Günter HANISCH

# Die $\chi^2$-Verteilung

**Zusammenfassung**

Ausgehend von einem Einführungsbeispiel wird das Testen mittels des $\chi^2$-Unabhängigkeitstests erläutert und die Dichte der $\chi^2$-Verteilung durch vollständige Induktion hergeleitet. Anschließend werden der $\chi^2$-Anpassungstest und der $\chi^2$-Unabhängigkeitstest besprochen. Den Abschluß bildet ein Computerprogramm zur Simulation der $\chi^2$-Verteilung, mit dessen Hilfe die $\chi^2$-Tabelle hergeleitet werden kann.

## 1 Ein Anwendungsbeispiel

Das folgende Beispiel behandelt die möglichen Auswirkungen von Dioxin. Die Daten stammen vom Institut zum Schutz für Mutter und Kind aus Hanoi in Vietnam. Bekanntlich gingen in Südvietnam Hunderte von Kilogramm Dioxin auf das Land nieder, die in den chemischen Kampfstoffen Agent Orange, Agent White, Agent Pink und Agent Purple enthalten waren. Untersucht wurde, ob die Herbizid-Exposition des Vaters – und nicht die der Mutter – für Mißbildungen bei Föten verantwortlich gemacht werden kann. Dazu wurden insgesamt 40 064 Frauen ausgesucht, die in drei nordvietnamesischen Distrikten zu Hause sind, in denen sie nicht mit Dioxin in Berührung gekommen waren. 29 041 dieser Frauen sind mit Männern verheiratet, die im Norden gekämpft hatten und daher ebenfalls nicht mit Dioxin in Berührung gekommen waren, wogegen die verbleibenden 11 023 Frauen mit solchen Männer verheiratet sind, die in Südvietnam eingesetzt waren und daher dem Einfluß von Dioxin ausgesetzt waren. Die Ergebnisse sind in Tab. 1 zusammengestellt.

Um zu untersuchen, ob es sich für die Schwangerschaft verhängnisvoll auswirkt, wenn der Vater unter Dioxinbelastung stand, ob also unterschiedliche Häufigkeiten bei den Ergebnissen der Schwangerschaften auftreten, ist es sinnvoll, sich vorerst auszurechnen, welche Häufigkeiten zu erwarten wären, wenn man annimmt, daß die Dioxinbelastung des Vaters keinen Einfluß auf die Schwangerschaft hat. Dazu werden die absoluten Häufigkeiten der jeweiligen Ergebnisse der Schwangerschaften berechnet und diese dann im Verhältnis 121 993 : 32 069 aufgeteilt. Diese so unter der Annahme „Kein Einfluß“ (gerundeten) erwarteten Häufigkeiten sind in Tab. 2 kursiv eingetragen.

Man sieht, daß die erwarteten Häufigkeiten $h_i^{(e)}$ von den beobachteten $h_i^{(b)}$ abweichen. Um festzustellen, ob dies überzufällig ist, brauchen wir eine Kenngröße; als

| Schwangerschaften in Nordvietnam | | |
|---|---|---|
| Ergebnis der Schwangerschaft | kein Elternteil Dioxin belastet | Vater Dioxin belastet |
| Normale Geburt | 110 992 | 28 795 |
| Angeborene Schädigung | 521 | 189 |
| Totgeburt | 2 512 | 576 |
| Spontaner Abortus | 7 148 | 2 271 |
| Kurretage | 750 | 210 |
| Eileiterschwangerschaft | 70 | 28 |
| Summe | 121 993 | 32 069 |

Tabelle 1: Ergebnisse von Schwangerschaften nordvietnamesischer Frauen, deren Männer teilweise Dioxin ausgesetzt waren

| Vergleich erwarteter mit beobachteten Häufigkeiten | | | | | |
|---|---|---|---|---|---|
| Ergebnis der Schwangerschaft | kein Elternteil Dioxin belastet | | Vater Dioxin belastet | | Summe |
| Normale Geburt | 110 992 | *110 689* | 28 795 | *29 098* | 139 787 |
| Angeborene Schädigung | 521 | *562* | 189 | *148* | 710 |
| Totgeburt | 2 512 | *2 445* | 576 | *643* | 3 088 |
| Spontaner Abortus | 7 148 | *7 458* | 2 271 | *1 961* | 9 419 |
| Kurretage | 750 | *760* | 210 | *200* | 960 |
| Eileiterschwangerschaft | 70 | *78* | 28 | *20* | 98 |
| Summe | 121 993 | | 32 069 | | 154 062 |

Tabelle 2: Ergebnisse von Schwangerschaften nordvietnamesischer Frauen: Vergleich erwarteter (kursiv) mit beobachteten Häufigkeiten

solche würde sich die Summe der Differenzen anbieten; diese ist jedoch 0, da sich die Summanden gegenseitig aufheben. Daher werden die Differenzen – wie beim Berechnen der Varianz – quadriert. Um allerdings die sich aus verschiedenen Aufgabenstellungen ergebenden Summen miteinander vergleichen zu können, werden die einzelnen Summanden dadurch normiert, daß sie durch die erwarteten Häufigkeiten dividiert werden. Wir bilden also für jede Zelle die Prüfgröße $\frac{(h_i^{(b)} - h_i^{(e)})^2}{h_i^{(e)}}$ und summieren anschließend auf. In Tab. 3 sind die so erhaltenen Kenngrößen

eingetragen.

| Prüfgrößen | | | |
|---|---|---|---|
| Ergebnis der Schwangerschaft | kein Elternteil Dioxin belastet | Vater Dioxin belastet | Summe |
| Normale Geburt | 0,827 | 3,146 | 3,973 |
| Angeborene Schädigung | 3,021 | 11,490 | 14,511 |
| Totgeburt | 1,824 | 6,939 | 8,764 |
| Spontaner Abortus | 12,916 | 49,133 | 62,049 |
| Kurretage | 0,136 | 0,517 | 0,653 |
| Eileiterschwangerschaft | 0,744 | 2,832 | 3,576 |
| Summe | 19,468 | 74,059 | 93,527 |

Tabelle 3: Ergebnisse von Schwangerschaften nordvietnamesischer Frauen: Prüfgrößen

Je größer die Summe der Prüfgrößen ist, desto unwahrscheinlicher ist es, daß die Abweichungen zufällig sind. Wir werden im folgenden zeigen, daß sich aus der Prüfgrößensumme die Wahrscheinlichkeit ergibt, daß das beobachtete Ereignis zufällig ist, wobei dieser Wert nur von der Zeilenanzahl $n_z$ und der Spaltenanzahl $n_s$ der Tabelle abhängt.

