Signifikanztest mit der Vierfeldertafel

Timm Grams, Fulda, 31. Juli 2010 (rev. 10.08.10)

Inhalt

Die Vierfeldertafel in den alltäglichen Statistiken

Jeder kennt Meldungen über mehr oder weniger wundersame Zusammenhänge: Bei Vollmond gibt es mehr Zwillingsgeburten als sonst. Italiener sind kinderfreundlicher als Deutsche. Besonders die Frauen bevorzugen Geländewagen. Ehen werden überdurchschnittlich oft zwischen Partnern desselben Sternbilds geschlossen. Solche Aussagen werden meist noch mit dem Hinweis auf Statistiken belegt. Und das naive Vertrauen in solche Zahlenwerke gibt einem dann den Rest: Man fällt auf die Sache herein.

Wir konzentrieren uns hier auf zwei Fragen: Wie *groß* ist ein in einer Meldung hochgejubelter Zusammenhang tatsächlich? Wird dieser Zusammenhang in der Statistik auch *deutlich* genug sichtbar?

Wir nehmen ein Beispiel aus der Aral-Studie „Trends beim Autokauf 2009“: 4% der Männer und 8% der Frauen würden als nächsten Wagen einen Ford kaufen. Befragt worden sind 301 Personen. Nehmen wir einmal an, dass es gleich viele Männer wie Frauen waren. Wir rechnen mit 150 befragten Frauen und ebenfalls 150 befragten Männern. Daraus lässt sich zurück schließen auf die zu Grunde liegende Statistik, die wir als *Vierfeldertafel* mit den Randsummen darstellen:

Tabelle 1 Die Automarke Ford bei Frauen besonders beliebt?

	Auto		
Geschlecht	Ford	andere	Zeilensumme
Frau	12	138	150
Mann	6	144	150
Spaltensumme	18	282	300

Wir haben es mit dem Merkmal Ford (ja oder nein) zu tun und mit zwei Statistiken, einer für die Frauen und einer für die Männer. Natürlich kann man auch so tun, als sei es nur eine Statistik, aufgeschlüsselt nach zwei Merkmalen: Automarke und Geschlecht.

Die Frage ist, was die Statistik tatsächlich über die *Größe* und *Deutlichkeit* der erhöhten Vorliebe der Frauen für die Automarke Ford aussagt. Könnte es sein, dass sowohl Männer als auch Frauen mit jeweils derselben 6-prozentigen Wahrscheinlichkeit Ford bevorzugen und dass die Stichprobe nur rein zufällig die Werte 12 und 6 der Tabelle anstelle von jeweils 9 produziert hat?

Unabhängigkeit zweier Merkmale

Die Nullhypothese

Für die weitere Analyse gehen wir von der generellen Vierfeldertafel der Tabelle 2 aus[1]. Sie dient der Auswertung einer Stichprobe von Objekten, denen gewisse Merkmale zukommen, oder auch nicht. Die Variable *N* bezeichnet den Umfang der Stichprobe. Wir beschränken uns auf die Untersuchung zweier Merkmale und benennen die Häufigkeiten, mit denen die Merkmalskombinationen auftreten mit *a, b, c* und *d*.

Falls die Tabelle das Ergebnis von zwei Stichproben ist und falls nur ein Merkmal erfasst worden ist, fassen wir die zwei Stichproben zu einer Gesamtstichprobe zusammen und nehmen die Zugehörigkeit zu den jeweiligen Stichproben als zweites Merkmal.

Tabelle 2 Generelle Vierfeldertafel

	Merkmal 1 liegt vor		
Merkmal 2 liegt vor	ja	nein	Zeilensumme
ja	*a*	*b*	*a+b*
nein	*c*	*d*	*c+d*
Spaltensumme	*a+c*	*b+d*	*a+b+c+d*

Wir fragen, ob die beiden Merkmale unabhängig voneinander auftreten oder ob wir davon ausgehen müssen, dass ein Zusammenhang zwischen ihnen besteht. Wir formulieren dazu die so genannte

> *Nullhypothese*: Die relative Häufigkeit des Vorliegens eines Merkmals ist unabhängig davon, ob gleichzeitig auch das andere Merkmal vorliegt. Abweichungen von dieser Annahme sind zufallsbedingt.

Die Hypothese besagt, dass das Merkmal 1 unter allen Objekten mit Merkmal 2 – abgesehen von den Zufälligkeiten der Stichprobenentnahme – relativ genau so häufig vorkommt, wie unter den Objekten, denen das Merkmal 2 fehlt. Im Idealfall gilt $a/b = c/d$ und damit auch $a/(a+b) = c/(c+d) = (a+c)/(a+b+c+d)$.

