Die Wahrscheinlichkeit für den Fehler 1. Art wird mit α bezeichnet, und man spricht deshalb gelegentlich vom α-Fehler. α heißt auch Signifikanzniveau des Tests.

In der Praxis ist es üblich, sich ein Signifikanzniveau α vorzugeben (übliche Werte hierfür sind $0{,}05$, $0{,}01$ oder $0{,}001$) und dann den Test so auszulegen (also den Ablehnungsbereich K so zu bestimmen), dass die Wahrscheinlichkeit für den Fehler 1. Art den Wert α besitzt.

Konstruktion eines einfachen Tests

Wir konstruieren einen Test für den Parameter p einer Bernoulli-verteilten Zufallsvariablen X. Wir setzen

$$H_0 : p \geq p_0, \qquad H_1 : p < p_0.$$

Als Testgröße verwenden wir

$$T := X_1 + \ldots + X_n.$$

Für größere Wahrscheinlichkeiten p erwarten wir auch größere Werte für T. Deshalb ist es sinnvoll, einen Ablehnungsbereich der Art $K := [0, k]$ für T zu wählen, wobei $k \in \mathbb{R}$ geeignet festzulegen ist. Wir konstruieren hier also einen einseitigen Test, während für eine Nullhypothese $H_0 : p = p_0$ sowohl zu kleine als auch zu große Werte von T zur Ablehnung von H_0 führen sollten und somit ein zweiseitiger Test vorzuziehen wäre.

T ist binomialverteilt. Da wir von einem großen Stichprobenumfang n ausgehen, bietet es sich an, die Verteilung von T nach dem Grenzwertsatz von de Moivre (siehe Korollar 109) durch die Normalverteilung zu approximieren.

Sei

$$\tilde{T} := \frac{T - np}{\sqrt{np(1-p)}}.$$

$\tilde{T}$ ist annähernd standardnormalverteilt.

Wir berechnen für jeden Wert von k das zugehörige Signifikanzniveau α des Tests.

$$\begin{aligned}
\text{Fehlerwahrscheinlichkeit 1. Art} &= \max_{p \in H_0} \Pr_p[T \in K] \\
&= \max_{p \in H_0} \Pr_p[T \leq k] \\
\text{Fehlerwahrscheinlichkeit 2. Art} &= \sup_{p \in H_1} \Pr_p[T \notin K] \\
&= \sup_{p \in H_1} \Pr_p[T > k]
\end{aligned}$$

Für den Fehler 1. Art α erhalten wir

$$\begin{aligned}
\alpha &= \max_{p \geq p_0} \Pr_p[T \leq k] = \Pr_{p=p_0}[T \leq k] \\
&= \Pr_{p=p_0}\left[\tilde{T} \leq \frac{k - np}{\sqrt{np(1-p)}}\right] \\
&= \Pr\left[\tilde{T} \leq \frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right] \approx \Phi\left(\frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right).
\end{aligned}$$

Unter Verwendung der Quantile der Standardnormalverteilung ergibt sich damit:

- Ist k so gewählt, dass $(k - np_0)/\sqrt{np_0(1-p_0)} = z_\alpha$, so ist das Signifikanzniveau gleich α.
- Ist das gewünschte Signifikanzniveau α des Tests vorgegeben, so erhält man den Wert $k = k(n)$ in Abhängigkeit vom Umfang n der Stichprobe durch

$$k = z_\alpha \cdot \sqrt{np_0(1-p_0)} + np_0. \tag{8}$$

Kleinere Werte für k verkleinern zwar den Fehler 1. Art, vergrößern jedoch den Annahmebereich und damit die Wahrscheinlichkeit für einen Fehler 2. Art.

Verhalten der Testfehler

Wie verhalten sich die möglichen Testfehler des konstruierten Verfahrens? Was geschieht beispielsweise, wenn p nur geringfügig kleiner als p_0 ist?

In diesem Fall betrachten wir beim Fehler 2. Art die Wahrscheinlichkeit

$$\Pr_{p=p_0-\varepsilon}[T \geq k] \approx \Pr_{p=p_0}[T \geq k] \approx 1 - \alpha \ .$$

Wenn sich also die „wahren“ Verhältnisse nur minimal von unserer Nullhypothese unterscheiden, so werden wir diese „im Zweifelsfall“ annehmen.

Bei echten Alternativtests werden für hinreichend große Stichproben und einen geeignet eingestellten Ablehnungsbereich beide Testfehler klein.