Grund für diese Behauptung ist, daß die Summe der Prüfgrößen asymptotisch einer bestimmten Verteilung, nämlich der $\chi^2$-Verteilung mit $(n_z - 1)(n_s - 1)$-Freiheitsgraden folgt. Mit ihr werden wir nachweisen, daß die Annahme, die Abweichungen seien nur zufälliger Natur, höchst unwahrscheinlich ist.

# 2 Dichte der $\chi^2$-Verteilung

Die $\chi^2$-Verteilung kann folgendermaßen definiert werden:

**Def. 2.1:** Die Verteilung einer Summe ${X_1}^2 + \ldots + {X_n}^2$, wobei $X_1, \ldots, X_n$ unabhängige standard-normalverteilte Zufallsvariablen sind, heißt *$\chi^2$-Verteilung mit $n$ Freiheitsgraden* oder kurz *$\chi^2(n)$-Verteilung*, wobei $n \in N$ ist. ∎

**Bemerkung:** Mit dieser Definition folgen wir auch dem historischem Gang, den im Zusammenhang mit der Fehlertheorie von GAUß untersuchte der Astronom Friedrich Robert HELMERT (1843–1917) Quadratsummen von normalverteilten Größen.

Die dabei nachgewiesene Verteilungsfunktion nannte Karl PEARSON (1857–1936) später $\chi^2$-Verteilung. Sie ist von einem Formparameter $n$, dem sogenannten *„Freiheitsgrad"* abhängig; d.h.: Die $\chi^2$-Verteilung wird durch eine einparametrige Kurvenschar repräsentiert (vgl.: die Normalverteilung $N(\mu;\, \sigma^2)$ ist zweiparametrig).

Satz 2.1: Die Dichte der $\chi^2(n)-$Verteilung ist gegeben durch

$$g_n(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\,\left(\frac{n}{2}-1\right)!}\, y^{\frac{n}{2}-1}\, e^{-\frac{y}{2}} & \text{für } y > 0 \\ 0 & \text{für } y \leq 0. \end{cases}$$

Dabei ist $n! = n \cdot (n-1)!$ und $(-\frac{1}{2})! = \sqrt{\pi}$. ∎

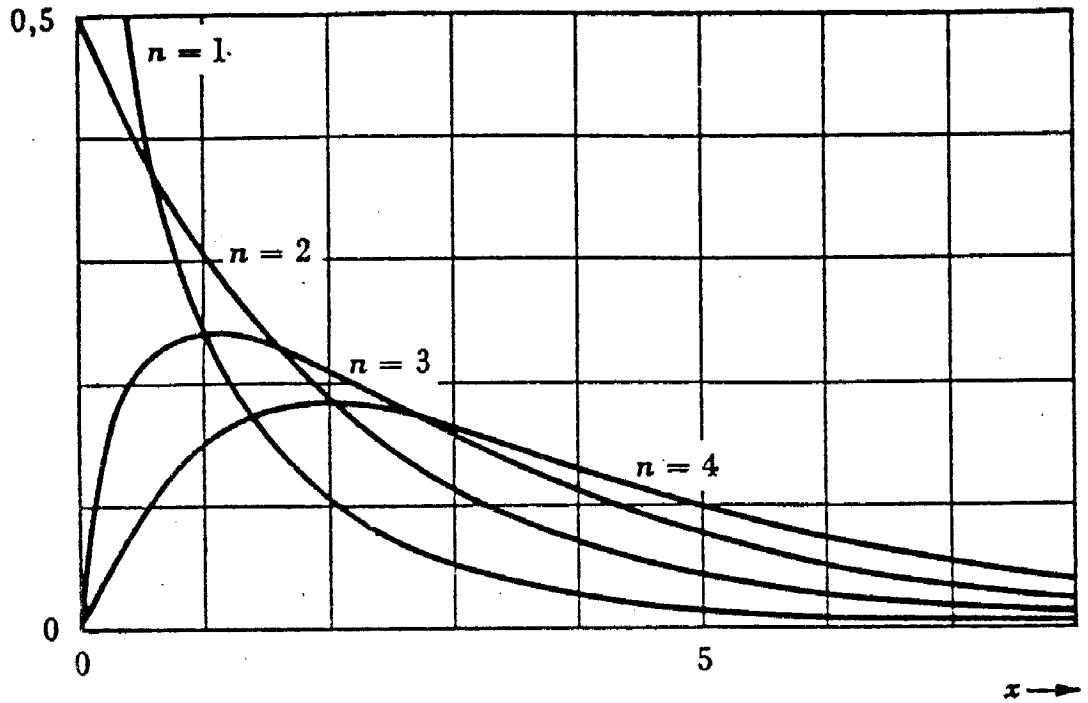


Abbildung 1: Dichte der $\chi^2$-Verteilung für verschiedene Freiheitsgrade

Bemerkung: Üblicherweise wird $n!$, wenn $n$ keine natürliche Zahl ist, durch die Gammafunktion

$$\Gamma(x) = \int_0^\infty e^{-t} \cdot t^{x-1}\, dx \qquad \text{mit} \qquad \Gamma(x+1) = x!$$

definiert. Da wir aber $n!$ nur für ganzzahlige Vielfache von $\frac{1}{2}$ brauchen, ist dies hier nicht notwendig.

**Beweis:** Der Beweis erfolgt durch vollständige Induktion:

*1. Induktionsanfang $n = 1$:*
Ist $X$ $N(0,1)$-verteilt, so hat $X$ die Dichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Es sei nun $Y = X^2$. Für die Verteilungsfunktion $G_1(y)$ muß daher gelten, da sie wegen $Y = X^2$ keine negativen Werte annehmen kann:

$$G_1(y) = \begin{cases} P(Y < y) & \text{für } y > 0 \\ 0 & \text{für } y \leq 0. \end{cases}$$

Da $G_1$ gleichmäßig stetig ist, können wir umformen:

$$P(Y < y) = P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y})$$

Daher ist

$$g_1(y) = \frac{dG_1(y)}{dy} = 2 \cdot \frac{\varphi(\sqrt{y})}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{y}} \cdot e^{-\frac{y}{2}} = \frac{1}{\sqrt{2\pi}} \cdot y^{-\frac{1}{2}} \cdot e^{-\frac{y}{2}} \text{ für } y > 0.$$

Somit ist

$$g_1(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} \cdot y^{-\frac{1}{2}} \cdot e^{-\frac{y}{2}} & \text{für } y > 0 \\ 0 & \text{für } y \leq 0. \end{cases}$$

und das ist die $\chi^2$-Verteilung mit einem Freiheitsgrad.

