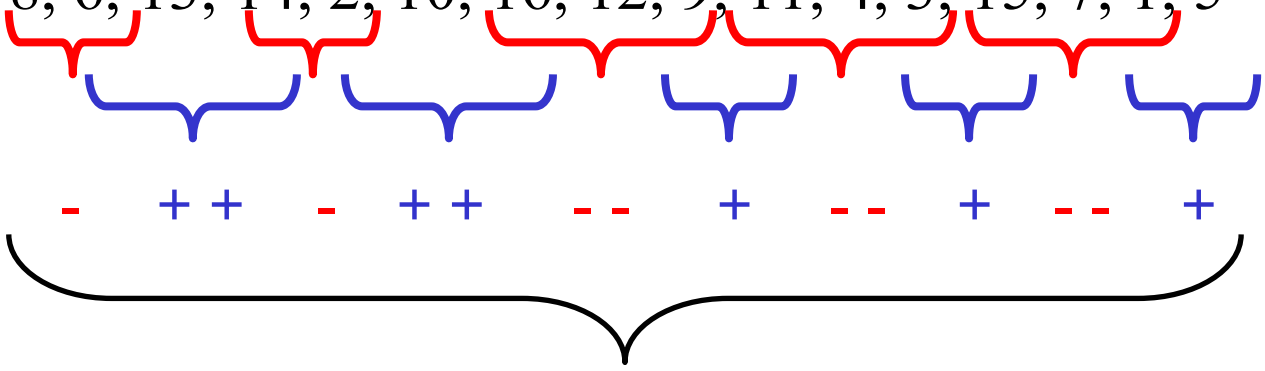
Skizze des Prinzips



Beispiel $x_{i+1} = (5 \cdot x_i) \pmod{17}$

Generierte ZZs für x_0=5:

8, 6, 13, 14, 2, 10, 16, 12, 9, 11, 4, 3, 15, 7, 1, 5



Laut Theorie: $E(R_{16}) = (2 \cdot 16 - 1)/3 = 10.333$
(Gute Übereinstimmung)

Für α=0.05 würde die Hypothese unabhängig [0,1)-verteilt (H_0)
für 8 bis 13 runs akzeptiert!

Verschiedene ähnliche Tests existieren
(zum Teil auch unter dem Namen runs-Test)

Test der Autokorrelation

Seien $X_1, ..., X_n$ ZVs mit Erwartungswert μ

Der Autokorrelationskoeffizient der Ordnung s $\rho(s)$ ist definiert als

$$\rho(s) = \frac{\sum_{i=1}^{n-s}(X_{i+s}-\mu)(X_i-\mu)}{\sum_{i=0}^{n-s}(X_i-\mu)^2} = \frac{C(X_i, X_{i+s})}{\sigma^2(X)} = \frac{C_s}{C_0}$$

Es gilt:

- $-1 \leq \rho(s) \leq 1$
- falls X_i und X_{i+s} unabhängig, so gilt $\rho(s) = 0$

Für [0,1)-verteilte ZV X gilt ferner

- $E(X) = 1/2$ und $\sigma^2(X) = C_0 = 1/12$,

Da $C_s = E(X_iX_{i+s}) - E(X_i)E(X_{i+s})$ gilt in diesem Fall auch

$$\rho(s) = \frac{E(X_i, X_{i+s}) - E(X_i)E(X_{i+s})}{\sigma^2(X)} = \frac{E(X_i, X_{i+s}) - 1/4}{1/12} = 12E(X_i, X_{i+s}) - 3$$

Ein Schätzer für $\rho(s)$ lautet:

$$\tilde{\rho}(s) = \frac{12}{h+1} \sum_{k=0}^{h} \left(X_{1+ks} \cdot X_{1+(k+1)s} \right) - 3 \qquad \text{mit } h = \lfloor (n-1)/s \rfloor - 1$$

Einsetzen der konkreten Werte x_i statt der Zufallsvariablen X_i liefert den konkreten Schätzwert $\hat{\rho}(s)$.

Für unabhängige [0,1)-verteilte X_i ist $\tilde{\rho}(s)$ normalverteilt mit Erwartungswert 0 und Standardabweichung $(13h+7)^{1/2}/(h+1)$

Testverfahren untersuchen, ob $\rho(s) \approx 0$ gilt, indem

$$z = \rho(s) \cdot (h + 1) / (13h+7)^{1/2}$$

gegen die kritischen Werte einer N(0,1)-Verteilung getestet wird

Theoretische Testverfahren

Erkennung der Struktur der generierten Sequenzen von ZZs:

Sei x_1, x_2, ... die Sequenz der erzeugten ZZs

Überlappende d-Tupel $(x_1,...,x_d)$, $(x_2,...,x_{d+1})$, ... definieren jeweils Punkte im d-dimensionalen Hyperraum

Beobachtung: Punkte fallen auf „relativ wenige“ Hyperebenen der Dimension d-1

Im zweidimensionalen Fall liegen die Punkte auf einem Gitter

Punkte außerhalb des Gitters sind nicht erreichbar

Anzahl der Gitterlinien bestimmt die Qualität des Generators

Verfahren wie der Spektraltest oder Gittertest bestimmen die Abdeckung des Hyperraums durch den Generator!

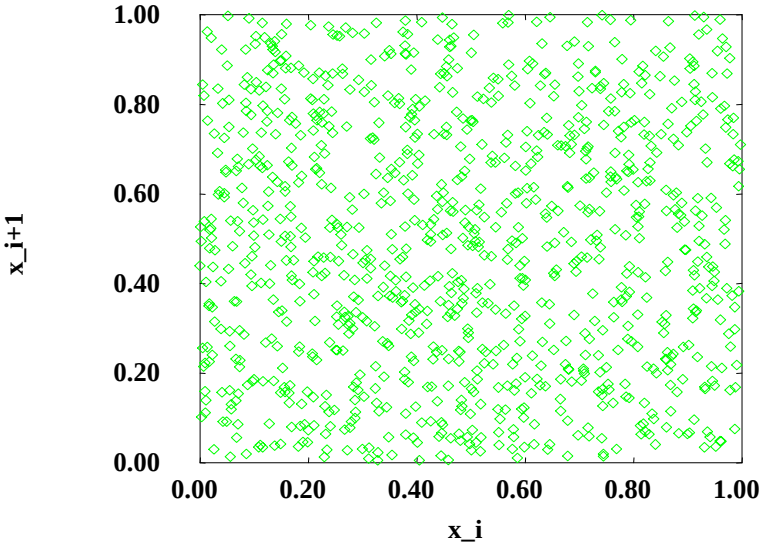
Visuelle Darstellung durch Tupel ..., (x_i, x_{i+1}), (x_{i+2}, x_{i+3}), ...

oder Tripel ..., (x_i, x_{i+1}, x_{i+2}), $(x_{i+3}, x_{i+4}, x_{i+5})$, ...

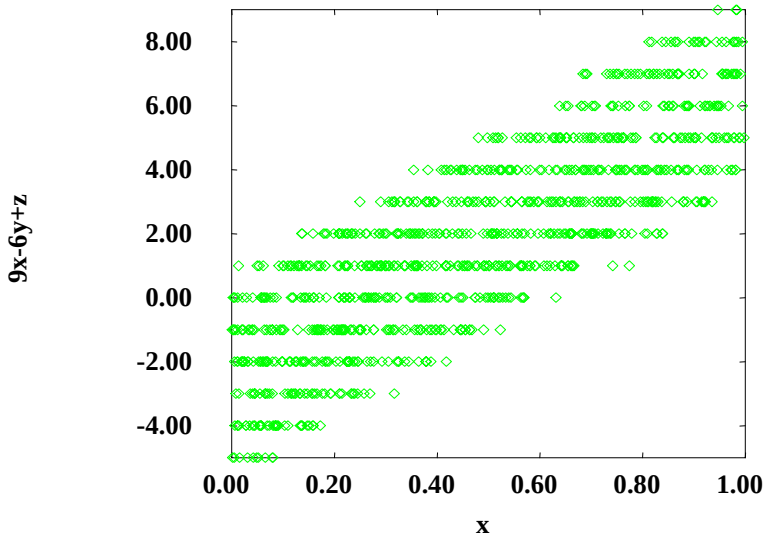
Beispiel für einen schlechten Generator RANDU

$$(x_{i+1} = 65539 \cdot x_i)(\mathrm{mod}\ 2^{31})$$

Darstellung für Tupel



Darstellung von Tripeln (x,y,z) als 9x-6y+z



Alle Werte liegen auf einer von 15 Hyperebenen, die durch die Gleichung 9x-6y+z gegeben sind

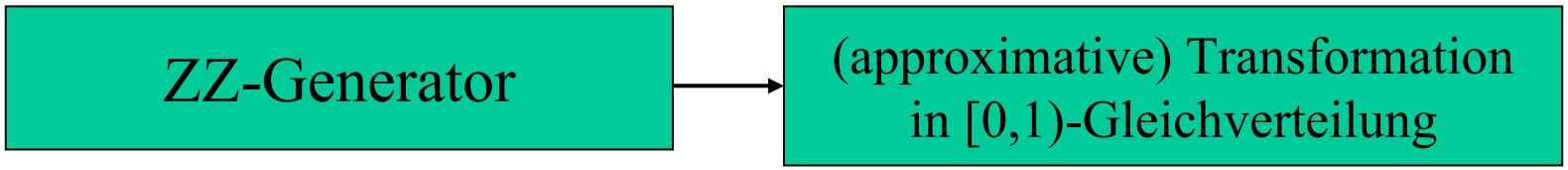
Bei einem guten Generator würden die Werte in einem Band um die Diagonale verteilt liegen!