Gilt die Nullhypothese, lassen sich offenbar die Werte der Vierfeldertafel aus den Randsummen ermitteln. Beispielsweise ist $a = (a+b)(a+c)/(a+b+c+d)$

Wegen der Zufälligkeit der Stichprobenbildung können wir nicht erwarten, dass diese Gleichung exakt gilt. Wir bilden aus den Randwerten einen Schätzwert $\tilde{a}$ für a:

$$\tilde{a} = (a+b)(a+c)/N$$

Die Größe der Gesamtstichprobe N ist dabei gegeben durch

$$N = a+b+c+d.$$

Für die Abweichung des Wertes der Vierfeldertafel von diesem Schätzwerte schreiben wir Δ.

---

[1] Prüfverfahren für die Vierfeldertafeln werden in den Statistik-Lehrbüchern behandelt. Meine Darstellung folgt dem Buch „Angewandte Statistik“ von Lothar Sachs (Springer-Verlag, Berlin, Heidelberg 1992, S. 449 ff.) und beschränkt sich konsequent auf die Anwendung der elementaren Wahrscheinlichkeitsrechnung. Alles wird hergeleitet. Nur beim zentralen Grenzwertsatz wird auf Herleitungen verzichtet. Die dafür erforderliche Ausrüstung geht deutlich über die Elementarmathematik hinaus. In diesem Fall bleibt der Alltagsmathematik nur, anhand von Beispielen Vertrauen zu schaffen.

Es ist

$$\Delta = a - \tilde{a} = \frac{ad - bc}{N}.$$

Für die Werte *b*, *c* und *d* ergeben sich betragsmäßig, also abgesehen vom Vorzeichen, dieselben Differenzen.

Die zentrale Frage lautet, ob sich die Differenz allein auf den Zufall zurückführen lässt. Falls das nicht der Fall ist, wenn also die betragsmäßige Abweichung $|\Delta|$ so groß ist, dass sie unter der Nullhypothese nahezu ausgeschlossen werden kann, dann wird die Nullhypothese verworfen und wir können davon ausgehen, dass die Merkmale statistisch voneinander abhängig sind.

Nehmen wir den Fall der Tabelle 1 (Beliebtheit der Automarke Ford): Für die Zahl der Frauen, die einen Ford kaufen würden, ergibt sich der Schätzwert 9. Die Abweichung des Umfrageergebnisses vom Schätzwert beträgt also 12-9 = 3.

Nun ist noch die Frage zu beantworten, ob diese Differenz noch dem Zufall angelastet werden kann, oder ob womöglich eine erhöhte Vorliebe der Frauen für Ford vorliegt.

Wie viele Stichproben gibt es überhaupt?

Bei der Stichprobenbildung wird aus einer Grundgesamtheit – beispielsweise aller Deutschen, die sich demnächst ein Auto kaufen wollen – eine Teilmenge als Stichprobe ausgewählt. Zunächst wollen wir errechnen, wie viele solche Stichproben es überhaupt gibt, genauer, wie groß ist die Anzahl aller *n*-elementigen Teilmengen einer *N*-elementigen Grundgesamtheit.

In einer Tabelle tragen wir von links nach rechts den Umfang der Teilmenge an und von oben nach unten den Umfang der Grundgesamtheit. In der Spalte *n* der Zeile *N* steht also die Anzahl aller möglichen Stichproben vom Umfang *n* aus einer Grundgesamtheit vom Umfang *N*. Es ergibt sich das *Pascalsche Dreieck,* dessen Anfang die Tabelle 3 zeigt.

Tabelle 3 Pascalsches Dreieck (Anfang)

N \ n	0	1	2	3	4	5	6	7	8	9
0	1									
1	1	1								
2	1	2	1							
3	1	3	3	1						
4	1	4	6	4	1					
5	1	5	10	10	5	1				
6	1	6	15	20	15	6	1			
7	1	7	21	35	35	21	7	1		
8	1	8	28	56	70	56	28	8	1	
9	1	9	36	84	126	126	84	36	9	1

Das Bildungsgesetz ist einfach: Man beginnt mit einer Grundgesamtheit von 0 und trägt die Anzahl der Teilmengen ein: Die einzige Menge vom Umfang – oder der Mächtigkeit – null ist die leere Menge. Und sie hat auch nur die leere Menge als Teilmenge. Damit erhalten wir in Zeile 0 und Spalte 0 den Wert 1.

Jede der anderen Zahlen der Tabelle ergibt sich, indem man die darüber stehende mit der links daneben addiert. Leere Felder haben den Wert null.