*2. Schluß von $n = 1$ auf $n = 2$*
Normalerweise würde jetzt der Schluß von $n$ auf $n + 1$ erfolgen, und so findet man es auch in den Lehrbüchern, die die $\chi^2$-Verteilung über vollständige Induktion herleiten. Allerdings ist der Schluß von $n$ auf $n + 1$ wesentlich aufwendiger als der von $n$ auf $n + 2$. Daher gehen wir diesen Weg.

Während wir oben zu untersuchen hatten, wie das Quadrat einer Zufallsvariablen verteilt ist, brauchen wir jetzt die Summe:

**Hilfssatz** *(Faltung):*

Seien $X_1$, $X_2$ unabhängige Zufallsvariablen mit den Dichten $f_1$ und $f_2$. Dann gilt für die Dichte von $X_1 + X_2$:

$$g(y) = \int_{-\infty}^{\infty} f_1(x_1)\, f_2(y - x_1)\, dx_1 \,.$$

**Beweis:** Hier bringen wir zwei Beweise und zwar vorerst einen anschaulichen Beweis, der Doppelintegrale vermeidet:

Es sei $Y = X_1 + X_2$. Im diskreten Fall erhalten wir

$$g(y) = P(Y = y)$$

dadurch, daß wir alle Wahrscheinlichkeiten berechnen, bei denen die Zufallsvariable $X_1$ den Wert $x_1$ annimmt und die Zufallsvariable $X_2$ den Wert $x_2$, wobei $x_1 + x_2 = y$ ist, und anschließend aufsummieren. Wegen der Unabhängigkeit erhalten wir deshalb

$$g(y) = P(Y = y) = \sum_{\forall x_1} P(X_1 = x_1) \cdot P(X_2 = y - x_1).$$

(Man denke sich etwa $X_1$ als das Ergebnis eines Münzwurfes mit den Ausgängen 1 und 2 und $X_2$ als das Ergebnis eines Würfelwurfes. Angenommen $y$ sei 4. Das kann durch $X_1 = 1$ und $X_2 = 3$ mit der Wahrscheinlichkeit $\frac{1}{2} \cdot \frac{1}{6}$ oder durch $X_1 = 2$ und $X_2 = 2$ mit derselben Wahrscheinlichkeit bewirkt werden.)

Im stetigen Fall geht die Summe in ein Integral über und wir erhalten die oben angegebene Formel.

Auf einer höheren Exaktheitsstufe erfolgt der Beweis des Satzes mit Hilfe der Verteilungsfunktionen:

$$F_{X_1+X_2}(t) = P(X_1 + X_2 \leq t) = P_{(X_1,X_2)}\{(x_1, x_2) : \; x_1 + x_2 \leq t\} \overset{Unabh.}{=}$$

$$= \iint_{x_1+x_2 \leq t \to x_2 \leq t-x_1} f_1(x_1)\, f_2(x_2)\, dx_1\, dx_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x_1} f_1(x_1)\, f_2(x_2)\, dx_2\, dx_1 =$$

Substitution: $x_2 = y - x_1$ mit $y \leq t$,

$$x_2 = y - x_1 \qquad \frac{dx_2}{dx_1} = 1 \qquad dx_2 = dy\ .$$

$$= \int_{-\infty}^{\infty} f_1(x_1) \int_{-\infty}^{t} f_2(y - x_1)\, dy\, dx_1 = \int_{-\infty}^{t} \underbrace{\left( \int_{-\infty}^{\infty} f_1(x_1)\, f_2(y - x_1)\, dx_1 \right)}_{f_1 * f_2(y)} dy\ .$$

$f_1 * f_2$ heißt *Faltung*, aus ihr ergibt sich die gesuchte Dichte. ∎

Nun können wir die Dichte der $\chi^2_2$ ermitteln, wobei sich die Grenzen des Integrals aus der Tatsache ergeben, daß $g_1(x_1) = 0$ für $x_1 < 0$ und $g_1(y - x_1) = 0$ für $y - x_1 < 0$, somit $g_1(x_1) \cdot g_1(y - x_1) \neq 0$ für $0 \leq x_1 \leq y$ ist:

$$\int_0^y \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x_1}{2}} \cdot \frac{1}{\sqrt{2\pi}} (y - x_1)^{-\frac{1}{2}} e^{-\frac{y-x_1}{2}}\, dx_1 =$$

$$= \frac{1}{2\pi} e^{-\frac{y}{2}} \int_0^y \frac{1}{\sqrt{x_1(y-x_1)}}\, dx_1$$

Um das Integral zu lösen, substituieren wir $u = \sqrt{1 - \frac{x_1}{y}}$, somit $x_1 = y(1-u^2)$ und $dx_1 = -2yu\, du$ und erhalten

$$\int_0^y \frac{1}{\sqrt{x_1(y-x_1)}}\, dx_1 = \int_1^0 \frac{-2yu\, du}{\sqrt{z(1-u^2)\cdot u^2 \cdot z}} = \int_0^1 \frac{-2\, du}{\sqrt{1-u^2}} = 2 \arcsin u \Big|_0^1 = 2 \cdot \frac{\pi}{2}$$

Somit erhalten wir als Dichte für die $\chi_2^2$:

$$g_2(y) = \begin{cases} \frac{1}{2} e^{-\frac{y}{2}} & \text{für } y > 0 \\ 0 & \text{für } y \leq 0. \end{cases}$$

**Bemerkung:** Obige Verteilungsfunktion läßt sich elementar auswerten, da

$$G_2(y) = \frac{1}{2} \int_0^y e^{-\frac{u}{2}} = -e^{-\frac{u}{2}} \Big|_0^y = 1 - e^{-\frac{y}{2}} \text{ für } y > 0$$