Aussagen zu Testverfahren un der Qualität von Generatoren:

- Es existiert eine Vielzahl von Testverfahren, aber es ist unklar, welcher Test das beste Ergebnis liefert
- Test können keine beweisbar guten Zufallsgeneratoren liefern, sondern nur schlechte aussondern
- Einige Generatoren wurden aufwändig getestet und haben sich bzgl. dieser Tests als „gut“ erwiesen. Möglichst diese Generatoren in der Simulation verwenden.
- Durch Nutzung mehrerer unterschiedlicher Generatoren, können Verzerrungen durch einzelne Generatoren aufgedeckt werden.

Generierung von Zufallszahlen der gewünschten Verteilung

Bisherige Schritte



Also $FU(u) = \begin{cases} 0 & \text{falls } u < 0 \\ u & \text{falls } 0 \le u < 1 \\ 1 & \text{falls } u \ge 1 \end{cases}$ $fU(u) = \begin{cases} 1 & \text{falls } 0 \le u < 1 \\ 0 & \text{sonst} \end{cases}$

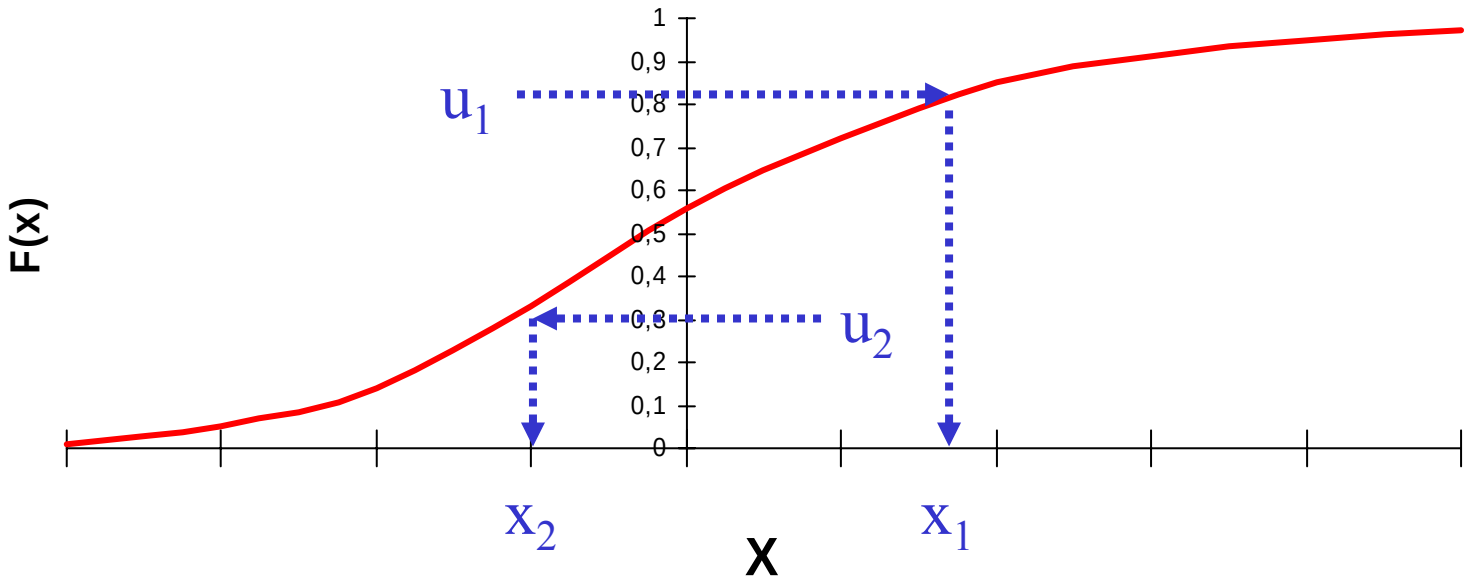
Benötigt werden aber ZZs mit Verteilungsfunktion FX(x)
$\Rightarrow$Transformation der [0,1)-gleichverteilten ZZs u_i in ZZs x_i, die nach FX(x) verteilt sind

Wir beginnen mit kontinuierlichen Verteilungen, sowie für $x_1<x_2$ mit $0<F(x_1)\leq F(x_2)<1$ gelte auch $F(x_1)< F(x_2)$

Damit existiert die Umkehrfunktion F^{-1} von F

Naheliegendes Vorgehen (**Inverse Transformation**):

1. Generiere u aus [0,1)-Gleichverteilung
2. Transformiere $x = F^{-1}(u)$



Formal gilt: $P[X\leq x] = P[F^{-1}(u)\leq x] = P[u\leq F(x)] = F(x)$

Beispiel: Exponentialverteilung

Für $x \geq 0$: $F(x) = 1 - e^{-\lambda x}$ Sei u ZZ aus (0,1)-Gleichverteilung

Transformationsschritte:

$$1 - e^{-\lambda x} = u \Rightarrow e^{-\lambda x} = 1 - u \Rightarrow -\lambda x = \ln(1-u) \Rightarrow x = (\ln(1-u))/-\lambda$$

Da 1-u ebenfalls (0,1)-gleichverteilt, kann es durch u ersetzt werden!

Erzeugung exponentiell vert. ZZs:

1. Generiere u aus (0,1)-Gleichverteilung
2. Transformier $x = \ln(u)/-\lambda$

Funktioniert dies für alle kontinuierlichen Verteilungen?

Nein, $F^{-1}(x)$ muss in geschlossener Form vorliegen oder einfach berechenbar sein! (gleich mehr dazu)

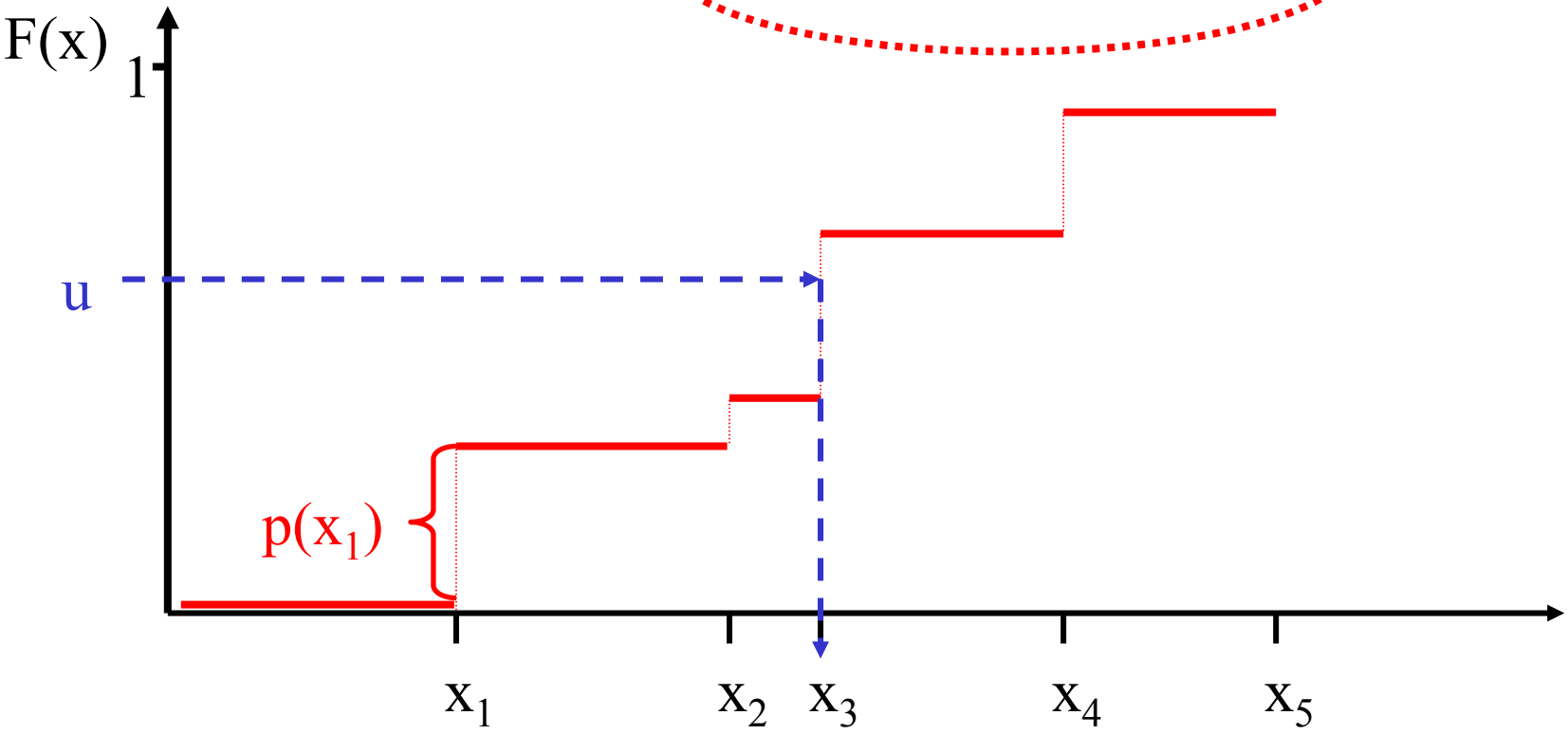
Inverse Transformation für diskrete Verteilungen