Beispiel 121

Die Abbruchrate p der Transaktionen in einem Online-Datenbanksystem wurde bereits früher einmal ermittelt. Allerdings sind die entsprechenden Daten verloren gegangen und die Entwickler erinnern sich nur noch, dass das Ergebnis entweder $p = 1/3$ oder $p = 1/6$ lautete. Unter dieser Annahme würde man den Test wie folgt ansetzen:

$$H_0 : p \geq 1/3, \qquad H_1' : p \leq 1/6.$$

Beispiel (Forts.)

Für den Fehler 2. Art erhält man nun:

$$\begin{aligned} \textit{Fehlerwahrsch. 2. Art} &= \max_{p \le 1/6} \Pr_p[T > k] \\ &\approx 1 - \Phi\left(\frac{k - (1/6) \cdot n}{\sqrt{(1/6) \cdot (5/6) n}}\right). \end{aligned}$$

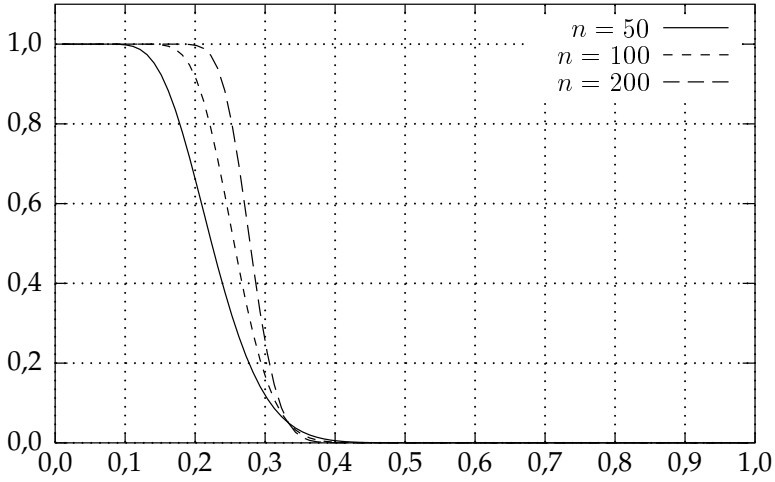
Mit den obigen Werten $k = 25$ *und* $n = 100$ *ergibt sich mit*

$$\Phi\left(\frac{150 - 100}{\sqrt{5} \cdot 10}\right) = \Phi(\sqrt{5}) \approx 0{,}9871$$

ein Fehler 2. Art der Größe $0{,}0129$*, während sich für die triviale Alternative* $H_1 : p < 1/3$ *ein Wert von etwa* $0{,}95$ *ergibt.*

Die so genannte Gütefunktion g gibt allgemein die Wahrscheinlichkeit an, mit der ein Test die Nullhypothese verwirft. Für unser hier entworfenes Testverfahren gilt

$$g(n,p) = \Pr_p[T \in K] = \Pr_p[T \leq k] \approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right).$$



Gütefunktion $g(n, p)$ für verschiedene Werte von n

Man erkennt deutlich, dass für alle n der Wert von $k = k(n)$ genau so gewählt wurde, dass $g(n, 1/3) = 0{,}05$ gilt. Dies wird durch den in Gleichung 8 angegebenen Ausdruck erreicht.

Für Werte von p größer als $1/3$ wird $H_0 : p \geq 1/3$ mit hoher Wahrscheinlichkeit angenommen, während für Werte deutlich unter $1/3$ die Hypothese H_0 ziemlich sicher abgelehnt wird.

Ferner ist auffällig, dass g für größere Werte von n schneller von Eins auf Null fällt. Daran erkennt man, dass durch den Test die Fälle „H_0 gilt“ und „H_0 gilt nicht“ umso besser unterschieden werden können, je mehr Stichproben durchgeführt werden. Für Werte von p, bei denen $g(n, p)$ weder nahe bei Eins noch nahe bei Null liegt, kann der Test nicht sicher entscheiden, ob die Nullhypothese abzulehnen ist.

4.2 Praktische Anwendung statistischer Tests

Das im vorhergehenden Abschnitt konstruierte Testverfahren taucht in der Literatur unter dem Namen approximativer Binomialtest auf.

Die folgende Tabelle 1 gibt einen Überblick über die Eckdaten dieses Tests.