*3. Schluß von $n$ auf $n+2$:*

Wir berechnen

$$\int_0^y g_n(x_1) \cdot g_2(y-x_1)\, dx_1 = \int_0^y \frac{1}{2^{\frac{n}{2}} \left(\frac{n}{2}-1\right)!}\, x_1^{\frac{n}{2}-1}\, e^{-\frac{x_1}{2}} \cdot \frac{1}{2} e^{-\frac{y-x_1}{2}}\, dx_1 =$$

$$= \frac{1}{2^{\frac{n}{2}+1} \left(\frac{n}{2}-1\right)!} e^{-\frac{y}{2}} \int_0^y x_1^{\frac{n}{2}-1}\, dx_1 = \frac{1}{2^{\frac{n}{2}+1} \left(\frac{n}{2}-1\right)!} e^{-\frac{y}{2}} \cdot x^{\frac{n}{2}} \cdot \frac{2}{n} \Big|_0^y =$$

$$= \frac{1}{2^{\frac{n}{2}+1} \left(\frac{n}{2}-1\right)! \cdot \frac{n}{2}} e^{-\frac{y}{2}} z^{\frac{n}{2}}$$

und das ist wegen

$$\left(\frac{n}{2}-1\right)! \cdot \frac{n}{2} = \left(\frac{n}{2}\right)!$$

die Dichte der $\chi_{n+2}^2$-Verteilung:

$$g_{n+2}(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}+1} \left(\frac{n}{2}\right)!}\, y^{\frac{n}{2}}\, e^{-\frac{y}{2}} & \text{für } y > 0 \\ 0 & \text{für } y \leq 0. \end{cases}$$

Und da der Satz für $n = 1$ und für $n = 2$ gilt und – wie jetzt gezeigt wurde – der Schluß von $n$ auf $n+2$ zulässig ist, gilt er für alle $n \in N$.

Ein anderer – meines Erachtens wesentlich eleganterer Beweis – betrachtet die $\chi^2$-Verteilung als Spezialfall der $\Gamma$-Verteilung, zeigt mit Hilfe der charakteristischen Funktion, daß für die $\Gamma$-Verteilung ein Additionstheorem existiert und wendet dieses – nachdem die $\chi^2_1$-Verteilung hergeleitet ist – auf die $\chi^2$-Verteilung an (siehe etwa FISZ, 1971). ∎

# 3 Der $\chi^2$-Anpassungstest

Die Bedeutung der $\chi^2$-Verteilung ergibt sich unter anderem aus zwei wichtigen Testverfahren, dem $\chi^2$-Anpassungstest und dem $\chi^2$-Unabhängigkeitstest. Bei ersterem wird untersucht, ob ein Ergebnis einer Stichprobe mit einer gegebenen Verteilung verträglich ist, beim zweiten, ob zwischen zwei (diskreten) Mermalen ein Zusammenhang besteht.

Vorerst wieder ein Beispiel: Im Jahr 1989 gab es in Wien 15 941 Geburten, davon waren 7768 Mädchen- und 8173 Knabengeburten. Widerspricht dieses Ergebnis der Annahme, daß Mädchen- und Knabengeburten gleichverteilt sind?

Solche Fragen stellt man im allgemeinen, wenn man gute Gründe hat anzunehmen, daß das beobachtete Ergebnis nicht zufällig von der Gleichverteilung 1 : 1 abweicht, sonst würde man ja kaum auf die Idee kommen, so ein Problem zu untersuchen. Um es zu lösen, geht man wie beim indirekten Beweisen vor: Wir nehmen an, obiges Ergebnis ließe sich durch Zufall erklären und zeigen, daß diese Annahme höchst unwahrscheinlich ist.

Sei $X_1$ die Zufallsvariable, die die Anzahl der Mädchengeburten zählt. Diese ist dann – wenn unsere Annahme „Mädchen- und Knabengeburten sind gleichwahrscheinlich" zutreffend ist – binomialverteilt mit $n = 15941$ und $p_1 = \frac{1}{2}$. Da $n$ sehr groß ist, ist $X_1$ annähernd $N(\mu = np_1,\ \sigma^2 = np_1(1-p_1))$-verteilt. Somit ist

$$Z = \frac{X_1 - \mu}{\sigma} \text{ annnähernd } N(0,1)\text{-verteilt.}$$

Dann aber ist

$$Z^2 = \frac{(X_1 - \mu)^2}{\sigma^2} = \frac{(X_1 - np_1)^2}{np_1(1-p_1)}$$

$\chi^2_1$-verteilt. Wir formen im folgenden $Z^2$ um und verwenden dabei $X_2 = n - X_1$, die Zufallsvariable, die die Anzahl der Knabengeburten zählt, und $p_2 = 1 - p_1$, die Wahrscheinlichkeit für eine Knabengeburt:

$$Z^2 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2(p_2 + p_1)}{np_1p_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{np_2} =$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{[X_1 - n(1-p_2)]^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{[X_1 - (X_1 + X_2) + np_2])^2}{np_2} =$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \sum_{i=1}^{2} \frac{(X_i - np_i)^2}{np_i}$$

Und dies ist $\chi^2_{2-1}$-verteilt. Dabei sind $np_i$ die erwarteten relativen Häufigkeiten.

Durch vollständige Induktion ergibt sich der Fall der Multinomialverteilung, wo jeweils eines von $k$ Ereignissen $A_i$ mit der Wahrscheinlichkeit $p_i$ eintritt. Wenn $X_i$ die absolute Häufigkeit von $A_i$ zählt, dann ist die Testgröße

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

$\chi^2_{k-1}$-verteilt.

In obigem Beispiel ist $p_1 = p_2 = \frac{1}{2}$ und somit

$$T = \frac{(7768 - 15941 \cdot \frac{1}{2})^2}{15941 \cdot \frac{1}{2}} + \frac{(8173 - 15941 \cdot \frac{1}{2})^2}{15941 \cdot \frac{1}{2}} = 10,3$$

und daraus kann man mit Hilfe einer Tabelle (siehe Tab. 8) schließen, daß obiges Ergebnis höchst unwahrscheinlich ist; exakter ausgedrückt: Die Wahrscheinlichkeit, daß die Testgröße T einen Wert von 10 oder einen noch größeren annimmt, ist sehr klein.