Es gilt $F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$

Datenstrukturen zum effizienten Suchen verwenden

Generiere u aus [0,1)-Gleichverteilung

Finde ganze Zahl I, so dass $\sum_{i<I} p(x_i) < u \leq \sum_{i \leq I} p(x_i)$



Konvolutionsverfahren

Falls ZV X als Summe von ZVs X_i (i=1,...I) mit Vfkt. $F_i(x)$ darstellbar ist und Generatoren für die X_i existieren, so können Realisierungen von X wie folgt generiert werden:

```
x = 0 ;
for i=1 to I do
        ziehe ZZ y aus F_i(y) ;
        x = x + y ;
end for
return (x) ;
```

Beispiel: Erlang-Verteilung

Kompositionsverfahren

Falls ZV X eine Vfkt F(x) hat, die sich wie folgt darstellen lässt

$F(x) = \Sigma_{i \le I} p_i \cdot F_i(x)$

und Generatoren existieren, die Realisierungen X_i aus $F_i(x)$ erzeugen, so können Realisierungen von X wie folgt generiert werden:

```
generiere i gemäß Verteilung p_i ;
ziehe ZZ x aus F_i(x) ;
return(x) ;
```

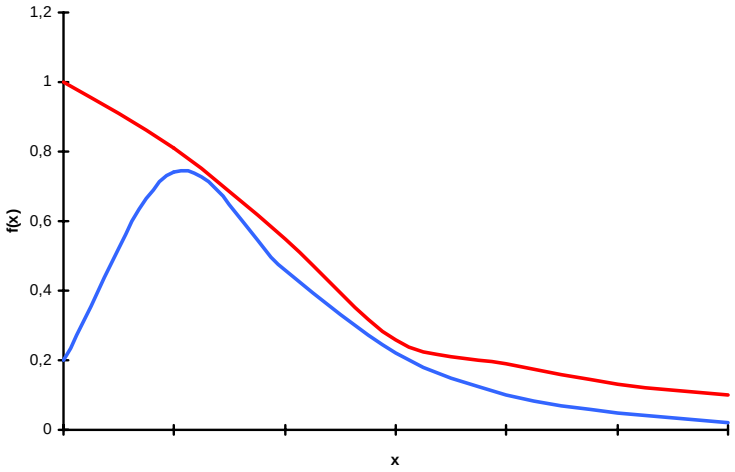
Beispiel: Hyperexponentialvert.

Verwerfungsmethode

Anwendbar für diskrete und kontinuierliche Verteilungen, wir betrachten den kontinuierlichen Fall!

Ges: Realisierungen ZV X mit bekannter Dfkt. $f_X(x)$
Voraussetzung: Generator für ZV Y mit Dfkt. $f_Y(x)$ bekannt und es existiert $\alpha \in (1,\infty)$ so dass für alle x gilt $f_X(x) \leq \alpha \cdot f_Y(x)$

Skizze:



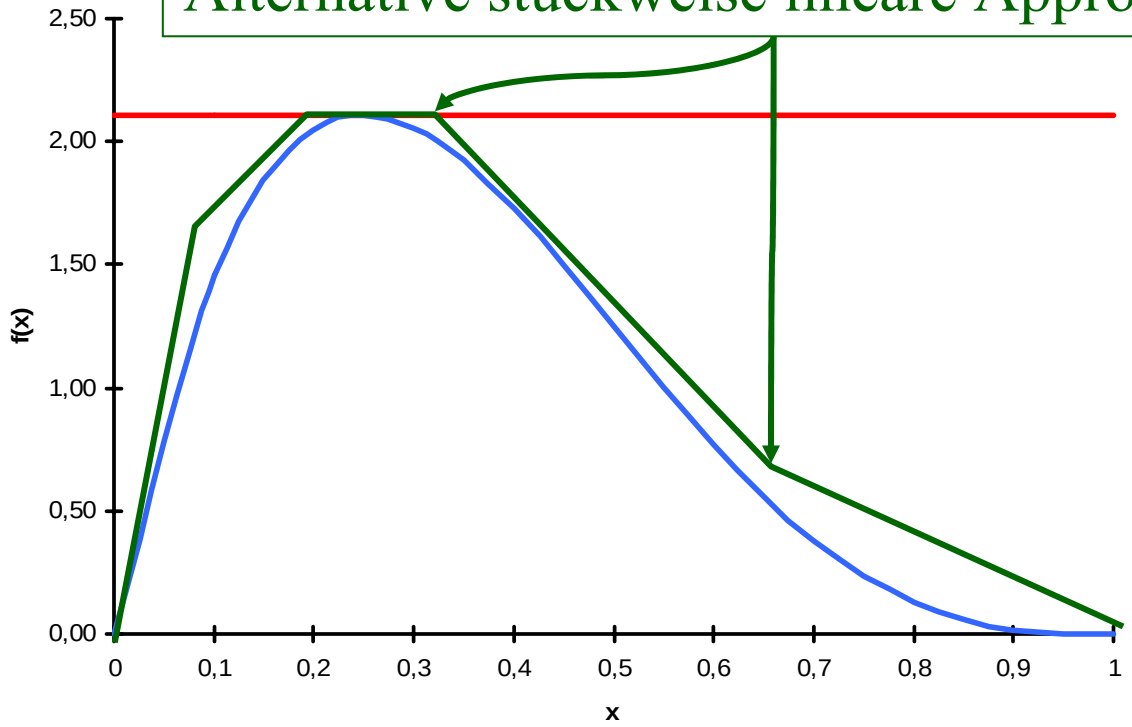
Generierungsmethode:

repeat
  ziehe ZZ y gemäß $f_Y(x)$;
  ziehe ZZ x aus $[0,\alpha \cdot f_Y(y))$-Gleichv.;
until $x \leq f_X(y)$;
return(y) ;

Beispiel: $\beta(2,4)$-Vert. mit $f(x) = 20 \cdot x \cdot (1-x)^3$ falls $0 \le x \le 1$ und 0 sonst

Dichtefunktion wird beschränkt durch Rechteck der Höhe 2.11

Alternative stückweise lineare Approximation



Generierungsmethode:

repeat
  ziehe y aus [0,1)-Gl.Vert. ;
  ziehe x aus [0,2.11)-Gl.Vert.;
until $x \le 20 \cdot y \cdot (1-y)^3$;

Effizienz der Methode hängt ab von

1. der W., dass eine ZZ akzeptiert wird $1 - \int(\alpha \cdot f_Y(x) - f_x(x))dx/\alpha$
2. der Effizienz der Generierung von ZZs mit Dfkt. $f_Y(x)$

Generierung von normalverteilten ZZs

Sei X eine N(0,1)-verteilte ZV, dann ist

$Y = \mu + \sigma \cdot X$ eine (μ,σ)-verteilte ZV

Methode zur Generierung von N(0,1)-verteilten ZZ ist ausreichend!
Verteilungsfunktion und auch inverse Verteilungsfunktion der Normalverteilung haben keine geschlossene Darstellung
$\Rightarrow$ inverse Transformation nicht einsetzbar!

Konvolution

(nach dem zentralen Grenzwertsatz!)

$$x = \frac{\left(\sum_{i=1}^{n} u_i\right) - n/2}{\sqrt{n/12}} \qquad \text{(oft n=12)}$$

u_i aus [0,1)-Gleichverteilung
$\Rightarrow$ x ist aus N(0,1)

Methode von Box-Muller (1958)

$$x_1 = \cos(2\pi u_1)\sqrt{-2\ln(u_2)}$$

$$x_2 = \sin(2\pi u_1)\sqrt{-2\ln(u_2)}$$

u_i aus [0,1)-Gleichverteilung
$\Rightarrow$ x_i aus N(0,1)

Weitere Methoden existieren !!