Tabelle : **Approximativer Binomialtest**

Annahmen:

$X_1, \ldots, X_n$ *seien unabhängig und identisch verteilt mit* $\Pr[X_i = 1] = p$ *und* $\Pr[X_i = 0] = 1-p$, *wobei* p *unbekannt sei.* n *sei hinreichend groß, so dass die Approximation aus Korollar 109 brauchbare Ergebnisse liefert.*

Hypothesen:

$$\begin{array}{lllll} a) & H_0 : p = p_0 & \textit{gegen} & H_1 : p \neq p_0, \\ b) & H_0 : p \geq p_0 & \textit{gegen} & H_1 : p < p_0, \\ c) & H_0 : p \leq p_0 & \textit{gegen} & H_1 : p > p_0. \end{array}$$

Testgröße:

$$Z := \frac{h - np_0}{\sqrt{np_0(1-p_0)}},$$

wobei $h := X_1 + \ldots + X_n$ *die Häufigkeit bezeichnet, mit der die Ereignisse* $X_i = 1$ *aufgetreten sind.*

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$\begin{array}{ll} a) & |Z| > z_{1-\alpha/2}, \\ b) & Z < z_\alpha, \\ c) & Z > z_{1-\alpha}. \end{array}$$

4.3 Allgemeines Vorgehen bei statistischen Tests

1. Schritt: Formulierung von Annahmen. Ganz ohne Annahmen kommt man meist nicht aus. Übliche Annahmen betreffen meist die Verteilung der Stichprobenvariablen und deren Unabhängigkeit.
2. Schritt: Formulierung der Nullhypothese.
3. Schritt: Auswahl des Testverfahrens.
4. Schritt: Durchführung des Tests und Entscheidung.

4.4 Ausgewählte statistische Tests

4.4.1 Wie findet man das richtige Testverfahren?

Statistische Tests kann man nach mehreren Kriterien in Klassen einteilen.

- **Anzahl der beteiligten Zufallsgrößen**
  Sollen zwei Zufallsgrößen mit potentiell unterschiedlichen Verteilungen verglichen werden, für die jeweils eine Stichprobe erzeugt wird (Zwei-Stichproben-Test), oder wird nur eine einzelne Zufallsgröße untersucht (Ein-Stichproben-Test)?

Bei der Fragestellung

> *Beträgt die mittlere Zugriffszeit auf einen Datenbankserver im Mittel höchstens 10ms?*

hat man es mit einem Ein-Stichproben-Test zu tun, während die Untersuchung der Frage

> *Hat Datenbankserver A eine kürzere mittlere Zugriffszeit als Datenbankserver B?*

auf einen Zwei-Stichproben-Test führt.

Bei mehreren beteiligten Zufallsgrößen wird zusätzlich unterschieden, ob aus voneinander unabhängigen Grundmengen Stichproben erhoben werden oder nicht. Beim vorigen Beispiel werden unabhängige Messungen vorgenommen, sofern die Server A und B getrennt voneinander arbeiten. Wenn man jedoch die Frage

> *Läuft ein Datenbankserver auf einer Menge festgelegter Testanfragen mit Query-Optimierung schneller als ohne?*

untersucht, so spricht man von verbundenen Messungen.

Gelegentlich betrachtet man auch den Zusammenhang zwischen mehreren Zufallsgrößen. Beispielsweise könnte man sich für die Frage interessieren:

> *Wie stark wächst der Zeitbedarf für eine Datenbankanfrage im Mittel mit der (syntaktischen) Länge der Anfrage, d. h. führen kompliziertere Formulierungen zu proportional längeren Laufzeiten?*

Mit solchen Fragenstellungen, bei denen ein funktionaler Zusammenhang zwischen Zufallsgrößen ermittelt werden soll, beschäftigt sich die Regressionsanalyse. Wenn überhaupt erst zu klären ist, ob ein solcher Zusammenhang besteht oder ob die Zufallsgrößen vielmehr unabhängig voneinander sind, so spricht man von Zusammenhangsanalyse.

- **Formulierung der Nullhypothese**
  Welche Größe dient zur Definition der Nullhypothese? Hierbei werden in erster Linie Tests unterschieden, die Aussagen über verschiedene so genannte Lageparameter treffen, wie z.B. den Erwartungswert oder die Varianz der zugrunde liegenden Verteilungen.

  Im Zwei-Stichproben-Fall könnte man beispielsweise untersuchen, ob der Erwartungswert der Zufallsgröße A größer oder kleiner als bei Zufallsgröße B ist.