# 4 Mehrfelder-Tafel – der $\chi^2$-Unabhängigkeitstest

Kehren wir wieder zu unserem Ausgangsbeispiel zurück: Bei unserem Beispiel sind wir davon ausgegangen, daß die Ereignisse voneinander unabhängig sind und haben berechnet, wie groß unter dieser Annahme bei festgehaltenen Randsummen die erwarteten Häufigkeiten wären.

Also: Wir beobachten die Ausprägungen zweier (verbundener) Zufallsvariabler $X$ und $Y$ mit den möglichen Werten $x_1 \ldots x_z$, bzw. $x_1 \ldots x_s$. Die in einer Stichprobe vom Umfang $n$ erhobenen (absoluten) Häufigkeiten $H_{ik}$ tragen wir in einer sogenannten Kontingenztafel ein:

Die Häufigkeiten $H_{ik}$ der Stichprobe sind natürlich Zufallsvariable; die Zeilensummen $H_{i0}$ stellen die Häufigkeiten $H(X = x_i)$ dar, die Spaltensummen $H_{0k}$ die Häufigkeiten $H(Y = y_k)$. Ihre Summe muß dann jeweils $n$ ergeben, womit eine gewisse „Probe“ gegen Schreib- und Zählfehler gegeben ist.

| Kontingenztafel | | | | | |
|---|---|---|---|---|---|
| $X$ | $Y$ | | | | $\Sigma$ |
| | $y_1$ | $y_2$ | $\dots$ | $y_s$ | |
| $x_1$ | $H_{11}$ | $H_{12}$ | $\dots$ | $H_{1s}$ | $H_{10}$ |
| $x_2$ | $H_{21}$ | $H_{22}$ | $\dots$ | $H_{2s}$ | $H_{20}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $H_{ik}$ | $\vdots$ | $\vdots$ |
| $x_z$ | $H_{z1}$ | $H_{z2}$ | $\dots$ | $H_{zs}$ | $H_{z0}$ |
| $\Sigma$ | $H_{01}$ | $H_{02}$ | $\dots$ | $H_{0s}$ | $H_{00} = n$ |

Tabelle 4: $n_z n_s$-Kontingenztabelle

Wir wissen: Für große $n$ approximieren die relativen Häufigkeiten $h_{ik} = \frac{H_{ik}}{n}$ die Wahrscheinlichkeiten $pik = P(X = x_i \vee Y = y_z)$. Analog gilt

$$\frac{H_{i0}}{n} = h_{i0} \approx P(X = x_i) = p_{i0}$$

und

$$\frac{H_{0k}}{n} = h_{0k} \approx P(Y = y_k) = p_{0k}.$$

Diese $p_{i0}$ und $p_{0k}$ sind die „Randverteilungen“ der zweidimensionalen Verteilung $\{p_{ik}\}$.

**Bemerkung:** Mathematisch exakter sind die $h_{ik}$, $h_{i0}$, $h_{0k}$ die *Maximum-Likelihood-Schätzungen* der an sich *unbekannten* Parameter $p_{ik}$, $p_{i0}$, $p_{0k}$ der Grundgesamtheit.

Nun zur eigentlichen Idee des Tests: Wir wissen, daß $X$ und $Y$ dann *unabhängig* sind, wenn $P(X = x_i \wedge Y = y_k) = P(X = x_i) \cdot P(Y = y_i)$ ist. Wenn also $X$ und $Y$ unabhängig sind, dann müßte für alle $i$, $k$ gelten:

$$p_{ik} = p_{i0} \cdot p_{0k}, \text{ also } h_{ik} = h_{i0} \cdot h_{0k} \text{ bzw. } H_{ik} = \frac{1}{n^2} H_{i0} \cdot H_{0k}$$

Wir gehen – wie stets – von der Richtigkeit der Annahme (= „*Hypothese*“) $H_0$: „$X$ und $Y$ sind unabhängig“ aus und vergleichen die in der Stichprobe *beobachteten* Werte $H_{ik}$ mit den unter $H_0$ „erwarteten“ Werten

$$E(H_{ik}) = np_{ik} = np_{i0} \cdot p_{0k} \approx nh_{i0} \cdot h_{0k} = \frac{1}{n} H_{i0} \cdot H_{0k}.$$

Wir betrachten allerdings nicht die Abweichungen $H_{ik} = E(H_{ik})$, sondern wir „gewichten“ diese, indem wir sie in Relation zu den Erwartungswerten $E(H_{ik})$

setzen und betrachten – damit sich entgegengesetzte Abweichungen nicht aufheben können – wie beim $\chi^2$-Anpassungstest als Testgröße T die Summe

$$T = \sum_{i=1}^{z}\sum_{k=1}^{s} \frac{[H_{ik} - E(H_{ik})]^2}{E(H_{ik})} = \sum_{i=1}^{z}\sum_{k=1}^{s} \frac{[H_{ik} - \frac{1}{n}H_{i0}\cdot H_{0k}]^2}{\frac{1}{n}H_{i0}\cdot H_{0k}} =$$

$$= n\sum_{i=1}^{z}\sum_{s}^{k=1} \frac{[H_{ik} - \frac{1}{n}H_{i0}\cdot H_{0k}]^2}{H_{i0}\cdot H_{0k}}$$

Die Analogie zum $\chi^2$-Anpassungstest ist klar: Wieder sollen die Ist-Werte $H_{ik}$ mit den Sollwerten $E(H_{ik})$ verglichen werden, und diese Erwartungswerte werden unter der Annahme, daß $H_0$ stimmt (Unabhängigkeit von $X$ und $Y$) durch $E(H_{ik}) = \frac{1}{n}H_{i0}\cdot H_{0k}$ berechnet. Wenn also $H_0$ stimmt, müßten alle $z\cdot s$ Summanden der Testgröße $T$ null sein oder nur wenig größer.

Die Summe kennen wir bereits, denn beim $\chi^2$-Anpassungstest hatten wir eine gleiche Summe als Testgröße verwendet, daher muß T $\chi^2$-verteilt sein.