Generierung von abhängigen Zufallsvariablen

Viele Parameter sind in der Realität korreliert, z.B.:

- Größe und Gewicht von Menschen
- Temperatur und Regenmenge
- Ein-/Ausgabeoperationen und CPU-Zeitbedarf

Dadurch bedingte Probleme:

- Verwendung unabhängiger Zufallsvariablen verfälscht Verhalten
- mehrdimensionale Verteilung muss spezifiziert werden
  - Abhängigkeiten sind schwer zu schätzen
  - in allgemeiner Form nicht kompakt darstellbar

Formale Darstellung als Zufallsvektor $(X_1,X_2,...,X_d)$ mit Verteilungsfunktion $F_{X_1,X_2,...,X_d}(x_1,x_2,...,x_d)$

Alternativ: Darstellung als bedingte Verteilung mit $F_i(x_i \mid x_1,...,x_{i-1})$

- falls bedingte Verteilungen bekannt, dann Generierung einfach
- i.d.R. ist diese detaillierte Information aber kaum ermittelbar

Beispiel bivariate Normalverteilung

X_1 und X_2 sind korrelierte normalverteilte ZVs

- mit Erwartungswert μ_i und Standardabweichung σ_i
- und Korrelation $\rho = COV(X_1,X_2) / (\sigma_1\sigma_2)$

Erzeugung der ZZs:

1. Erzeuge z_1 und z_2 als unabhängige N(0,1) verteilte ZZs (z.B. mit der Box-Muller-Methode)
2. $x_1 = \mu_1 + \sigma_1 z_1$
3. $x_2 = \mu_2 + \sigma_2(\rho z_1 + (1-\rho^2)^{1/2} z_2)$

- Generalisierung auf multivariate Normalverteilungen möglich
- für andere Verteilungen sind andere Methoden notwendig
- oft verwendet stochastische Prozesse (MA-,AR-Modelle) etc.

2.4 Modellierung von Eingabedaten

- Repräsentation der Parameter spielt zentrale Rolle in jedem Simulationsprogramm!
- Oft werden stochastische Größen zur Abbildung der Realität verwendet

Wie kommt man an Verteilungen und Parameter?

Wird in diesem Abschnitt untersucht

Zwei wesentliche Quellen:

1. a priori Wissen
   - aus vorherigen Modellierungen
   - aus ähnlichen Modellen
   - aus Erfahrung
   - aus der Theorie
   - ...
2. Messungen (d.h. Stichproben)
   - am realen System
   - an ähnlichen Systemen
   - an anderen Modellen
   - ...

Schritte bei der Modellierung von Eingabedaten

1. Datensammlung
2. Entscheidung über die Darstellung der gemessenen Daten
    a. Deterministische Größe
    b. Diskrete empirische Verteilung
    c. Kontinuierliche empirische Verteilung
    d. Stochastische Verteilung

Falls 2d. gewählt

3. Auswahl eines Verteilungstyps
4. Schätzung der Verteilungsparameter
5. Überprüfung der Passgüte durch Anpassungstest

Sammlung/Messung von Daten

Datenerhebung ist aufwändig und oft frustrierend, aber eminent wichtig für alle nachfolgenden Schritte!

GIGO-Prinzip (garbage-in garbage-out)

Probleme bei der Datenerhebung

- Zu wenige Daten
  - geringer Stichprobenumfang
  - nur summarische Statistiken
  - lediglich qualitative Informationen
- Zu viele Daten
  - Daten aus automatischen Messungen
  - vollständige Traces
- Falsche Daten
  - falscher Aggregationszustand (z.B. Monat statt Tag)
  - korrelierte Daten
  - falscher zeitlicher Bezug (z.B. aus der Vergangenheit)
  - falscher räumlicher Bezug (z.B. vom falschen System)
  - ungenauer sachlicher Bezug (z.B. Umsatz statt Nachfrage)

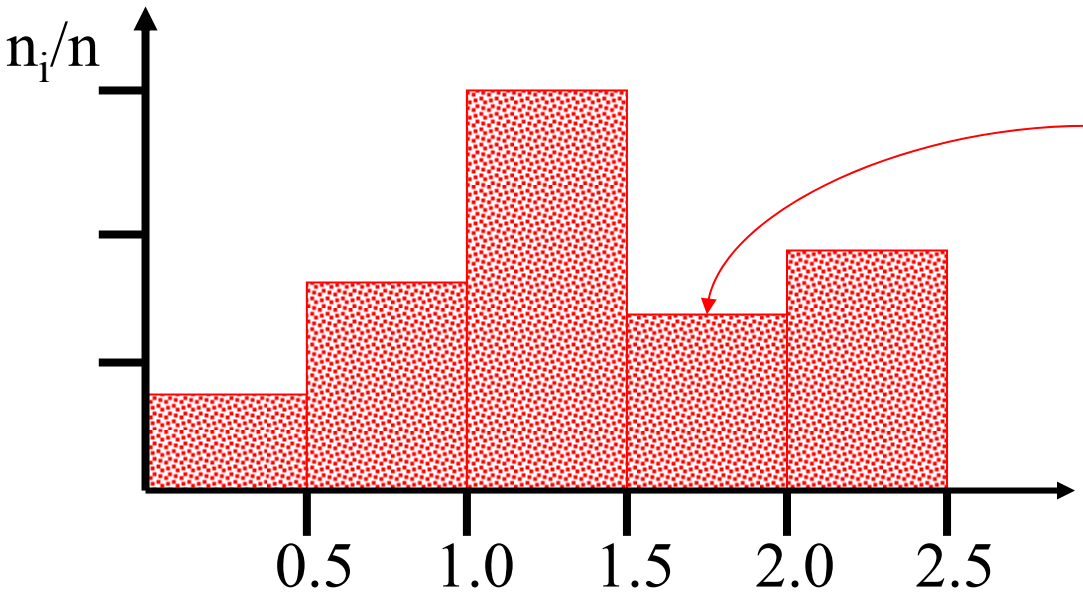
Regeln zur Datenerhebung:

- Vorläufe zur Bestimmung des Erhebungsintervalls, der Auflösung, des Stichprobenumfangs
- Falls möglich Daten schon während der Erhebung analysieren
- Bei mehreren Stichproben erst Homgenität und Herkunft aus einer Verteilung prüfen (z.B. mit Testverfahren)
- Auf zensierte oder verfälschte Daten achten
- Daten auf Korrelation untersuchen
- Zwischen Ein- und Ausgabedaten bei der Messung unterscheiden

Aufbereitung und Repräsentation von Daten

- Daten seien über den Beobachtungszeitraum identisch verteilt (ansonsten Darstellung des Verlauf über der Zeit)
- Zur Interpretation von Daten eignen sich graphische Repräsentationen und Maßzahlen
- Annahme: n Daten x_i nach Erhebungszeit geordnet $y_1,..., y_n$ Stichprobe nach Größe geordnet (d.h. $y_i \leq y_{i+1}$)

Histogramm (als Approximation der Dfkt.)



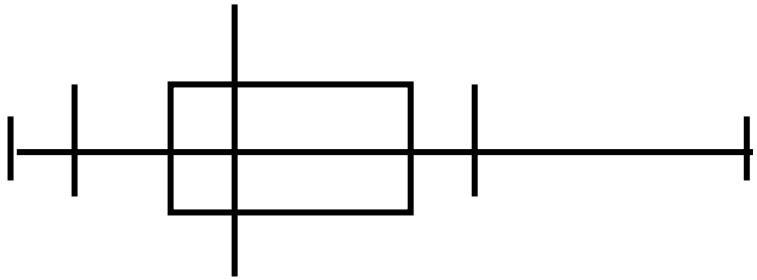
Höhe proportional zur Anzahl Werte im Intervall dividiert durch die Anzahl der Werte der Stichprobe (Schätzer für den Wert der Dichtefunktion)

Intervallbreite frei wählbar, mehrere probieren!

Maßzahlen

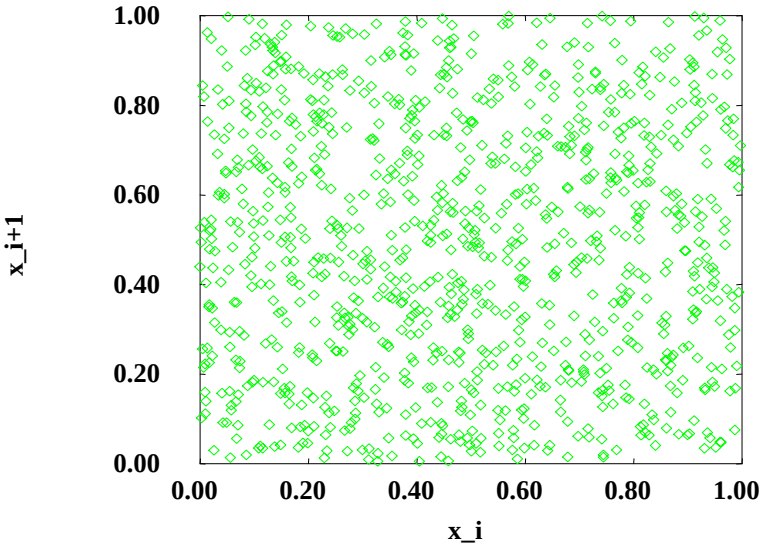
- Momente, Varianz
- Quantilschätzer aus der geordneten Stichprobe $(y_1,\ldots,y_n)$:
  - Median y_i mit $\lfloor i = (n+1)/2 \rfloor$
  - Quartile y_j und y_{n-j+1} mit $j = \lfloor (n+1)/4 \rfloor$
  - Octile y_k und y_{n-k+1} mit $k = \lfloor (n+1)/8 \rfloor$
  - Extremwerte y_1 und y_n

