

# Verwendung von Fourier-Koeffizienten zum Testen auf Gleichverteilung

Jan Behrens

1. Juni 2008

Ergänzung/Überarbeitung vom 14. November 2008

Ziel soll es sein, zu einer Stichprobe von Zufallszahlen den eindeutig definierten Wert  $g(h)$  zwischen 0 und 1 zu berechnen, der als Maß für das Abweichen dieser Stichprobe von einer Gleichverteilung dient. Kleine Werte von  $g(h)$  stehen hierbei für eine hohe Abweichung von der Gleichverteilung, während Werte gegen 1 für eine gute Übereinstimmung der Stichprobe mit einer Gleichverteilung stehen. Die Größe  $g(h)$  soll außerdem quantitativ mit dem Signifikanzniveau korrelieren; d.h. bei einer Stichprobe, die einer Gleichverteilung entstammt, soll die Wahrscheinlichkeit, dass  $g(h)$  einen Wert  $x$  unterschreitet genau  $x$  sein.

Hierzu wird die Summe  $h$  der Quadrate von Fourier-Koeffizienten (unter Auslassung des Koeffizienten für den Gleichanteil) gebildet, da diese bei einer konstanten Funktion (entsprechend einer Gleichverteilung) gleich 0 ist. Konvergenz wird dadurch gewährleistet, dass mit zunehmendem Index  $n$  die Quadrate der Fourier-Koeffizienten schwächer gewichtet werden, nämlich mit  $\frac{1}{n^2}$ . Zur Berechnung der Größe  $g(h)$  aus der Summe  $h$  wird diese Summe für den Fall einer zugrundeliegenden Gleichverteilung ( $h_{gleich}$ ) zunächst mit Hilfe des Zentralen Grenzwertsatzes von LINDBERG-LEVY durch eine unendliche Summe von gewichteten, quadrierten normalverteilten Zufallsgrößen angenähert. Die der unendlichen Summe dieser Zufallsgrößen entsprechenden unendlichen Faltungen der Wahrscheinlichkeitsdichten werden anschließend berechnet, um die Wahrscheinlichkeitsdichte  $f(x)$  von  $h_{gleich}$  zu ermitteln. Aus dieser Wahrscheinlichkeitsdichte wird anschließend durch Integration die kumulative Verteilungsfunktion  $F(x)$  und daraus dann eine Formel für die gesuchte Größe  $g(h)$  gebildet.

Abschließend wird noch eine Möglichkeit erwähnt, wie durch eine einfache Transformation der Stichprobe anstelle des Testens auf Gleichverteilung auf jede andere Wahrscheinlichkeitsverteilung mit gegebenen Parametern geprüft werden kann.

Gegeben sei eine Folge  $a$  von  $M$  reellen Zahlen, die im Intervall  $[0; 1]$  liegen. Durch Berechnung von

$$h := \frac{1}{M} \cdot \sum_{n=1}^{\infty} \left[ \frac{1}{n^2} \cdot \left( \left( \sum_{m=1}^M \cos(2\pi \cdot n \cdot a_m) \right)^2 + \left( \sum_{m=1}^M \sin(2\pi \cdot n \cdot a_m) \right)^2 \right) \right]$$

erhalten wir eine Zahl, die umso größer wird, desto ungleich-verteilter die Zahlen der Folge  $a$  im Intervall  $[0; 1]$  sind, da bei einer Gleichverteilung sich die positiven und negative Werte des Sinus und Cosinus in den (inneren) Summen aufheben.

Sind die Zahlen der Folge  $a$  zufällig und genügen einer Gleichverteilung, dann weisen die Zufallsgrößen

$$U_{n,m} := \cos(2\pi \cdot n \cdot a_m)$$

folgenden Mittelwert  $\mu$  und folgende Varianz  $\sigma^2$  auf:

$$\mu = \frac{1}{1-0} \cdot \int_{\Theta=0}^1 \cos(2\pi n \Theta) d\Theta = 0$$

$$\sigma^2 = \frac{1}{1-0} \cdot \int_{\Theta=0}^1 \cos^2(2\pi n \Theta) d\Theta = \frac{1}{2}$$

