Verwendung von Fourier-Koeffizienten zum Testen auf Gleichverteilung

Jan Behrens

1. Juni 2008
Ergänzung/Überarbeitung vom 9. August 2009

Ziel soll es sein, zu einer Stichprobe von Zufallszahlen im Bereich von 0 bis 1 den eindeutig definierten Wert $g(h)$ zu berechnen, der als Maß für das Abweichen dieser Stichprobe von einer Gleichverteilung dient. $g(h)$ nimmt ebenfalls Werte zwischen 0 und 1 an, wobei kleine Werte eine hohe Abweichung von der Gleichverteilung bedeuten, während Werte gegen 1 für eine gute Übereinstimmung der Stichprobe mit einer Gleichverteilung stehen. Die Größe $g(h)$ soll außerdem mit dem Signifikanzniveau übereinstimmen; 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 Lindeberg-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.

Es wird weiterhin 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.

Da die Signifikanz bei Erhöhung des Stichproblenumfanges von einer nicht exakt gleichverteilten Zufallsgröße beliebig steigt, und $g(h)$ entsprechend gegen 0 strebt, wird zum Abschluss noch eine weitere Größe k untersucht, die ebenfalls ein Maß für die Ungleichverteilung ist, die sich jedoch bei einer Erhöhung des Stichproblemumfanges einem Grenzwert annähert, der die Abweichung zwischen der zugrundeliegenden Verteilung und einer Gleichverteilung wiederspiegelt.

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 μ und folgende Varianz σ^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 $M \to \infty$ gemäß dem Grenzwertsatz von Lindeberg-Levy 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

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

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

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 darstellen:

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

Die Zufallsgröße $X_n^2 + Y_n^2$ genügt einer χ^2-Verteilung mit 2 Freiheitsgraden, welche die Wahrscheinlichkeitsdichte $\frac{1}{2} \cdot \mathrm{e}^{-\frac{1}{2} \cdot x}$ 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_{gleich} den folgenden Wahrscheinlichkeitsdichtefunktionen:

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

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

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

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 ($\alpha \neq \beta$):

$$(A \cdot \mathrm{e}^{-\alpha x}) * (B \cdot \mathrm{e}^{-\beta x}) = \int_{\Theta=0}^{x} A \cdot \mathrm{e}^{-\alpha(x-\Theta)} \cdot B \cdot \mathrm{e}^{-\beta\Theta}\, d\Theta$$

$$= AB \cdot \int_{\Theta=0}^{x} \mathrm{e}^{-\alpha x + \alpha\Theta - \beta\Theta}\, d\Theta = AB \cdot \mathrm{e}^{-\alpha x} \cdot \int_{\Theta=0}^{x} \mathrm{e}^{(\alpha-\beta)\cdot\Theta}\, d\Theta$$

$$= AB \cdot \mathrm{e}^{-\alpha x} \cdot \frac{1}{\alpha-\beta} \cdot \left(\mathrm{e}^{(\alpha-\beta)\cdot x} - \mathrm{e}^{(\alpha-\beta)\cdot 0}\right) = \frac{AB}{\alpha-\beta} \cdot \left(\mathrm{e}^{-\beta x} - \mathrm{e}^{-\alpha x}\right)$$

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 $A \cdot \mathrm{e}^{-\alpha x}$, so ist der Koeffizient B eines Summanden $B \cdot \mathrm{e}^{-\beta x}$ im Ergebnis mit $\frac{A}{\alpha-\beta}$ zu multiplizieren. Aus diesem Grunde lässt sich $f(x)$ für $x > 0$ darstellen als:

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

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:

$$p_1(n) = \prod_{m=1}^{n-1} \frac{m^2}{m^2-n^2} = \prod_{m=1}^{n-1} \left((-1) \cdot \frac{m^2}{n^2-m^2} \right)$$

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

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

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:

$$p_3(n) = \prod_{m>2n} \frac{m^2}{m^2-n^2} = \prod_{m>2n} \left(\frac{m}{m-n} \cdot \frac{m}{m+n} \right)$$

$$= \prod_{m>0} \left(\frac{m+2n}{m+n} \cdot \frac{m+2n}{m+3n} \right) = \prod_{m=1}^{n} \frac{m+2n}{m+n} \cdot \prod_{m>n} \frac{m+2n}{m+n} \cdot \prod_{m>0} \frac{m+2n}{m+3n}$$

$$= \prod_{m=1}^{n} \frac{m+2n}{m+n} \cdot \underbrace{\prod_{m>0} \frac{m+3n}{m+2n} \cdot \prod_{m>0} \frac{m+2n}{m+3n}}_{=1} = \prod_{m=1}^{n} \frac{m+2n}{m+n}$$

Das Gesamtprodukt $p(n)$ ist somit:

$$p(n) = \overbrace{(-1)^{n-1} \cdot \prod_{m=1}^{n-1} \frac{m}{m+n}}^{p_1(n)} \cdot \overbrace{\prod_{m=1}^{n} \frac{(m+n)^2}{m(m+2n)}}^{p_2(n)} \cdot \overbrace{\prod_{m=1}^{n} \frac{m+2n}{m+n}}^{p_3(n)}$$

$$= (-1)^{n-1} \cdot \prod_{m=1}^{n-1} \left(\underbrace{\frac{m}{m+n} \cdot \frac{(m+n)^2}{m(m+2n)} \cdot \frac{m+2n}{m+n}}_{=1} \right) \cdot \underbrace{\frac{(n+n)^2}{n(n+2n)} \cdot \frac{n+2n}{n+n}}_{=2}$$

$$\underline{\underline{= (-1)^{n-1} \cdot 2}}$$

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

$$f(x) = \sum_{n=1}^{\infty} p(n) \cdot \left(n^2 \cdot \mathrm{e}^{-n^2 \cdot x} \right)$$

$$= \sum_{n=1}^{\infty} \left((-1)^{n-1} \cdot 2 \right) \left(n^2 \cdot \mathrm{e}^{-n^2 \cdot x} \right) = 2 \cdot \sum_{n=1}^{\infty} (-1)^{n-1} \cdot n^2 \cdot \mathrm{e}^{-n^2 \cdot x}$$