Graphische Darstellung Box-Plots

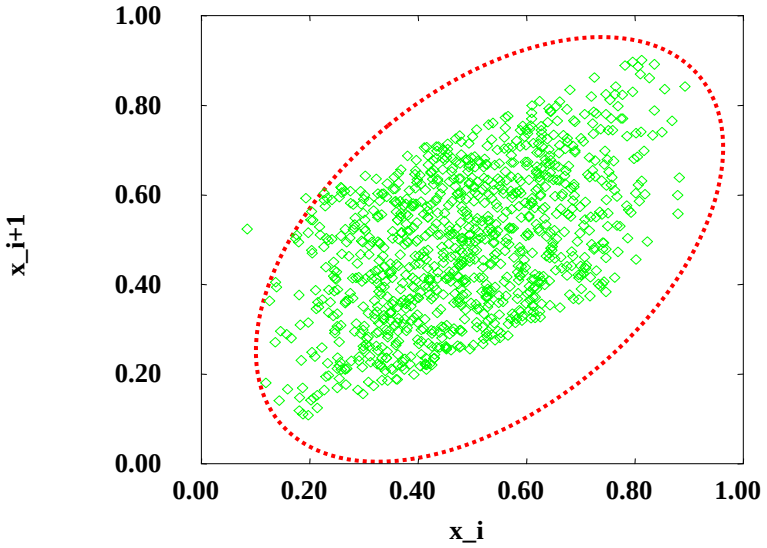


Graphische Darstellung von Abhängigkeiten durch Tupel (x_i, x_{i+1})

keine Korrelation erkennbar



positive Korrelation



- Visueller Eindruck zeigt Korrelationen
- Auch zur Darstellung der Abhängigkeiten zwischen X_i und X_{i+k} (k>1) nutzbar
- Im Prinzip auch 3D-Darstellung möglich

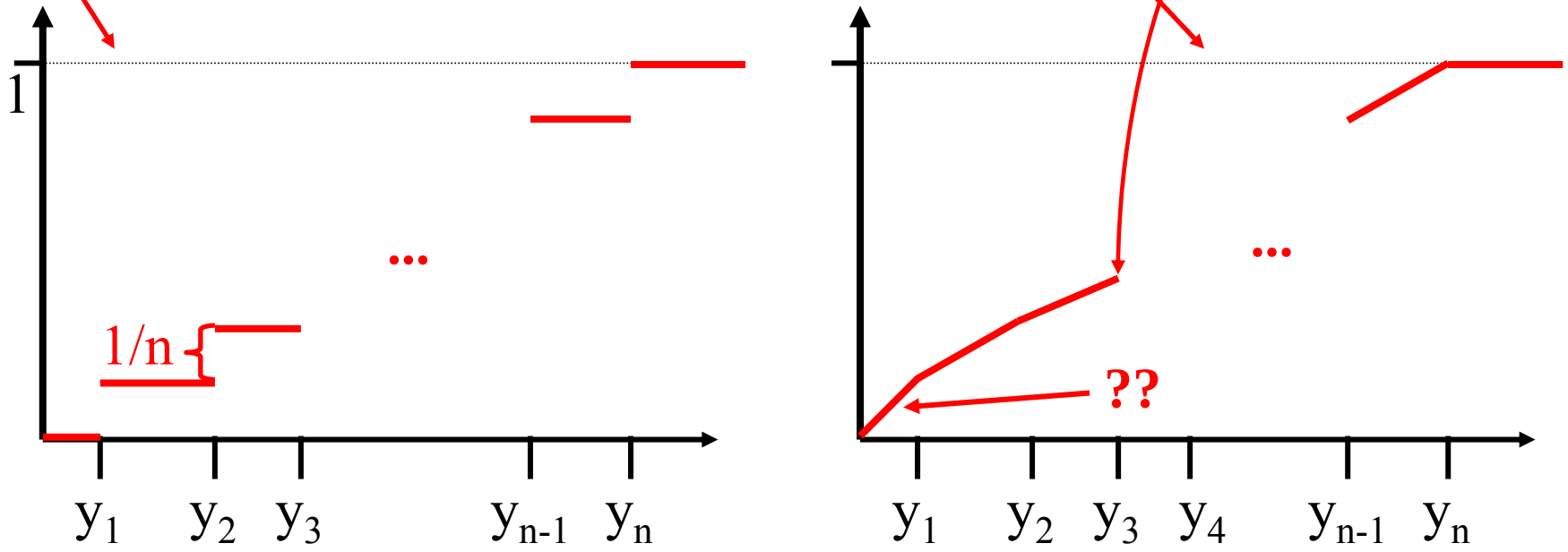
Darstellungsalternativen

Konstante, wenn die Daten alle (fast) identisch sind

Ansonsten:

- stochastische Verteilung
- diskrete empirische Verteilung
- kontinuierliche empirische Verteilung

lineare Interpolation zwischen diskreten Punkten



Für empirische Verteilung spricht:

- Anpassung an theoretische Vfkt. ist immer mit Informationsverlust verbunden
- Wahl der Verteilungsfamilie unklar
- Parameterschätzung oft nicht robust
- Für manche theoretischen Verteilungen existieren keine effiziente Methoden zur ZZ-Generierung

Für theoretische Verteilung spricht:

- Schwankungen können empirischer Vert. verzerren
- Realisierungen nur innerhalb der Bandbreite gemessener Werte
- Kompaktere Darstellung
- Oft effizientere Generierung
- Theoretische fundierte Gründe zur Auswahl eines Verteilungstyps existieren
- U.U. Verwendung alternativer Analyseansätze

Insgesamt kontrovers in der Literatur behandelt
Simulationsmodelle benutzen oft (aus Effizienzgründen) theoretische Verteilungen

Anpassung theoretischer Verteilungen

Auszuführende Schritte:

1. Bestimmung des Verteilungstyps
2. Schätzung der Parameter
3. Bestimmung der Anpassungsgüte

Typische Szenarien

a) Verteilungstyp aus der Theorie, Parameterschätzung aus der Stichprobe
b) Verteilungstyp aus der Stichprobe, Parameterschätzung aus der Stichprobe
c) Weder theoretische Erkenntnisse, noch Stichprobe vorhanden

a) und b) erfordern jeweils Parameterschätzung und Test der Anpassungsgüte

Fall a): Theoretische Erkenntnisse über den Verteilungstyp:

- ZV X, welche aus einer größeren Anzahl zufälliger Ereignisse resultiert, könnte normalverteilt sein (zentraler Grenzwertsatz)
- ZV X, welche das Minimum einer größeren Anzahl zufälliger Ereignisse ist, könnt Weibull-verteilt sein
- ZV X, welche zeitliche Abstände aufeinanderfolgender Ereignisse darstellt könnte exponentiell-verteilt sein, wenn anzunehmen ist, dass Ereignisse
  - einzeln auftreten
  - mit konstanter Rate λ auftreten
- ZV X, welche das Produkt einer größeren Anzahl zufälliger Einflüsse ist, könnte log-normal-verteilt sein
- ...

Zusätzlich oft Erkenntnisse aus dem Anwendungsgebiet

Fall b): Bestimmung des Verteilungstyps aus der Stichprobe

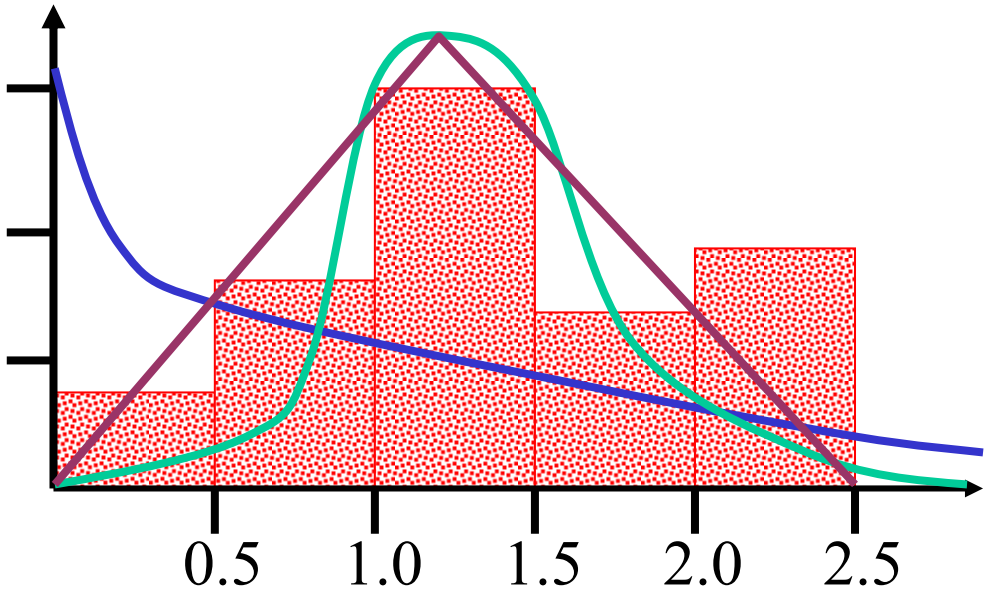
Erste Hinweise auf Ausschluss bestimmter Verteilungstypen auf Basis von Verteilungscharakteristika

- Variationskoeffizient VK(Y)= σ(Y) / E(Y)
  Ausschluss bestimmter Verteilungstypen z.B.
  Exponentialverteilung hat VK 1 ⇒ VK(Y) deutlich von 1 abweichend, keine Exponentialverteilung wählen
  (Schätzung von σ(Y) und E(Y) später)
- andere Charakteristika
  - Verhältnis Erwartungswert zu Median
  - höhere Momente (z.B. Schiefe $\upsilon = \frac{E((Y - E(Y))^3)}{\left(\sigma^2\right)^{3/2}}$)
  - ...