  Gelegentlich wird zur Formulierung der Nullhypothese auch der so genannte Median betrachtet: Der Median einer Verteilung entspricht dem (kleinsten) Wert x mit $F(x) = 1/2$.

  Neben solchen Tests auf Lageparameter gibt es z.B. auch Tests, die auf eine vorgegebene Verteilung oder auf ein Maß für die Abhängigkeit verschiedener Zufallsgrößen testen.

- **Annahmen über die Zufallsgrößen**
  Was ist über die Verteilung der untersuchten Größe(n) bekannt? Bei entsprechenden Annahmen könnte es sich z.B. um die Art der Verteilung, den Erwartungswert oder die Varianz handeln.

4.4.2 Ein-Stichproben-Tests für Lageparameter

Beim approximativen Binomialtest wird ausgenutzt, dass die Binomialverteilung für große n nach dem Grenzwertsatz von de Moivre (Korollar 109) gegen die Normalverteilung konvergiert. Aus diesem Grund kann man diesen Test auch als Spezialfall eines allgemeineren Testverfahrens ansehen, nämlich des Gaußtest, der nun dargestellt wird.

Tabelle : **Gaußtest**

Annahmen:

$X_1, \ldots, X_n$ *seien unabhängig und identisch verteilt mit* $X_i \sim \mathcal{N}(\mu, \sigma^2)$, *wobei* σ^2 *bekannt ist. Alternativ gelte* $\mathbb{E}[X_i] = \mu$ *und* $\mathrm{Var}[X_i] = \sigma^2$, *und* n *sei groß genug.*

Hypothesen:

$$\begin{array}{lllll} a) & H_0 : \mu = \mu_0 & \textit{gegen} & H_1 : \mu \neq \mu_0, \\ b) & H_0 : \mu \geq \mu_0 & \textit{gegen} & H_1 : \mu < \mu_0, \\ c) & H_0 : \mu \leq \mu_0 & \textit{gegen} & H_1 : \mu > \mu_0. \end{array}$$

Testgröße:

$$Z := \frac{\overline{X} - \mu_0}{\sigma} \sqrt{n}\,.$$

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$\begin{array}{ll} a) & |Z| > z_{1-\alpha/2}, \\ b) & Z < z_{\alpha}, \\ c) & Z > z_{1-\alpha}. \end{array}$$

Der Gaußtest hat den Nachteil, dass man die Varianz σ^2 der beteiligten Zufallsgrößen kennen muss.

Wenn diese unbekannt ist, so liegt es nahe, die Varianz durch die Stichprobenvarianz S^2 (siehe Definition 114) anzunähern. Dies führt auf den so genannten t-Test, der in der folgenden Übersicht dargestellt ist.

Tabelle : *t*-Test

Annahmen:

$X_1, \ldots, X_n$ *seien unabhängig und identisch verteilt mit* $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
Alternativ gelte $\mathbb{E}[X_i] = \mu$ *und* $\mathrm{Var}[X_i] = \sigma^2$, *und* n *sei groß genug.*

Hypothesen:

$$\begin{array}{lllll} a) & H_0 : \mu = \mu_0 & \textit{gegen} & H_1 : \mu \neq \mu_0, \\ b) & H_0 : \mu \geq \mu_0 & \textit{gegen} & H_1 : \mu < \mu_0, \\ c) & H_0 : \mu \leq \mu_0 & \textit{gegen} & H_1 : \mu > \mu_0. \end{array}$$

Testgröße:

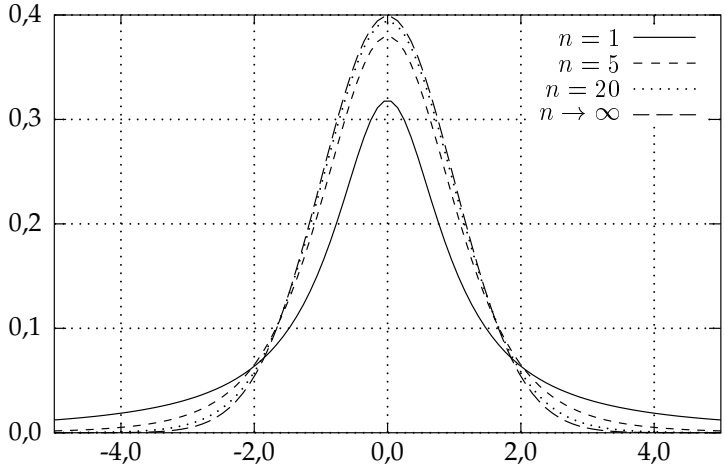
$$T := \frac{\overline{X} - \mu_0}{S}\sqrt{n}.$$

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$\begin{array}{ll} a) & |T| > t_{n-1,1-\alpha/2}, \\ b) & T < t_{n-1,\alpha}, \\ c) & T > t_{n-1,1-\alpha}. \end{array}$$

Hierbei gibt $t_{n-1,1-\alpha}$ das $(1-\alpha)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden an. Die t-Verteilung taucht manchmal auch unter dem Namen Student-Verteilung auf, da sie ursprünglich unter dem Pseudonym „Student“ publiziert wurde.