Wir bilden nun die Zufallsgröße:

$$V_n := \frac{\sum_{m=1}^M U_{n,m}}{\sqrt{M}} = \sum_{m=1}^M \frac{\cos(2\pi \cdot n \cdot a_m)}{\sqrt{M}}$$

$V_n$  nähert sich für  $M \rightarrow \infty$  gemäß dem Grenzwertsatz von LINDBERGLÉVY einer Normalverteilung mit dem Mittelwert  $\mu = 0$  und der Varianz  $\sigma^2 = \frac{1}{2}$  an. Für  $M > 30$  erhalten wir mit der Normalverteilung bereits eine hinreichend gute Näherung. Seien  $X_n$  und  $Y_n$  (0,1)-normalverteilte Zufallsgrößen, dann läßt sich  $V_n$  (näherungsweise) auch darstellen als:

$$V_n = \mu + \sigma X_n = \sqrt{\frac{1}{2}} \cdot X_n$$

Die Zufallsgrößen

$$\begin{aligned} W_n &:= \frac{1}{n^2} \cdot \left( \left( \sum_{m=1}^M \frac{\cos(2\pi \cdot n \cdot a_m)}{\sqrt{M}} \right)^2 + \left( \sum_{m=1}^M \frac{\sin(2\pi \cdot n \cdot a_m)}{\sqrt{M}} \right)^2 \right) \\ &= \frac{1}{M} \cdot \frac{1}{n^2} \cdot \left( \left( \sum_{m=1}^M \cos(2\pi \cdot n \cdot a_m) \right)^2 + \left( \sum_{m=1}^M \sin(2\pi \cdot n \cdot a_m) \right)^2 \right) \end{aligned}$$

ergeben sich entsprechend zu:

$$W_n = \frac{1}{n^2} \cdot \left( \left( \sqrt{\frac{1}{2}} \cdot X_n \right)^2 + \left( \sqrt{\frac{1}{2}} \cdot Y_n \right)^2 \right) = \frac{1}{n^2} \cdot \frac{X_n^2 + Y_n^2}{2}$$

$h$  lässt sich somit bei gleichverteilt-zufälligen  $a_m$  als unendliche Summe von gewichteten, quadrierten  $(0,1)$ -normalverteilten Zufallsgrößen  $X_n$  und  $Y_n$  schreiben:

$$h_{\text{gleich}} = \sum_{n=1}^{\infty} W_n = \sum_{n=1}^{\infty} \frac{1}{n^2} \cdot \frac{X_n^2 + Y_n^2}{2}$$

Die Zufallsgröße  $X_n^2 + Y_n^2$  genügt einer  $\chi^2$ -Verteilung mit 2 Freiheitsgraden, welche die Wahrscheinlichkeitsdichte  $\frac{1}{2} \cdot e^{-\frac{1}{2} \cdot x}$  aufweist. Multipliziert man eine Zufallsgröße mit einem Faktor, so bedeutet dies für die Wahrscheinlichkeitsdichte eine entsprechende Streckung in  $x$ -Richtung sowie eine Stauchung in  $y$ -Richtung. Demnach entsprechen die einzelnen Summanden

$$\frac{1}{n^2} \cdot \frac{X_n^2 + Y_n^2}{2}$$

von  $h_{\text{gleich}}$  den folgenden Wahrscheinlichkeitsdichtefunktionen:

$$n^2 \cdot e^{-n^2 \cdot x}$$

Ein Addieren von Zufallsgrößen entspricht der Faltung (\*) ihrer Wahrscheinlichkeitsdichten. Bezeichnen wir mit  $f(x)$  die Wahrscheinlichkeitsdichte der Zufallsgröße  $h_{\text{gleich}}$  an der Stelle  $x$ , so ist:

$$f(x) = (e^{-x}) * (2^2 \cdot e^{-2^2 \cdot x}) * (3^2 \cdot e^{-3^2 \cdot x}) * (4^2 \cdot e^{-4^2 \cdot x}) * \dots$$