Wie kommt es zu diesem Bildungsgesetz für die Tabelle? Nun: Gegenüber der vorherigen Zeile hat sich die Grundmenge um ein Element vergrößert. Die Anzahl aller Teilmengen des

gegebenen Umfangs, in denen dieses letzte Element nicht enthalten ist, steht genau darüber, denn die Elemente müssen alle aus der um ein Element verringerten Gesamtheit ausgewählt werden. Und die Anzahl aller Teilmengen, in denen das letzte Element enthalten ist, steht links neben der eben gefundenen Zahl, denn jetzt ist zur Auffüllung der Auswahl ein Element weniger aus der um ein Element verringerten Grundgesamtheit auszuwählen.

Aus dem Bildungsgesetz folgt sofort, dass sich die Zeilensummen von Zeile zu Zeile verdoppeln. Denn jede der Zahlen einer Zeile wird in der Zeile darunter genau zweimal berücksichtigt.

Die Zahlen des Pascalschen Dreiecks heißen *Binomialkoeffizienten*. Der Binomialkoeffizient aus Zeile N und Spalte n wird üblicherweise mit $\binom{N}{n}$ bezeichnet und „N über n“ gelesen.

Zur Namensgebung: Der Name der Binomialkoeffizienten kommt daher, weil sie als Beiwerte (Koeffizienten) beim Auswerten der Potenzen von Binomen auftreten. Hier gelten nämlich dieselben Gesetzmäßigkeiten, die uns beim Aufbau der Tabelle geleitet haben. Nehmen wir als Beispiel den Ausdruck $(a+b)^3$. Er ist äquivalent zu $1a^0b^3+3a^1b^2+3a^2b^1+1a^3b^0$. Die Beiwerte sind die Binomialkoeffizenten zu N=3.

Wir halten fest: Die Anzahl der n-elementigen Teilmengen einer N-elementigen Menge ist gleich N über n. Oder auch so: Die Anzahl der Möglichkeiten, eine Auswahl von n Elementen aus einer Menge mit N Elementen zu treffen, ist gleich N über n.

Die fundamentale Relation für Binomialkoeffizienten $\binom{N}{n} = \binom{N-1}{n-1} + \binom{N-1}{n}$ gilt für beliebige natürliche Zahlen N und n, wenn wir uns an folgende Übereinkunft halten[2]: Ein Binomialkoeffizient N über n wird gleich null gesetzt, wenn $N < n$.

Die Auswertung von Vierfeldertafeln mittels Kombinatorik

Wahrscheinlichkeiten für Merkmalshäufigkeiten

Jetzt tun wir so, als sei die Statistik der Vierfeldertafel alles, was wir kriegen können. Alle erfassten Elemente – also unsere Stichprobe – bilden gleichzeitig die Grundgesamtheit. Wir teilen diese Stichprobe in zwei Teilstichproben auf: Die 1. Teilstichprobe umfasst alle Objekte mit dem Merkmal 2. In der zweiten sind die restlichen Objekte, also diejenigen, denen das Merkmal 2 fehlt.

Wir gehen von der Gesamtstichprobe mit $N = a+b+c+d$ Objekten aus, von denen $M = a+c$ das Merkmal 1 haben. Die erste Teilstichprobe hat den Umfang $n = a+b$. Und die Nullhypothese besagt, dass diese Objekte rein zufällig aus der Gesamtstichprobe gewählt sind. Wir wollen nun festhalten, dass unter diesen n ausgewählten Objekten genau k das Merkmal 1 tragen (Trefferzahl) und dann die Wahrscheinlichkeit bestimmen, mit der es zu dieser Auswahl kommt.

Dazu bilden wir alle möglichen Fälle, in denen genau k der Merkmal-1-Objekte in der Auswahl landen. Es gibt $\binom{M}{k}$ Möglichkeiten, genau k Merkmal-1-Objekte aus der Gesamtmenge auszuwählen. Zu jeder dieser Möglichkeiten gibt es $\binom{N-M}{n-k}$ Möglichkeiten, die Teilstich-

[2] Wir nennen eine ganze Zahl n natürlich, wenn sie größer als null ist. Also: Hier zählen wir die Null nicht zu den natürlichen Zahlen.

probe mit Objekten aufzufüllen, denen das Merkmal-1 fehlt. Das Produkt dieser beiden Zahlen ist noch durch die Anzahl sämtlicher Möglichkeiten der Auswahl von n Elementen aus der Gesamtheit, nämlich durch $\binom{N}{n}$, zu teilen.