Wir gehen an dieser Stelle nicht darauf ein, wieso die Testgröße die t-Verteilung besitzt, sondern weisen nur darauf hin, dass die Dichte dieser Verteilung (eigentlich handelt es sich um eine ganze Familie von Verteilungen, da die Anzahl der Freiheitsgrade jeweils noch gewählt werden kann) der Dichte der Normalverteilung ähnelt. Für große n (Faustregel: $n \geq 30$) liegen die beiden Dichten so genau übereinander, dass man in der Praxis die t-Verteilung durch die Normalverteilung annähert.



Dichte der t-Verteilung mit n Freiheitsgraden

Als weitere Beispiele für gängige Ein-Stichproben-Tests zu Lageparametern seien der Wilcoxon-Test und der χ^2-Varianztest genannt. Ersterer dient zum Testen von Hypothesen zum Median, während der zweite Test Hypothesen zur Varianz beinhaltet.

4.4.3 Zwei-Stichproben-Tests für Lageparameter

Bei Zwei-Stichproben-Tests wollen wir das Verhältnis von Lageparametern untersuchen. Besonders wichtig sind hierbei Tests zum Erwartungswert. Für zwei Zufallsgrößen X und Y könnten wir beispielsweise die Frage untersuchen, ob für die Erwartungswerte μ_X und μ_Y gilt, dass $\mu_X = \mu_Y$ ist.

Tabelle : **Zwei-Stichproben-t-Test**

Annahmen:

$X_1, \ldots, X_m$ *und* $Y_1, \ldots, Y_n$ *seien unabhängig und jeweils identisch verteilt, wobei* $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *und* $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ *gelte. Die Varianzen seien identisch, also* $\sigma_X^2 = \sigma_Y^2$.

Hypothesen:

$$\begin{array}{lllll} a) & H_0 : \mu_X = \mu_Y & \text{gegen} & H_1 : \mu_X \neq \mu_Y, \\ b) & H_0 : \mu_X \geq \mu_Y & \text{gegen} & H_1 : \mu_X < \mu_Y, \\ c) & H_0 : \mu_X \leq \mu_Y & \text{gegen} & H_1 : \mu_X > \mu_Y. \end{array}$$

Testgröße:

$$T := \sqrt{\frac{n+m-2}{\frac{1}{m}+\frac{1}{n}}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{(m-1)\cdot S_X^2 + (n-1)\cdot S_Y^2}}.$$

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$\begin{array}{ll} a) & |T| > t_{m+n-2,1-\alpha/2}, \\ b) & T < t_{m+n-2,\alpha}, \\ c) & T > t_{m+n-2,1-\alpha}. \end{array}$$

Vom Zwei-Stichproben-t-Test findet man in der Literatur noch zusätzliche Varianten, die auch dann einsetzbar sind, wenn die beteiligten Zufallsgrößen nicht dieselbe Varianz besitzen. Der beim Ein-Stichproben-Fall erwähnte Wilcoxon-Test kann ebenfalls auf den Zwei-Stichproben-Fall übertragen werden.

4.4.4 Nicht an Lageparametern orientierte Tests

Wir betrachten in diesem Abschnitt exemplarisch den χ^2-Anpassungstest. Bei einem Anpassungstest wird nicht nur der Lageparameter einer Verteilung getestet, sondern es wird die Verteilung als Ganzes untersucht.

Beim approximativen Binomialtest (siehe Tabelle 1) haben wir streng genommen bereits einen Anpassungstest durchgeführt. Bei der Nullhypothese $H_0 : p = p_0$ wird untersucht, ob es sich bei der betrachteten Zufallsgröße um eine Bernoulli-verteilte Zufallsvariable mit Parameter p_0 handelt. Beim χ^2-Test gehen wir nun einen Schritt weiter: Wir nehmen an, dass die Zufallsgröße X genau k verschiedene Werte annimmt. Ohne Beschränkung der Allgemeinheit sei $W_X = \{1, \ldots, k\}$. Die Nullhypothese lautet nun

$$H_0 : \Pr[X = i] = p_i \quad \text{für } i = 1, \ldots, k.$$

Tabelle : χ^2-Anpassungstest

Annahmen:

$X_1, \ldots, X_n$ *seien unabhängig und identisch verteilt mit* $W_{X_i} = \{1, \ldots, k\}$.