mit

$$r(x) * s(x) = \int_{\Theta=0}^x r(x - \Theta) \cdot s(\Theta) d\Theta$$

Wir berechnen zunächst die Faltung von zwei unterschiedlichen Exponentialfunktionen ( $\alpha \neq \beta$ ):

$$\begin{aligned}
 (A \cdot e^{-\alpha x}) * (B \cdot e^{-\beta x}) &= \int_{\Theta=0}^x A \cdot e^{-\alpha(x-\Theta)} \cdot B \cdot e^{-\beta\Theta} d\Theta \\
 &= AB \cdot \int_{\Theta=0}^x e^{-\alpha x + \alpha\Theta - \beta\Theta} d\Theta = AB \cdot e^{-\alpha x} \cdot \int_{\Theta=0}^x e^{(\alpha-\beta)\Theta} d\Theta \\
 &= AB \cdot e^{-\alpha x} \cdot \frac{1}{\alpha - \beta} \cdot (e^{(\alpha-\beta)x} - e^{(\alpha-\beta)\cdot 0}) = \frac{AB}{\alpha - \beta} \cdot (e^{-\beta x} - e^{-\alpha x})
 \end{aligned}$$

Wir stellen fest, dass die Faltung zweier Exponentialfunktionen mit unterschiedlich skaliertem Exponenten eine Linearkombination dieser Exponentialfunktionen ergibt.

Faltet man eine Summe von Exponentialfunktionen mit einer weiteren Exponentialfunktion  $A \cdot e^{-\alpha x}$ , so ist der Koeffizient  $B$  eines Summanden  $B \cdot e^{-\beta x}$  im Ergebnis mit  $\frac{A}{\alpha-\beta}$  zu multiplizieren. Aus diesem Grunde lässt sich  $f(x)$  für  $x > 0$  schreiben als:

$$f(x) = \sum_{n=1}^{\infty} \underbrace{\left( \prod_{m \in \mathbb{N} \setminus \{n\}} \frac{m^2}{m^2 - n^2} \right)}_{=: p(n)} (n^2 \cdot e^{-n^2 \cdot x})$$

Zur Berechnung des unendlichen Produktes  $p(n)$  spalten wir dieses zunächst in drei getrennte Produkte  $p_1(n)$ ,  $p_2(n)$  und  $p_3(n)$  auf:

$$p(n) = \prod_{m \in \mathbb{N} \setminus \{n\}} \frac{m^2}{m^2 - n^2} = \underbrace{\prod_{m=1}^{n-1} \frac{m^2}{m^2 - n^2}}_{=: p_1(n)} \cdot \underbrace{\prod_{m=n+1}^{2n} \frac{m^2}{m^2 - n^2}}_{=: p_2(n)} \cdot \underbrace{\prod_{m>2n} \frac{m^2}{m^2 - n^2}}_{=: p_3(n)}$$

Umformung von  $p_1(n)$  ergibt:

$$\begin{aligned}
 p_1(n) &= \prod_{m=1}^{n-1} \frac{m^2}{m^2 - n^2} = \prod_{m=1}^{n-1} \left( (-1) \cdot \frac{m^2}{n^2 - m^2} \right) \\
 &= (-1)^{n-1} \cdot \prod_{m=1}^{n-1} \frac{m^2}{(n-m)(n+m)} = (-1)^{n-1} \cdot \prod_{m=1}^{n-1} \frac{1}{n-m} \cdot \prod_{m=1}^{n-1} \frac{m^2}{n+m} \\
 &= (-1)^{n-1} \cdot \prod_{m=1}^{n-1} \frac{1}{m} \cdot \prod_{m=1}^{n-1} \frac{m^2}{n+m} = (-1)^{n-1} \cdot \prod_{m=1}^{n-1} \frac{m}{m+n}
 \end{aligned}$$

Umformung von  $p_2(n)$  ergibt:

$$p_2(n) = \prod_{m=n+1}^{2n} \frac{m^2}{m^2 - n^2} = \prod_{m=n+1}^{2n} \frac{m^2}{(m-n)(m+n)} = \prod_{m=1}^n \frac{(m+n)^2}{m(m+2n)}$$