Die Anzahl der Freiheitsgrade erhalten wir aus folgender Plausibilitätsüberlegung, die gleichzeitig die Verwendung des Wortes „Freiheitsgrad" einsichtig macht:

Beim $\chi^2$-Anpassungstest hatten wir bei $N$ verschiedenen Parametern $p_i$ genau $(N-1)$ viele Freiheitsgrade für die Schätzung. Da die Summe $\sum_{i=1}^{N} p_i = 1$ ist, können nur $(N-1)$ viele $p_i$ geschätzt werden. Die Testgröße T war daher $\chi^2_{N-1}$-verteilt. Nun betrachten wir unsere Testgröße

$$T = \sum_{i=1}^{z}\sum_{k=1}^{s} \frac{[H_{ik} - \frac{1}{n}H_{i0}\cdot H_{0k}]^2}{\frac{1}{n}H_{i0}\cdot H_{0k}}$$

Die $H_{ik} = nh_{ik}$ schätzen die Parameter $p_{ik}$. Es können aber nicht alle $p_{ik}$ frei geschätzt werden. Wegen $\sum_i p_{i0} = 1$ und $\sum_i p_{0k} = 1$ haben wir bei der Schätzung der $p_{ik}$ *weniger* Freiheitsgrade, nämlich nur $(n_z - 1)(n_s - 1)$ viele!

**Bemerkung:** Anders gedacht: Die Summe T aus den $n_z \cdot n_s$ Summanden der Testgrößenwerte der Stichprobe wäre zunächst $\chi^2_{n_z\cdot n_s - 1}$-verteilt, die Freiheitsgrade verringern sich aber nach einem bekannten Satz um die Anzahl der zu schätzenden Parameter, hier also um $(n_z-1)+(n_s-1)$, und es ist $n_z\cdot n_s - 1 - [(n_z-1)+(n_s-1)] = (n_z-1)\cdot(n_s-1)$.

## 4.1 Tabelle der Verteilungsfunktion der $\chi^2$-Verteilung

Um eine Tabelle der Verteilungsfunktion der $\chi^2$-Verteilung zu erhalten, braucht man „nur" folgendes Problem zu lösen:

Gesucht ist bei gegebener Irrtumswahrscheinlichkeit $\alpha$ (etwa =0,05) ein Wert $x$, sodaß

$$\int_0^x \frac{1}{2^{\frac{n}{2}}\left(\frac{n}{2}-1\right)!}\, y^{\frac{n}{2}-1}\, e^{-\frac{y}{2}}\, dy = 1-\alpha$$

Eine Möglichkeit wäre numerisch zu integrieren und zu interpolieren. Wir gehen hier aber einen ganz anderen Weg, der die früheren Überlegungen entbehrlich macht.

Im Prinzip geht es beim $\chi^2$-Anpassungstest darum, zu untersuchen, wie die Testgröße

$$\sum_{i=1}^{n} \frac{(h_i^{(b)} - h_i^{(e)})^2}{h_i^{(e)}}$$

verteilt ist. Dies können wir aber einfach mit einem kleinen Computerprogramm simulieren:

```
100 REM Simulation der Chiquadratverteilung
110 DIM It%(99)
120 INPUT " Eingabe der Freiheitsgrade f= ";F%
130 PRINT " Simulation der Chiquadratverteilung"
140 PRINT mit ";F%;" Freiheitsgraden"
150 PRINT " Eingabe der Wahrscheinlichkeiten:"
160 Pr%(0)=0
170 FOR I%=1 TO F%
180 PRINT p";I%;"=";: INPUT P!(I%)
190 Pr!(I%)=Pr!(I%-1)+P!(I%): REM Intervallgrenzen
200 NEXT I%
210 IF Pr!(F%)>1 THEN PRINT " Summe der Wahrscheinlichkeiten ist >1;
    noch einmal!": GOTO 150
220 P!(F%+1)=1-Pr!(F%)
230 Pr!(F%+1)=1
240 INPUT "Wie groß ist die Stichpobe? ";Ns%
250 INPUT " Und wie oft soll sie durchgeführt werden? ";Nt%
260 FOR In%=1 TO 20:It%(In%)=0: NEXT In%
270 FOR It%=1 TO Nt%
280 FOR I%=1 TO F%+1:Prs%(I%)=0: NEXT I%: REM Nullsetzen
290 FOR Is%=1 TO Ns%
300 X!= RND(1)
310 FOR I%=1 TO F%+1
320 IF X!<Pr!(I%) THEN Prs%(I%)=Prs%(I%)+1: GOTO 335
330 NEXT I%
335 REM
```

```
340 NEXT Is%
350 REM Berechne T
360 T!=0
370 FOR I%=1 TO F%+1
380 T!=T!+(Prs%(I%)-Ns%*P!(I%))^2/Ns%/P!(I%)
390 NEXT I%
400 REM Feststellen in welches der Intervalle [0;0,5],[0,5;1]... T fällt
410 Ii%=2*(T!+.5)
420 It%(Ii%)=It%(Ii%)+1
430 NEXT It%
440 FOR I%=1 TO 20
450 PRINT "[";(I%-1)/2;";";I%/2;") ";It%(I%)
460 NEXT I%
470 END
```

Dieses Programm läßt uns folgende Vermutungen erhalten und testen:

Wenn die Wahrscheinlichkeiten, die man testet, stimmen, dann ist die Verteilung der Testgröße T

- weitgehend unabhängig von der Versuchszahl (siehe Tab. 5),
- weitgehend unabhängig von den einzelnen Wahrscheinlichkeiten $p_i$ (siehe Tab. 5) und
- strebt mit wachsendem Versuchsumfang einer Grenzverteilung, die unabhängig von den $p_i$ ist, zu (siehe Tab. 5, 6 und 7).

Wir wählen zwei Freiheitsgrade, weil wir zur Kontrolle das Ergebnis elementar berechnen können (siehe Bemerkung in Abschnitt 2).