Hypothesen:

$$H_0 \ : \ \Pr[X = i] = p_i \quad \textit{für } i = 1, \ldots, k,$$
$$H_1 \ : \ \Pr[X = i] \neq p_i \quad \textit{für mindestens ein } i \in \{1, \ldots, k\},$$

Testgröße:

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

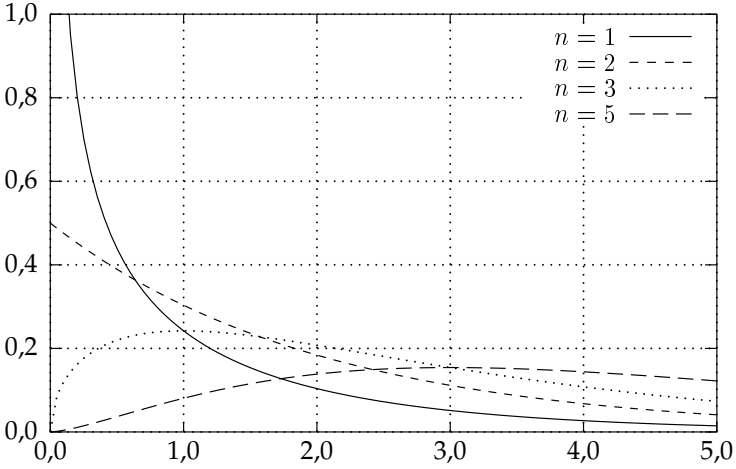
wobei h_i *die Häufigkeit angibt, mit der* $X_1, \ldots, X_n$ *den Wert* i *angenommen haben.*

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$T > \chi^2_{k-1,1-\alpha};$$

dabei sollte gelten, dass $np_i \geq 1$ *für alle* i *und* $np_i \geq 5$ *für mindestens* 80% *der Werte* $i = 1, \ldots, k$.

Für die Testgröße T wird näherungsweise eine χ^2-Verteilung mit $k-1$ Freiheitsgraden angenommen. Die Werte dieser Verteilung finden sich in entsprechenden Tabellen in der Literatur. Damit diese Approximation gerechtfertigt ist, sollte gelten, dass $np_i \geq 1$ für alle i und $np_i \geq 5$ für mindestens 80% der Werte $i = 1, \ldots, k$. Das γ-Quantil einer χ^2-Verteilung mit k Freiheitsgraden bezeichnen wir mit $\chi^2_{k,\gamma}$.



Dichte der χ^2-Verteilung mit n Freiheitsgraden

Beispiel 122

Als Anwendung für den χ^2-Test wollen wir überprüfen, ob der Zufallszahlengenerator von Maple eine gute Approximation der Gleichverteilung liefert. Dazu lassen wir Maple $n = 100000$ Zufallszahlen aus der Menge $\{1, \ldots, 10\}$ generieren. Wir erwarten, dass jede dieser Zahlen mit gleicher Wahrscheinlichkeit $p_1 = \ldots = p_{10} = 1/10$ auftritt. Dies sei unsere Nullhypothese, die wir mit einem Signifikanzniveau von $\alpha = 0{,}05$ testen wollen.
Beispiel:

i	1	2	3	4	5	6	7	8	9	10
h_i	10102	10070	9972	9803	10002	10065	10133	9943	10009	9901

Für den Wert der Testgröße gilt $T = 8{,}9946$. Ferner erhalten wir $\chi^2_{9,0,95} \approx 16{,}919$. Der Test liefert also keinen Grund, die Nullhypothese abzulehnen.

Das Prinzip des χ^2-Anpassungstests kann in leicht abgewandelter Form auch noch zum Testen einiger anderer Hypothesen verwendet werden: Beim χ^2-Homogenitätstest wird überprüft, ob zwei oder mehrere Verteilungen identisch sind, während beim χ^2-Unabhängigkeitstest zwei Zufallsgrößen auf Unabhängigkeit untersucht werden. Beschreibungen dieser Tests findet man in der Literatur.