Und eine Umformung des unendlichen Produktes  $p_3(n)$  ergibt:

$$\begin{aligned}
 p_3(n) &= \prod_{m>2n} \frac{m^2}{m^2 - n^2} = \prod_{m>2n} \left( \frac{m}{m-n} \cdot \frac{m}{m+n} \right) \\
 &= \prod_{m>0} \left( \frac{m+2n}{m+n} \cdot \frac{m+2n}{m+3n} \right) = \prod_{m=1}^n \frac{m+2n}{m+n} \cdot \prod_{m>n} \frac{m+2n}{m+n} \cdot \prod_{m>0} \frac{m+2n}{m+3n} \\
 &= \prod_{m=1}^n \frac{m+2n}{m+n} \cdot \underbrace{\prod_{m>0} \frac{m+3n}{m+2n} \cdot \prod_{m>0} \frac{m+2n}{m+3n}}_{=1} = \prod_{m=1}^n \frac{m+2n}{m+n}
 \end{aligned}$$

Das Gesamtprodukt  $p(n)$  ist somit:

$$\begin{aligned}
 p(n) &= \overbrace{(-1)^{n-1} \cdot \prod_{m=1}^{n-1} \frac{m}{m+n}}^{p_1(n)} \cdot \overbrace{\prod_{m=1}^n \frac{(m+n)^2}{m(m+2n)}}^{p_2(n)} \cdot \overbrace{\prod_{m=1}^n \frac{m+2n}{m+n}}^{p_3(n)} \\
 &= (-1)^{n-1} \cdot \prod_{m=1}^{n-1} \left( \underbrace{\frac{m}{m+n} \cdot \frac{(m+n)^2}{m(m+2n)} \cdot \frac{m+2n}{m+n}}_{=1} \right) \cdot \underbrace{\frac{(n+n)^2}{n(n+2n)} \cdot \frac{n+2n}{n+n}}_{=2} \\
 &= \underline{\underline{(-1)^{n-1} \cdot 2}}
 \end{aligned}$$

Setzen wir dieses Ergebnis in die Gleichung für  $f(x)$  ein, so erhalten wir ( $x > 0$  vorausgesetzt):

$$\begin{aligned} f(x) &= \sum_{n=1}^{\infty} p(n) \cdot (n^2 \cdot e^{-n^2 \cdot x}) \\ &= \sum_{n=1}^{\infty} ((-1)^{n-1} \cdot 2) (n^2 \cdot e^{-n^2 \cdot x}) = 2 \cdot \sum_{n=1}^{\infty} (-1)^{n-1} \cdot n^2 \cdot e^{-n^2 \cdot x} \end{aligned}$$

Aus der Dichteverteilung  $f$  lässt sich die kumulative Verteilungsfunktion  $F$  durch Integration berechnen:

$$F(x) = \int_{t=0}^x f(t) dt = \begin{cases} 0 & \text{für } x \leq 0 \\ 1 - 2 \cdot \sum_{n=1}^{\infty} (-1)^{n-1} \cdot e^{-n^2 \cdot x} & \text{für } x > 0 \end{cases}$$

Mit der gegebenen Verteilungsfunktion  $F$  lässt sich nun leicht die Wahrscheinlichkeit  $g(x)$  berechnen mit der die Größe  $h$  für den Fall, dass die Zahlen der Folge  $a$  die Stichprobe einer gleichverteilten Zufallsgröße sind, einen gegebenen Wert  $x$  überschreitet:

$$g(x) = \mathbb{P}(h_{gleich} > x) = 1 - F(x) = \begin{cases} 1 & \text{für } x \leq 0 \\ 2 \cdot \sum_{n=1}^{\infty} (-1)^{n-1} \cdot e^{-n^2 \cdot x} & \text{für } x > 0 \end{cases}$$

Ist  $g(h) < 5\%$ , dann weicht die Stichprobe signifikant von einer Gleichverteilung ab, d.h. die Wahrscheinlichkeit beträgt nur 5%, dass im Falle des Zugrundeliegens einer Gleichverteilung  $h$  zufällig so groß bzw.  $g(h)$  zufällig so klein ist.