| Simulation der Chiquadratverteilung mit 2 Freiheitsgraden | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Wahrscheinlichkeiten | Stichprobe=30 | | | | Stichprobe=100 | | | |
| | $p_1 = 0,333$ $p_2 = 0,333$ $p_3 = 0,333$ | | $p_1 = 0,500$ $p_2 = 0,300$ $p_3 = 0,200$ | | $p_1 = 0,333$ $p_2 = 0,333$ $p_3 = 0,333$ | | $p_1 = 0,500$ $p_2 = 0,300$ $p_3 = 0,200$ | |
| | Treffer | Vtlg. | Treffer | Vtlg. | Treffer | Vtlg. | Treffer | Vtlg. |
| [0,0; 0,5) | 11 | 11 | 12 | 12 | 6 | 6 | 19 | 19 |
| [0,5; 1) | 11 | 22 | 24 | 36 | 27 | 33 | 17 | 36 |
| [ 1; 1,5) | 8 | 30 | 14 | 50 | 8 | 41 | 18 | 54 |
| [ 1,5; 2) | 18 | 48 | 9 | 59 | 10 | 51 | 11 | 65 |
| [ 2; 2,5) | 9 | 57 | 13 | 72 | 11 | 62 | 7 | 72 |
| [ 2,5; 3) | 14 | 71 | 5 | 77 | 8 | 70 | 8 | 80 |
| [ 3; 3,5) | 3 | 74 | 1 | 78 | 5 | 75 | 2 | 82 |
| [ 3,5; 4) | 0 | 74 | 9 | 87 | 5 | 80 | 5 | 87 |
| [ 4; 4,5) | 12 | 86 | 0 | 87 | 6 | 86 | 3 | 90 |
| [ 4,5; 5) | 0 | 86 | 3 | 90 | 3 | 89 | 1 | 91 |
| [ 5; 5,5) | 2 | 88 | 3 | 93 | 3 | 92 | 3 | 94 |
| [ 5,5; 6) | 5 | 93 | 3 | 96 | 0 | 92 | 1 | 95 |
| [ 6; 6,5) | 1 | 94 | 1 | 97 | 0 | 92 | 1 | 96 |
| [ 6,5; 7) | 0 | 94 | 2 | 99 | 4 | 96 | 1 | 97 |
| [ 7; 7,5) | 2 | 96 | 0 | 99 | 0 | 96 | 1 | 98 |
| [ 7,5; 8) | 0 | 96 | 0 | 99 | 2 | 98 | 0 | 98 |
| [ 8; 8,5) | 1 | 97 | 0 | 99 | 0 | 98 | 0 | 98 |
| [ 8,5; 9) | 0 | 97 | 0 | 99 | 0 | 98 | 1 | 99 |
| [ 9; 9,5) | 1 | 98 | 0 | 99 | 0 | 98 | 0 | 99 |
| [ 9,5; 10) | 0 | 98 | 0 | 99 | 1 | 99 | 1 | 100 |

Tabelle 5: $\chi^2$-Simulation; Versuchsanzahl=100

| Simulation der Chiquadratverteilung mit 2 Freiheitsgraden | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Wahrscheinlichkeiten | Stichprobe=30 | | | | Stichprobe=100 | | | |
| | $p_1 = 0,333$ $p_2 = 0,333$ $p_3 = 0,333$ | | $p_1 = 0,500$ $p_2 = 0,300$ $p_3 = 0,200$ | | $p_1 = 0,333$ $p_2 = 0,333$ $p_3 = 0,333$ | | $p_1 = 0,500$ $p_2 = 0,300$ $p_3 = 0,200$ | |
| | Treffer | Vtlg. | Treffer | Vtlg. | Treffer | Vtlg. | Treffer | Vtlg. |
| [0,0; 0,5) | 155 | 15,5 | 151 | 15,1 | 78 | 7,8 | 128 | 12,8 |
| [0,5; 1) | 125 | 28,0 | 195 | 34,6 | 227 | 30,5 | 177 | 30,5 |
| [ 1; 1,5) | 134 | 41,4 | 154 | 50,0 | 125 | 43,0 | 168 | 47,3 |
| [ 1,5; 2) | 183 | 59,7 | 122 | 62,2 | 109 | 53,9 | 138 | 61,1 |
| [ 2; 2,5) | 66 | 66,3 | 96 | 71,8 | 132 | 67,1 | 56 | 66,7 |
| [ 2,5; 3) | 144 | 80,7 | 52 | 77,0 | 77 | 74,8 | 81 | 74,8 |
| [ 3; 3,5) | 32 | 83,9 | 22 | 79,2 | 56 | 80,4 | 45 | 79,3 |
| [ 3,5; 4) | 0 | 83,9 | 88 | 88,0 | 43 | 84,7 | 39 | 83,2 |
| [ 4; 4,5) | 82 | 92,1 | 13 | 89,3 | 39 | 88,6 | 37 | 96,9 |
| [ 4,5; 5) | 0 | 92,1 | 26 | 91,9 | 16 | 90,2 | 25 | 89,4 |
| [ 5; 5,5) | 7 | 92,8 | 14 | 93,3 | 26 | 92,8 | 26 | 92,0 |
| [ 5,5; 6) | 26 | 95,4 | 23 | 95,6 | 9 | 93,7 | 16 | 93,6 |
| [ 6; 6,5) | 18 | 97,2 | 14 | 97,0 | 23 | 96,0 | 21 | 95,7 |
| [ 6,5; 7) | 0 | 97,2 | 8 | 97,8 | 12 | 97,2 | 14 | 97,1 |
| [ 7; 7,5) | 3 | 97,5 | 8 | 98,6 | 5 | 97,7 | 8 | 97,9 |
| [ 7,5; 8) | 10 | 98,5 | 3 | 98,9 | 9 | 98,6 | 3 | 98,2 |
| [ 8; 8,5) | 4 | 98,9 | 3 | 99,2 | 2 | 98,8 | 4 | 98,6 |
| [ 3,5; 9) | 4 | 99,3 | 1 | 99,3 | 1 | 98,9 | 3 | 98,9 |
| [ 9; 9,5) | 0 | 99,3 | 2 | 99,5 | 3 | 99,2 | 2 | 99,1 |
| [ 9,5; 10) | 1 | 99,4 | 2 | 99,7 | 3 | 99,5 | 4 | 99,5 |