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

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

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:

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

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 Ψ auf die geprüft werden soll zu transformieren:

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

Möchte man z.B. auf eine Normalverteilung mit dem Mittelwert μ und der Standardabweichung σ 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 Stichprobenumfanges 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 $a_{m+M_0} = a_m$ und $N \in \mathbb{N}$ gilt $k|_{M=N \cdot M_0} = k|_{M=M_0}$. Falls eine Stichprobe mit M optimal gleichverteilten Werten vorliegt ($a_m = \frac{m}{M} + d$ mit $d \in [0, \frac{m}{M}]$), ist die Größe $k = \frac{1}{M}$. Dies lässt sich wie folgt beweisen:

$$k|_{a_m = \frac{m}{M} + d} = \sqrt{\frac{6}{\pi^2 M^2} \cdot \sum_{n=1}^{\infty} \left(\frac{1}{n^2} \cdot k_m\right)}$$

$$k_m := \left(\sum_{m=1}^{M} \cos\left(2\pi n(\frac{m}{M} + d)\right)\right)^2 + \left(\sum_{m=1}^{M} \sin\left(2\pi n(\frac{m}{M} + d)\right)\right)^2$$

Man formt den Term k_m mittels Ausmultiplizieren und Additionstheoremen um:

$$k_m = \sum_{m=1}^{M}\sum_{m'=1}^{M} \cos\left(2\pi n(\frac{m}{M}+d)\right)\cdot\cos\left(2\pi n(\frac{m'}{M}+d)\right)+\sum_{m=1}^{M}\sum_{m'=1}^{M}\sin(...)\cdot\sin(...)$$

$$= \sum_{m=1}^{M}\sum_{m'=1}^{M}\left[\cos\left(2\pi n(\frac{m}{M}+d)\right)\cdot\cos\left(2\pi n(\frac{m'}{M}+d)\right)+\sin(...)\cdot\sin(...)\right]$$

$$= \sum_{m=1}^{M}\sum_{m'=1}^{M}\cos\left(2\pi n(\frac{m}{M}+d)-2\pi n(\frac{m'}{M}+d)\right)$$

Dann lässt sich die Variable d entfernen:

$$k_m = \sum_{m=1}^{M}\sum_{m'=1}^{M}\cos\left(2\pi n\frac{m}{M}-2\pi n\frac{m'}{M}\right)$$

Anschließend wird der Term zurück transformiert:

$$k_m = \sum_{m=1}^{M}\sum_{m'=1}^{M}\left[\cos\left(2\pi n\frac{m}{M}\right)\cdot\cos\left(2\pi n\frac{m'}{M}\right)+\sin(...)\cdot\sin(...)\right]$$

$$= \sum_{m=1}^{M}\sum_{m'=1}^{M}\cos\left(2\pi n\frac{m}{M}\right)\cdot\cos\left(2\pi n\frac{m'}{M}\right)+\sum_{m=1}^{M}\sum_{m'=1}^{M}\sin(...)\cdot\sin(...)$$

$$= \left(\sum_{m=1}^{M}\cos\left(2\pi n\frac{m}{M}\right)\right)^2+\left(\sum_{m=1}^{M}\sin\left(2\pi n\frac{m}{M}\right)\right)^2$$

Falls n ein Vielfaches von M ist, ist $k_m = M^2$, ansonsten sind beide Summen 0 und damit auch $k_m = 0$. Beschreiben wir die Vielfachen von M mit $i \cdot M$, dann ergibt sich für $k|_{a_m=\frac{m}{M}+d}$:

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

$$= \sqrt{\frac{6}{\pi^2M^2}\cdot\sum_{i=1}^{\infty}\frac{1}{i^2}} = \sqrt{\frac{6}{\pi^2M^2}\cdot\frac{\pi^2}{6}} = \frac{1}{M}$$

Dieses entspricht genau der oben aufgestellten Behauptung.

Folgendes Haskell-Programm kann die oben definierten Größen berechnen:

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

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

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

Hinweise zum Programm: `fromIntegral` dient nur der Typ-Konvertierung von ganzzahligen Zahlen in Gleitkommazahlen und hat nichts mit der Berechnung von mathematischen Integralen zu tun. Die Variablen `i` und `n` stellen beide die gleiche Zahl dar, jedoch unterschiedlichen Typs: `i` ist vom Typ `Integer`, während `n` einen Typ der Klasse `RealFloat` aufweist. Die Fallunterscheidung `if i < 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 die Größe $g(h)$ für eine gegebene Zahlenfolge, d.h. wenn das Ergebnis kleiner als 0.05 ist, dann liegt eine signifikante Abweichung von der Gleichverteilung vor. Die Funktion `inhomogeneity` berechnet die Größe k, wobei 1 für maximale Ungleichverteilung steht (nur ein Wert kommt in der Zahlenfolge vor) und ein Wert nahe 0 gute Gleichverteilung angibt.