Das beschriebene Verfahren kann nicht nur zur Prüfung einer Stichprobe auf Gleichverteilung, sondern ebenso zur Prüfung einer Stichprobe  $a'$  auf jede beliebige Verteilung (mit gegebenen Parametern) angewendet werden. Hierzu ist die Stichprobe vorher mit Hilfe der entsprechenden kumulativen Verteilungsfunktion  $\Psi$  auf die geprüft werden soll zu transformieren:

$$a_m = \Psi(a'_m)$$

Möchte man z.B. auf eine Normalverteilung mit dem Mittelwert  $\mu$  und der Standardabweichung  $\sigma$  prüfen, dann ist:

$$a_m = \Psi(a'_m) = \Phi_{\mu, \sigma}(a'_m) = \frac{1}{2} \cdot \left( 1 + \operatorname{erf} \left( \frac{a'_m - \mu}{\sigma \cdot \sqrt{2}} \right) \right)$$

$h$  bzw.  $g(h)$  eignen sich, um festzustellen wie signifikant eine Stichprobe von einer Gleichverteilung abweicht. Da, falls die Stichprobe keiner Gleichverteilung folgt, die Signifikanz bei Erhöhung des Stichprobenumfangs beliebig steigt, definieren wir noch ein Maß  $k \in [0, 1]$  für den Grad der Abweichung von einer Gleichverteilung, welches sich bei Erhöhung des Stichprobenumfangs einem Grenzwert annähert:

$$k := \sqrt{\frac{6}{\pi^2 M^2} \cdot \sum_{n=1}^{\infty} \left[ \frac{1}{n^2} \cdot \left( \left( \sum_{m=1}^M \cos(2\pi \cdot n \cdot a_m) \right)^2 + \left( \sum_{m=1}^M \sin(2\pi \cdot n \cdot a_m) \right)^2 \right) \right]}$$

Insbesondere wirkt sich ein Duplizieren der Stichprobe nicht auf die neu definierte Größe  $k$  aus, d.h. für  $a_{m+N} = a_m$  und  $N \in \mathbb{N}$  gilt  $k|_{M=N \cdot M_0} = k|_{M=M_0}$ . Ich vermute, dass falls eine Stichprobe mit  $M$  optimal gleichverteilten Werten vorliegt ( $a_m = \frac{m}{M} + d$  mit  $d \in [0, \frac{m}{M}]$ ), die Größe  $k = \frac{1}{M}$  ist. Diese Vermutung begründe ich wie folgt:

$$\begin{aligned} k|_{a_m = \frac{m}{M} + d} &= \sqrt{\frac{6}{\pi^2 M^2} \cdot \sum_{n=1}^{\infty} \left( \frac{1}{n^2} \cdot k_m \right)} \\ k_m &:= \left( \sum_{m=1}^M \cos \left( 2\pi n \left( \frac{m}{M} + d \right) \right) \right)^2 + \left( \sum_{m=1}^M \sin \left( 2\pi n \left( \frac{m}{M} + d \right) \right) \right)^2 \\ &= \sum_{m=1}^M \sum_{m'=1}^M \cos \left( 2\pi n \left( \frac{m}{M} + d \right) \right) \cdot \cos \left( 2\pi n \left( \frac{m'}{M} + d \right) \right) + \sum_{m=1}^M \sum_{m'=1}^M \sin \left( \dots \right) \cdot \sin \left( \dots \right) \\ &= \sum_{m=1}^M \sum_{m'=1}^M \left[ \cos \left( 2\pi n \left( \frac{m}{M} + d \right) \right) \cdot \cos \left( 2\pi n \left( \frac{m'}{M} + d \right) \right) + \sin \left( \dots \right) \cdot \sin \left( \dots \right) \right] \\ &= \sum_{m=1}^M \sum_{m'=1}^M \cos \left( 2\pi n \left( \frac{m}{M} + d \right) - 2\pi n \left( \frac{m'}{M} + d \right) \right) \\ &= \sum_{m=1}^M \sum_{m'=1}^M \cos \left( 2\pi n \cdot \frac{m - m'}{M} \right) \\ &= \sum_{m=1}^M \cos \left( 2\pi n \frac{0}{M} \right) + \sum_{m=1}^{M-1} (M-m) \cos \left( 2\pi n \frac{m}{M} \right) + \sum_{m=1}^{M-1} (M-m) \cos \left( 2\pi n \frac{-m}{M} \right) \\ &= M + 2 \sum_{m=1}^{M-1} (M-m) \cos \left( 2\pi n \frac{m}{M} \right) \stackrel{?}{=} \begin{cases} M^2, & \text{falls } n \text{ Vielfaches von } M \\ \text{ansonsten } 0 \end{cases} \end{aligned}$$