Ausschluss von Verteilungstypen, i.a. aber **keine** Festlegung auf einen Verteilungstyp!

Bestimmung des Verteilungstyps aus dem Histogramm der Stichprobe

Histogramm ist erwartungstreuer Schätzer des Dichtefunktion!



- Visueller Vergleich Dfkt. – Histogramm auf Basis der Form (also erst einmal ohne Kenntnis der Parameter)

Basis: Erfahrung/Wissen

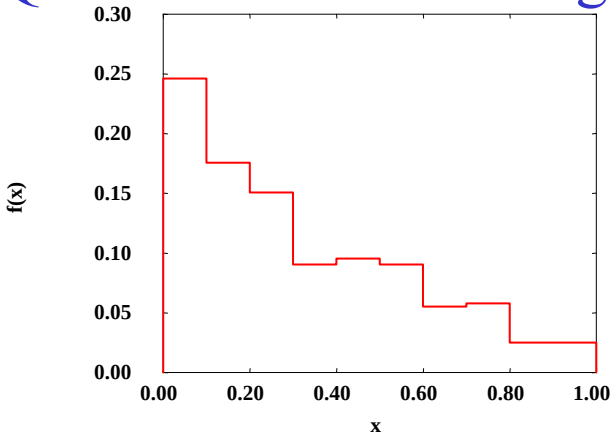
Heute Unterstützung durch Softwaretools (ExpertFit, Arena Input-Analyzer, ..) mit teilweiser automatischer Verteilungswahl und Parameterschätzung

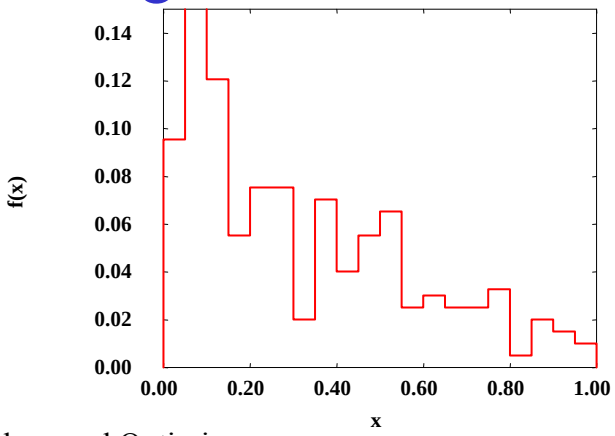
Vorsicht: Ergebnis kann durch (frei wählbare) Histogrammparameter beeinflusst werden!

Freiheitsgrade bei der Histogrammerstellung

- Anzahl und Breite der Zellen
  - i.d.R. Zellen gleicher Breite, ansonsten Höhe anpassen
  - Zahl der Zellen so wählen, dass
    - Form der Dichte erkennbar
    - jede Zelle mehrere Werte enthält (≥ 10)
- überdeckter Bereich
  - Start der ersten Zelle, Ende der letzten Zelle
  - Behandlung von Werten außerhalb des Histogramms (Ausreißern)

Quantitative Bewertung der Anpassung erst nach Parameterschätzung





Parameterschätzung

Jede Verteilungsfamilie weist gewisse Parameter auf:

- Exponentialverteilung: Rate λ
- Normalverteilung: Mittelwert μ, Standardabweichung σ
- Dreiecksverteilung: linke Grenze a, rechte Grenze b, Modalwert c
- ...

Allgemeine Form der Dichtefunktion $f(x,\Theta)$
mit Parametervektor $\Theta=(\Theta_1, ..., \Theta_p)$

Ziel: Bestimme die Werte von Θ so, dass die Dichtefunktion der Verteilung und die Stichprobe „möglichst gut korrespondieren“.

Zahlreiche Methoden existieren, wir betrachten

- Momentenmethode
- Maximum-Likelihood-Methode

Momentenmethode

Zur Notation:

- $(y_1, ..., y_n)$ ist die konkrete Stichprobe
- jeder Wert y_i ist eine Realisierung der ZV Y_i
  (alle Y_i sind identisch verteilt!)

Sei $\tilde{Y}^i$ Schätzer für $E(Y^i)$ und $\hat{Y}^i$ der konkrete Schätzwert
Entsprechend definieren wir

- $\tilde{S}^2$ und $\hat{S}^2$ als Schätzer und Schätzwert für die Varianz
- $\tilde{v}$ und $\hat{v}$ als Schätzer und Schätzwert für den Variationskoeffizienten

Erwartungstreue Schätzer:

$$\tilde{Y}^i = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j\right)^i \text{ und } \tilde{S}^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j - \tilde{Y}^1\right)^2$$

Parameter Θ_j lassen sich oft als Funktion der Momente darstellen

$$\Theta_j = \phi_j\left(E(Y^1),...,E(Y^p)\right)$$

Momentenmethode substituiert die Momentenschätzer für die Momente und liefert damit einen Parameterschätzer

$$\tilde{\Theta}_j = \phi_j\left(\tilde{Y}^1,...,\tilde{Y}^p\right)$$

Schätzer oft nicht erwartungstreu, aber asymptotisch erwartungstreu und konsistent

Trotzdem oft keine guten Schätzer, da die „Form“ der Verteilung nicht berücksichtigt wird

Beispiele:

Exponentialverteilung $E(Y^1)=\lambda^{-1}$

$$\Rightarrow \tilde{\lambda} = 1/\tilde{Y}^1 \left(= n/\left(\sum_{j=1}^{n} Y_j\right)\right)$$

Normalverteilung:

$E(Y^1)=\mu$, $\sigma^2 = E(Y^1)^2-E(Y^2)$

$$\Rightarrow \mu = \tilde{Y}^1 \text{ und } \sigma = \tilde{S}$$

Maximum-Likelihood-Methode (ML-Methode)

Suche nach den plausibelsten Parametern

Vorstellung der ML-Methode am Beispiel:

- Diskrete Verteilung mit Parameter θ, so dass $p_\theta(x)$ W. für Wert x bei Parameter θ
- Stichprobe $(y_1, ..., y_n)$
- Likelihoodfunktion $L(\theta) = p_\theta(y_1) \cdot ... \cdot p_\theta(y_n)$
- Ziel der ML-Methode: Wähle θ so, dass $L(\theta)$ maximal
- also $\max_\theta(L(\theta))$ Punkt der maximalen Beobachtungswahrscheinlichkeit
- analoges Vorgehen bei Parametervektor (mehrdimensionales Optimierungsproblem)

Kontinuierlicher Fall nicht ganz so intuitiv, da Wahrscheinlichkeit in jedem Punkt 0 ist
Verwendung der Dichtefunktion $f_\theta(x)$ mit Parameter θ

$$L(\theta) = \prod_{j=1}^{n} f_\theta(y_j)$$ finde θ_{max}, so dass $L(\theta_{max}) \geq L(\theta)$ für alle θ

Beispiel Exponentialverteilung:

$$L(\lambda) = \prod_{j=1}^{n} \lambda \cdot e^{-\lambda \cdot y_j} = \lambda^n \cdot e^{-\lambda \cdot \sum_{j=1}^{n} y_j}$$

Optimierung des natürlichen Logarithmus ist einfacher (natürlicher Logarithmus ist eine Transformation, welche die Lage des Optimums nicht verändert)

$$l(\lambda) = \ln(L(\lambda)) = n \cdot \ln \lambda - \lambda \cdot \sum_{j=1}^{n} y_j \qquad l'(\lambda) = n/\lambda - \sum_{j=1}^{n} y_j$$

Nullstelle der Ableitung $\hat{\lambda} = n / \left(\sum_{j=1}^{n} y_j \right)$

Schätzer in diesem Fall identisch zur Momentenmethode

Verfahren auch für Parametervektoren anwendbar

Allgemein Maximierung von l(θ) statt L(θ)

$$l(\theta) = \ln(L(\theta)) = \sum_{j=1}^{n} \ln(f_\theta(y_j))$$

Im Allgemeinen Optimierungsproblem ohne geschlossene Lösung
$\Rightarrow$ Anwendung von (nichtlinearen) Optimierungsverfahren

Eigenschaften der ML-Schätzer:

1. Asymptotisch erwartungstreu
2. Konsistent
3. Asymptotisch normalverteilt (Berechnung von Konfidenzintervallen)
4. I.d.R. nicht schlechter bzw. besser als Momentenschätzer

ML-Schätzer für die Normalverteilung: $\mu = \tilde{Y}^1$ und $\sigma^2 = \frac{n-1}{n}\tilde{S}^2$