Tabelle 6: $\chi^2$-Simulation; Versuchsanzahl=1000

| Simulation der Chiquadratverteilung mit 2 Freiheitsgraden | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Wahrscheinlichkeiten | Stichprobe=30 $p_1 = 0,333$ $p_2 = 0,333$ $p_3 = 0,333$ | | Stichprobe=30 $p_1 = 0,500$ $p_2 = 0,300$ $p_3 = 0,200$ | | Stichprobe=100 $p_1 = 0,333$ $p_2 = 0,333$ $p_3 = 0,333$ | | Stichprobe=100 $p_1 = 0,500$ $p_2 = 0,300$ $p_3 = 0,200$ | |
| | Treffer | Vtlg. | Treffer | Vtlg. | Treffer | Vtlg. | Treffer | Vtlg. |
| [0,0; 0,5) | 1759 | 17,6 | 1373 | 13,7 | 1003 | 10,3 | 1131 | 11,3 |
| [0,5; 1) | 1183 | 29,4 | 1838 | 32,1 | 2332 | 33,4 | 1931 | 30,6 |
| [ 1; 1,5) | 1142 | 40,8 | 1469 | 46,8 | 1315 | 46,5 | 1590 | 46,5 |
| [ 1,5; 2) | 1590 | 56,7 | 1098 | 57,8 | 1117 | 57,7 | 1221 | 58,7 |
| [ 2; 2,5) | 651 | 63,3 | 1141 | 69,2 | 1136 | 69,0 | 804 | 66,8 |
| [ 2,5; 3) | 1409 | 77,3 | 741 | 76,6 | 582 | 74,9 | 825 | 75,0 |
| [ 3; 3,5) | 322 | 80,7 | 146 | 78,1 | 464 | 79,5 | 506 | 80,1 |
| [ 3,5; 4) | 0 | 80,7 | 850 | 86,6 | 523 | 84,7 | 412 | 84,2 |
| [ 4; 4,5) | 951 | 90,2 | 149 | 88,1 | 360 | 88,3 | 425 | 88,5 |
| [ 4,5; 5) | 0 | 90,2 | 273 | 90,8 | 236 | 90,7 | 171 | 90,2 |
| [ 5; 5,5) | 141 | 91,6 | 165 | 92,4 | 255 | 93,2 | 267 | 92,8 |
| [ 5,5; 6) | 336 | 94,6 | 244 | 94,9 | 145 | 94,7 | 111 | 94,0 |
| [ 6; 6,5) | 167 | 96,6 | 123 | 96,1 | 124 | 95,9 | 165 | 95,6 |
| [ 6,5; 7) | 0 | 96,6 | 78 | 96,9 | 99 | 96,9 | 104 | 96,7 |
| [ 7; 7,5) | 43 | 97,1 | 45 | 97,3 | 50 | 97,74 | 90 | 97,6 |
| [ 7,5; 8) | 86 | 97,9 | 39 | 97,7 | 55 | 98,0 | 52 | 98,1 |
| [ 8; 8,5) | 69 | 98,6 | 89 | 98,6 | 61 | 98,6 | 50 | 98,6 |
| [ 8,5; 9) | 44 | 99,0 | 29 | 98,9 | 32 | 98,9 | 22 | 98,8 |
| [ 9; 9,5) | 0 | 99,0 | 15 | 99,1 | 11 | 99,0 | 33 | 99,1 |
| [ 9,5; 10) | 11 | 99,2 | 8 | 99,1 | 22 | 99,2 | 24 | 99,4 |

Tabelle 7: $\chi^2$-Simulation; Versuchsanzahl=10 000

Betrachtet man Tab. 7 erkennt man: Ist der Wert der Testgröße größer als 6 kommen im allgemeinen weniger als 0,05% der Fälle vor. Macht man das für verschiedene Freiheitsgrade und mit einer feineren Unterteilung (dazu braucht man allerdings einen größeren Stichprobenumfang), kann man die Tab. 8 gewinnen.

Nun können wir unsere beiden Beispiele lösen: Im Dioxinbeispiel waren $(6-1)(2-1) = 5$ Freiheitsgrade. Bei einer Irrtumswahrscheinlichkeit $\alpha = 1\%$ müßte die Prüfgröße $< 15,09$ sein. Tatsächlich ist sie aber $93,53$, also weit größer. Daher ist die Annahme, eine Dioxinbelastung des Mannes hat keinen Einfluß auf die Schwangerschaft, höchst unwahrscheinlich und es kann das Gegenteil angenommen

| Freiheits- | Irrtumswahrscheinlichkeit | | Freiheits- | Irrtumswahrscheinlichkeit | |
|---|---|---|---|---|---|
| grad | $\alpha = 5\%$ | $\alpha = 1\%$ | grad | $\alpha = 5\%$ | $\alpha = 1\%$ |
| 1 | 3,84 | 6,64 | 10 | 18,31 | 23,21 |
| 2 | 5,99 | 9,21 | 15 | 25,00 | 30,58 |
| 3 | 7,82 | 11,35 | 20 | 31,41 | 37,57 |
| 4 | 9,49 | 13,28 | 25 | 37,65 | 44,31 |
| 5 | 11,07 | 15,09 | 30 | 43,77 | 50,89 |
| 6 | 12,59 | 16,81 | 40 | 55,76 | 63,69 |
| 7 | 14,07 | 18,48 | 50 | 67,50 | 76,20 |
| 8 | 15,51 | 20,09 | 60 | 79,08 | 88,38 |
| 9 | 16,92 | 21,67 | 120 | 146,57 | 158,95 |

Tabelle 8: $\chi^2$-Tabelle

werden: Eine Dioxinbelastung des Mannes hat auf den Verlauf der Schwangerschaft einen Einfluß.

Beim Geburtenbeispiel trat 1 Freiheitsgrad auf. Bei einer Irrtumswahrscheinlichkeit $\alpha = 1\%$ müßte die Prüfgröße somit $< 6,64$ sein. Tatsächlich ist sie aber mit $10,3$ weit größer. Daher ist die Annahme „Mädchen- und Knabengeburten sind gleichwahrscheinlich“ höchst unwahrscheinlich und es kann auch hier das Gegenteil angenommen werden.

# 5 Anwendung in der Schule

Meines Erachtens ist die Herleitung der $\chi^2$-Verteilung zu aufwendig, als daß sie (trotz des Tricks mit dem Schluß von $n$ auf $n+2$) im Mathematikunterricht einer höheren Schule Platz hätte. Trotzdem ist die $\chi^2$-Verteilung eine so wichtige Verteilung, daß zu überlegen wäre, sie mit Hilfe des angegebenen Computerprogramms zu simulieren und die beiden Testverfahren zu besprechen. Ein ähnliches Problem tritt ja auch bei der Normalverteilung auf. Sicher wird man ohne diese schwer auskommen können, obwohl auch bei dieser die Herleitung sehr aufwendig ist.

Etwas anderes ist, wenn im Wahlpflichtfach Mathematik auf „Stochastik“ der Schwerpunkt gelegt wird. Hier sollte sowohl die Herleitung der Normal- als auch die der $\chi^2$-Verteilung besprochen werden.