Für die letzte Gleichung habe ich noch keinen Beweis gefunden, jedoch deuten numerische Berechnungen für alle  $n$  und  $M$  bis 100 darauf hin, dass sie erfüllt ist. Daraus würde dann folgen:

$$\begin{aligned} k|_{a_m=\frac{m}{M}+d} &= \sqrt{\frac{6}{\pi^2 M^2} \cdot \sum_{i=1}^{\infty} \left( \frac{1}{(i \cdot M)^2} \cdot M^2 \right)} \\ &= \sqrt{\frac{6}{\pi^2 M^2} \cdot \sum_{i=1}^{\infty} \frac{1}{i^2}} = \sqrt{\frac{6}{\pi^2 M^2} \cdot \frac{\pi^2}{6}} = \frac{1}{M} \end{aligned}$$

Es wird unentgeltlich jeder Person das Recht eingeräumt, Kopien dieses Dokumentes und des dazugehörigen Haskell-Programmes zu benutzen, zu kopieren, zu modifizieren, mit anderen Werken zu kombinieren, zu veröffentlichen, zu verbreiten, unterzulizensieren, zu verkaufen und/oder weiteren Personen diese Rechte einzuräumen, solange der Autor und dieser Hinweis in allen Kopien oder wesentlichen Teilen davon erwähnt werden. Fehlerfreiheit und Funktionsfähigkeit werden nicht zugesichert; Benutzung erfolgt auf eigenes Risiko.

Folgendes Haskell-Programm kann die oben definierten Größen berechnen:

```
inhomoSum :: [Double] -> Double
inhomoSum xs
  = sum [ ((s1 n)**2 + (s2 n)**2) / n**2 | n <- [1..huge] ]
  where
    s1 n    = sum [ cos (2 * pi * n * x) | x <- xs ]
    s2 n    = sum [ sin (2 * pi * n * x) | x <- xs ]
    huge    = 1000

inhomoSignificance :: [Double] -> Double
inhomoSignificance xs = 1 / prob (inhomoSum xs / count)
  where
    count    = fromIntegral (length xs)
    prob x   = (min 1 . max 0) (
      sum [ (if n < huge then 2 else 1) *
        (-1)**(n-1) * exp (-n**2 * x)
        | n <- [1..huge] ]
    )
    huge     = 1000

inhomogeneity :: [Double] -> Double
inhomogeneity xs = sqrt (6 / pi**2 / count**2 * inhomoSum xs)
  where
    count    = fromIntegral (length xs)
    huge     = 1000
```

Hinweise: `fromIntegral` dient nur der Typ-Konvertierung von ganzzahligen Zahlen in Gleitkommazahlen und hat nichts mit der Berechnung von mathematischen Integralen zu tun. Die Fallunterscheidung `if n < huge then 2 else 1` dient der Vermeidung von Berechnungsfehlern für  $x$  gleich oder nahe 0, indem das letzte Glied der (eigentlich unendlichen) Summe nur halb angerechnet wird.

`inhomoSum` ist eine Hilfsfunktion. Die Funktion `inhomoSignificance` berechnet für eine gegebene Zahlenfolge den Kehrwert der Größe  $g(h)$ , d.h. wenn das Ergebnis größer als 20 ist, liegt eine signifikante Abweichung von der Gleichverteilung vor. Die Funktion `inhomogeneity` berechnet die Größe  $k$ , wobei 1 für maximale Ungleichverteilung (nur ein Wert kommt in der Zahlenfolge vor) steht und ein Wert nahe 0 gute Gleichverteilung angibt.