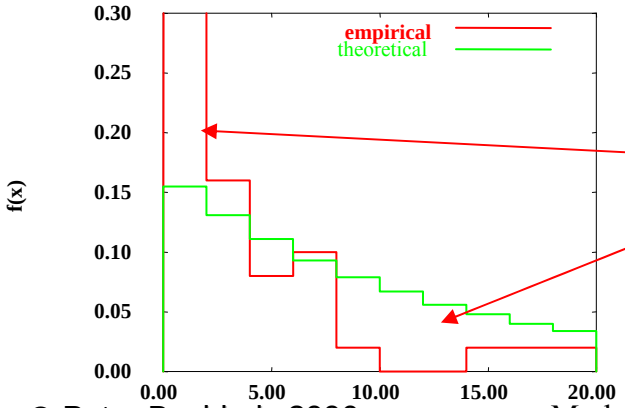
Bestimmung der Anpassungsgüte

Nach Verteilungsbestimmung und Parameterschätzung kann die Anpassungsgüte der theoretischen Verteilung quantifiziert oder getestet werden

Methoden zur Quantifizierung

Vergleich der Histogramme

- n_i Anzahl Werte der Stichprobe im i-ten Intervall und $h_i = n_i/n$
- $p_i = \int_{\Delta_i} f(x)\,dx$ wobei Δ_i das i-te Intervall ist
- Abstandsmaß $D = \Sigma_i | h_i - p_i|$ oder $D' = \Sigma_i (h_i - p_i)^2$



Differenz der Histogramme

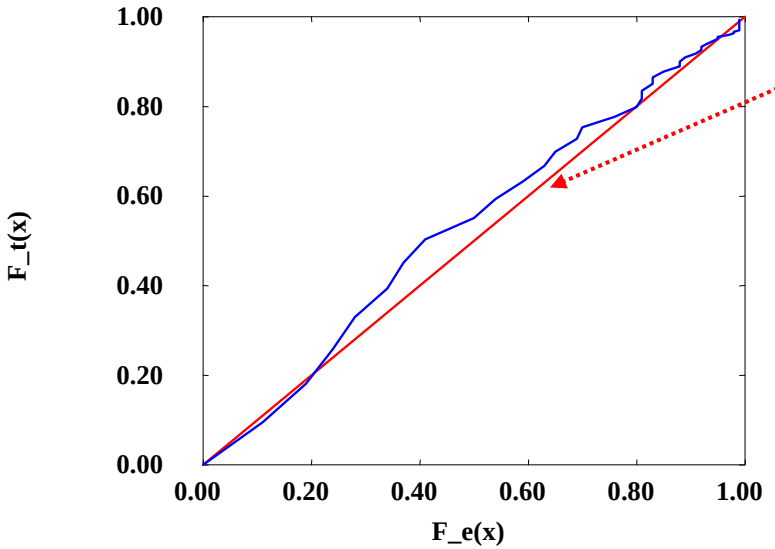
Wahrscheinlichkeitsplots

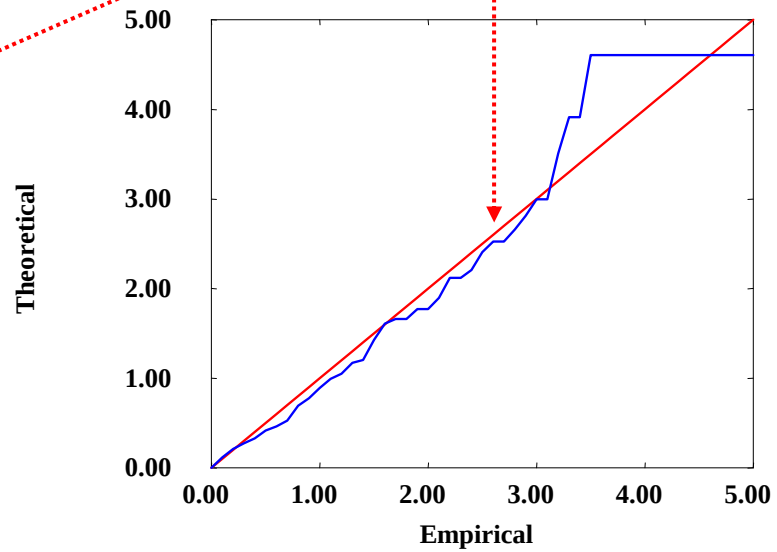
Grundidee: Vergleich der Werte der Verteilungsfunktion der empirischen und theoretischen Verteilung

- Für die empirische Verteilung gilt $F_e(y) = \max_{y_i \le y}(i/n)$
- $F_t(y)$ sei der Wert der theoretischen Verteilungsfunktion
- Q-Q-Plot Darstellung der Quantile (y_i, x_i) mit $F_e(y_i) = F_t(x_i)$
- P-P-Plot Darstellung $(F_e(y_i), F_t(y_i))$

y_i geordnet

Abweichung von einer Geraden beschreibt Abweichungen der Vfkts.





Anpassungstests

Zentrale Frage: Wann ist die Modellierung der Stichprobe durch eine theoretische Verteilung als adäquat anzusehen?

- Bisherige Ansätze erlauben Bewertung der Anpassung auf Basis des visuellen Vergleichs oder ausgesuchter Maßzahlen
- Auf Grund stochastischer Schwankungen ist zu erwarten, dass empirische und theoretische Verteilung immer Abweichungen aufweisen (müssen)
- Bleibt die Frage, welche Abweichungen (noch) tolerierbar sind

Alternative/Ergänzung zu den bisherigen Maßzahlen sind Tests:
Hypothese H_0: $(y_1,..., y_n)$ wurde aus Verteilung F(x) gezogen
Test liefert Antwort, ob H_0 verworfen oder angenommen werden soll

Chi-Quadrat Test

Sehr altes Testverfahren (ca. 1900) erlaubt einen formalen Vergleich der Histogramme empirischer und theoretischer Verteilungen

Definiere $b_0 < b_1 < ... < b_k$ als Intervallgrenzen

- $n_i = |\{y_j \mid b_{i-1} \leq y_j < b_i\}|$
- $p_i = \int_{b_{i-1}}^{b_i} f(y)dy$

Differenz $\Sigma_{i=1,..,k}|n_i - p_i \cdot n|$ liefert ein Maß für die Abweichung der beiden Histogramme: Je kleiner der Wert, desto wahrscheinlicher ist es, dass die Stichprobe aus der theoretischen Verteilung gezogen wurde.

Um ein Testverfahren anzuwenden, muss eine Teststatistik mit bekannter Verteilung definiert werden!

Teststatistik $d = \sum_{j=1}^{k} \frac{\left(n_j - n \cdot p_j\right)^2}{n \cdot p_j}$ Welche Verteilung hat d?

χ^2-Verteilung:

Seien $Y_1, ..., Y_k$ unabhängig, identisch N(0,1)-verteilte ZVs, dann ist

$$Y = \sum_{i=1}^{k} Y_i$$

χ^2-verteilt mit k Freiheitsgraden.

Familie der –Verteilungen liegt in vertafelter Form vor (keine explizite funktionale Form)

Wert d ist Realisierung einer ZV D

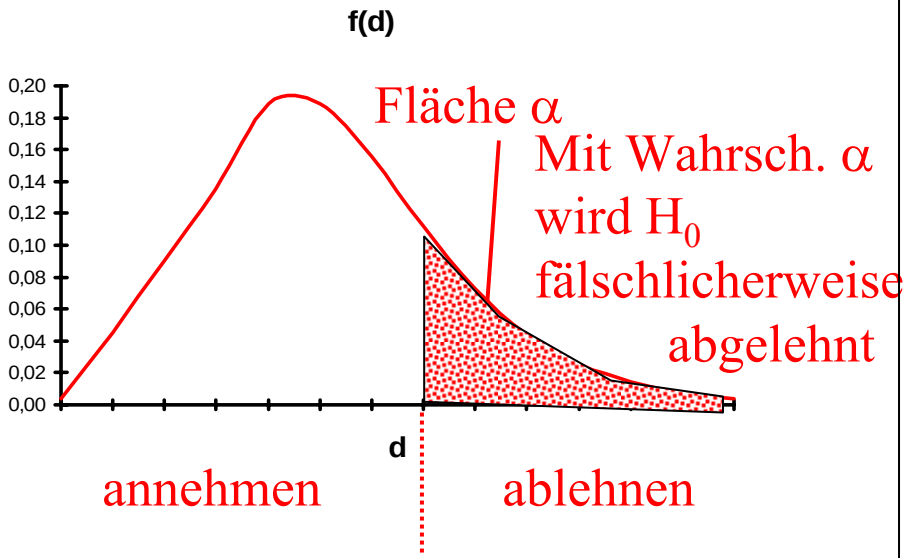
Falls Hypothese H_0 gilt:

- dann ist D asymptotisch χ^2-verteilt (d.h. für „genügend große“ n ist D approximativ χ^2-verteilt)
- Fallunterscheidung zur Bestimmung der Freiheitsgrade
  - falls keine Verteilungsparameter aus der Stichprobe geschätzt wurde mit k-1 Freiheitsgraden
  - falls p Verteilungsparameter aus der Stichprobe geschätzt wurden mit k-p-1 Freiheitsgraden

Vorgehen

- d berechnen und
- mit kritischen Werten zum gewählten Signifikanzniveau vergleichen (vertafelt)
- Hypothese annehmen/ablehnen

Skizze des Vorgehens:



Subjektive Komponenten:
Lage, Größe und Anzahl der Intervalle
(d.h. Festlegung der b_i)

Hinweise

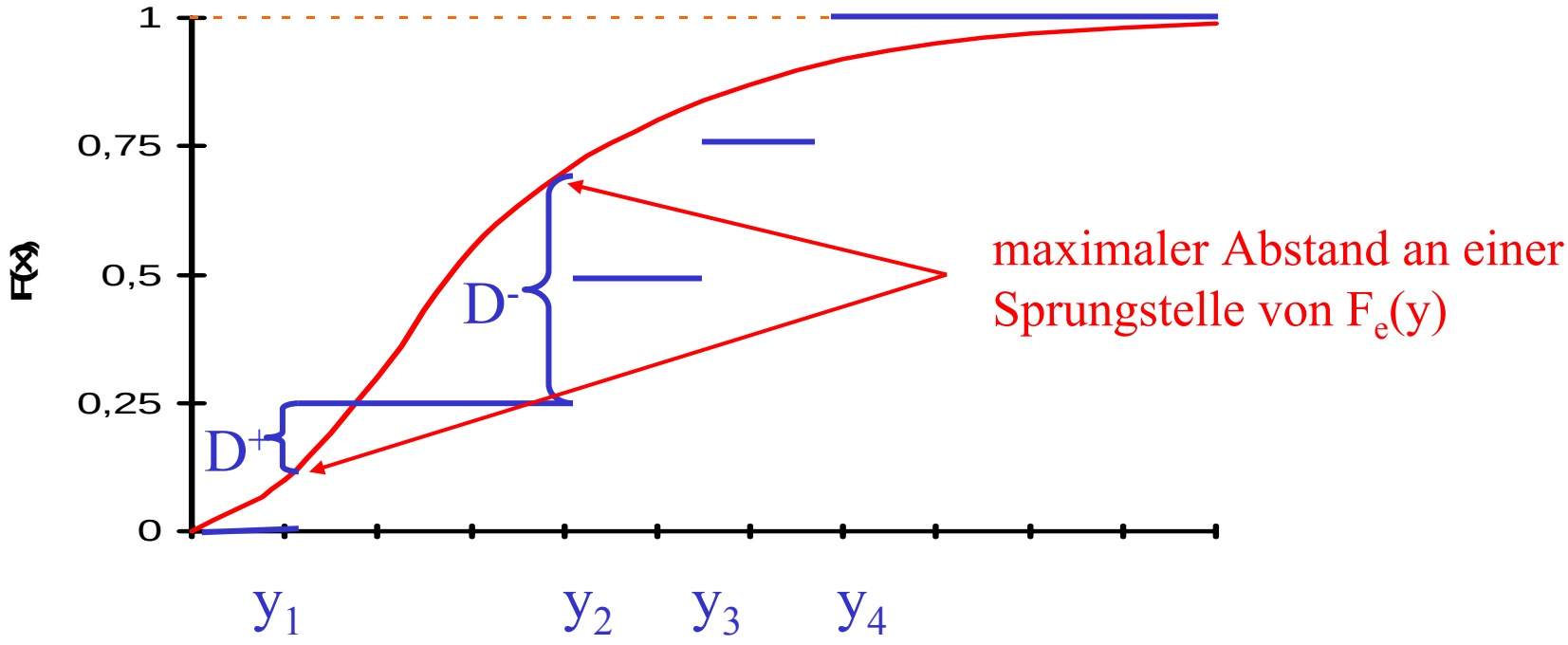
- Wähle Intervalle so, dass Werte p_j identisch/ähnlich sind
  (also unterschiedliche Intervallbreiten)
- Wähle Intervalle so, dass $n \cdot p_j \geq 5$

Kolmogorov-Smirnov Test

Grundidee: Vergleich der empirischen und theoretischen Verteilungsfunktion:

$F_e(y) = |\{y_i \leq y\}|/n$ (Treppenfunktion)

Teststatistik $D_n = \max_y(|F_e(y)-F_t(y)|)$ (falls nötig sup statt max)



Formale Definition von $D_n = \max \{D_n^-, D_n^+\}$ mit

$D_n^+ = \max_{1 \le i \le n} \{i/n - F_t(y_i)\}$ und $D_n^- = \max_{1 \le i \le n} \{F_t(y_i) - (i-1)/n\}$

Großer Wert von D_n deutet auf eine schlechte Anpassung hin

Fallunterscheidung bei der Anwendung des Tests:

- Falls keine Verteilungsparameter aus der Stichprobe geschätzt wurden, ist H_0 zu verwerfen, wenn

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) \cdot D_n \ge c_{1-\alpha}$$ Werte $c_{1-\alpha}$ sind vertafelt

- Falls Verteilungsparameter aus der Stichprobe geschätzt wurden, so sind Teststatistiken nur für spezielle Verteilungen bekannt z.B. Normalverteilung, Exponentialverteilung, Weibull-Verteilung

Fall c): Keine Information und keine Stichprobe:

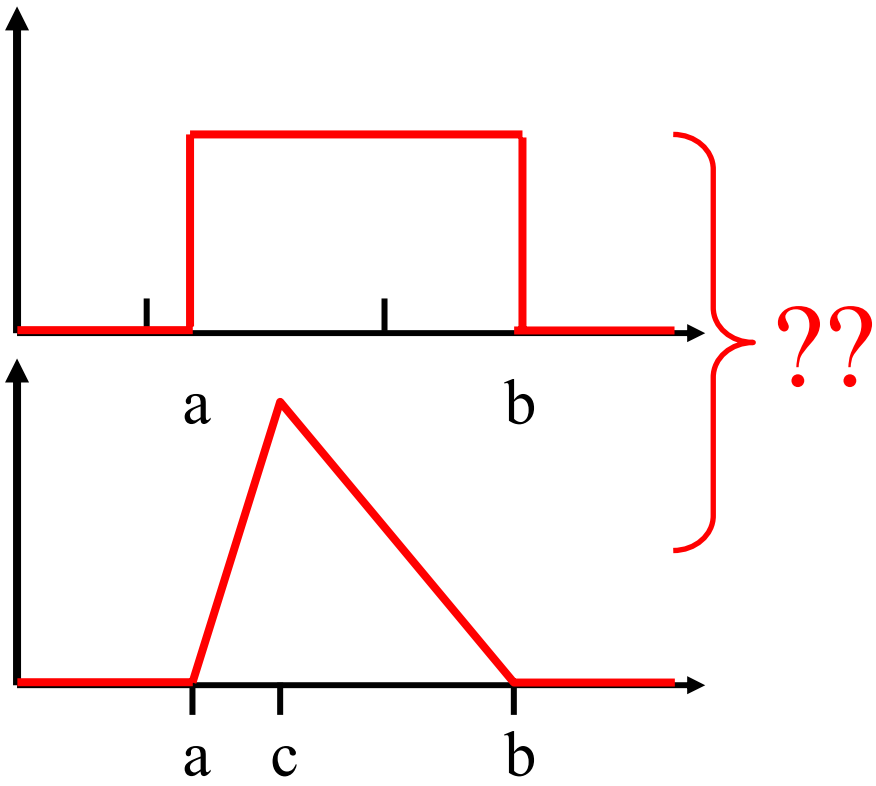
Äußerst „ungemütliche“ Situation, da eigentlich nicht genügend Information vorhanden

Heuristische Ansätze aus der Praxis

1. Raten des minimalen Wertes a und des maximalen Wertes b, so dass $P[x<a] \approx P[x>b] \approx 0$
2. Evtl. zusätzliches Raten des Mittelwertes c

Falls 1. vorliegt wähle [a,b]-Gleichverteilung

Falls 1. und 2. vorliegt wähle Dreiecksverteilung



Weitere Aspekte bei der Modellierung von Eingabedaten

Hier behandelt: unabhängige, identisch verteilte Daten

Reale Daten sind oft

- korreliert bzgl. der Zeitintervalle
  (z.B. Ankunftsprozess im Rechnernetz)
- korreliert bzgl. verschiedener Messgrößen
  (z.B. Größe und Gewicht von Menschen)
- über einen längeren Zeitraum nicht identisch verteilt
  (Ankunftsprozess in einem Restaurant)

In diesen Fällen sind andere Verteilungen zu verwenden
z.B. Zufallsvektoren, Markovsche Ankunftsprozesse, nichtstationäre Poisson Prozesse, autoregressive Modelle, bivariate Normalverteilungen, ...

Beispiel Schätzung der Korrelation einer bivariaten Normalverteilung

$$\widetilde{COV}(X_1, X_2) = \frac{1}{n-1}\sum_{j=1}^{n}(X_{1j} - \tilde{X}_1)(X_{2j} - \tilde{X}_2)$$

$$= \frac{1}{n-1}\left(\sum_{j=1}^{n} X_{1j}X_{2j} - n\tilde{X}_1\tilde{X}_2\right)$$

$$\tilde{\rho} = \frac{\widetilde{COV}(X_1, X_2)}{\tilde{S}_1\tilde{S}_2}$$

Verwendung der Werte zur ZZ-Generierung (siehe Folie 127)
Prinzip auf den n-dimensionalen Fall übertragbar!

Parameterbestimmung im allgemeinen Fall oft schwierig und hier nicht behandelt!