Die Wahrscheinlichkeit p_k, bei der Stichprobe vom Umfang n auf genau k Elemente mit dem Merkmal 1 zu treffen, ist demzufolge gleich

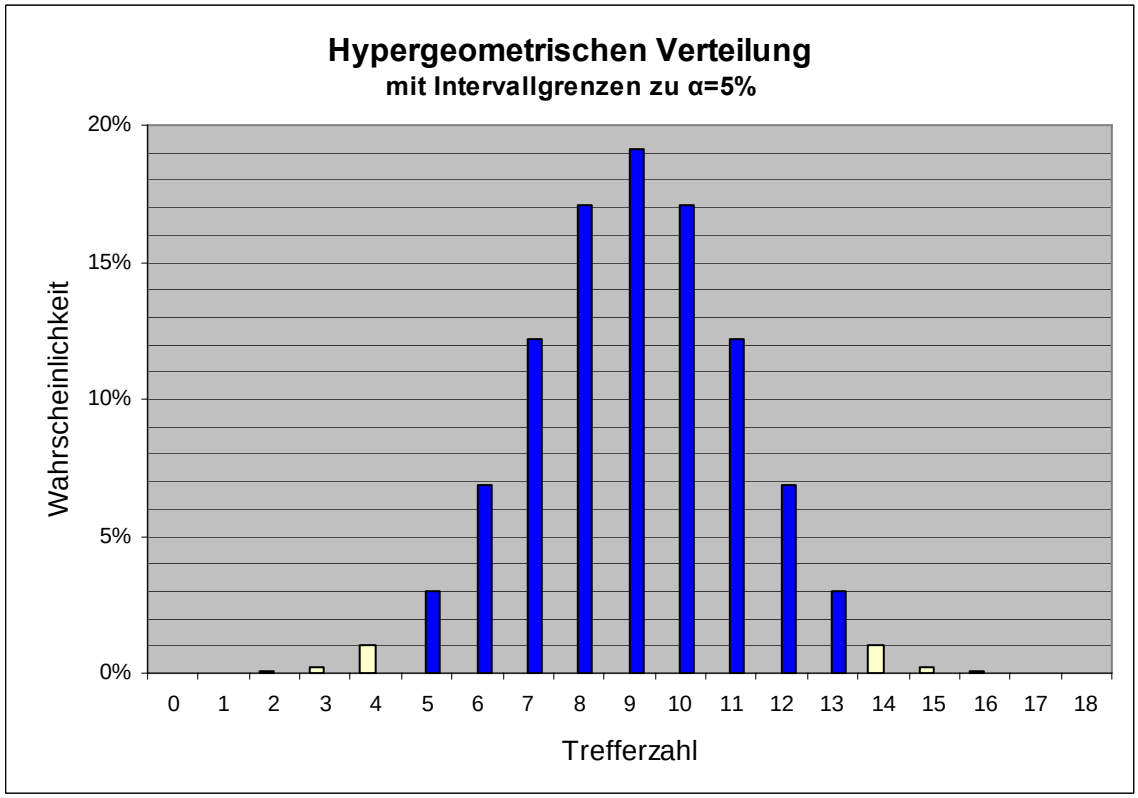
$$p_k = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}},$$

mit $k = 0, 1, 2, \ldots, n$. Das ist die Wahrscheinlichkeitsverteilung der *hypergeometrischen Verteilung.*

Mit einem Tabellenkalkulationsprogramm wie Excel verschaffen wir uns eine Darstellung dieser Verteilung: Als Beispiel dient wieder die Vorliebe der Frauen für die Automarke Ford. Wir stellen fest, dass das Maximum der Verteilung beim Schätzwert $\tilde{a} = (a+b)(a+c)/N$ liegt und dass links und rechts davon die Wahrscheinlichkeiten *glockenförmig* abfallen.

Die wahrscheinlicheren Werte liegen in der Nähe des Maximums der Glockenkurve. Die an den Rändern der Glockenkurve liegenden Werte treten seltener auf. Die geringeren Wahrscheinlichkeiten sind in der Grafik als helle Balken dargestellt.

Wenn die Nullhypothese gilt, sollte der tatsächliche Wert $k = a$ zu den wahrscheinlichen gehören. Oder andersherum ausgedrückt: Wenn der tatsächliche Wert zu den unwahrscheinlichen gehört, wollen wir die Nullhypothese ablehnen und sagen, dass die Abweichung des Wertes vom Schätzwert deutlich – oder wissenschaftlich ausgedrückt: signifikant – ist und nicht mehr dem Zufall allein zugeschrieben werden sollte.



Präzisierung des Tests

Wir teilen den Wertebereich in einen inneren und einen äußeren Teil auf. Der innere Teil enthält die wahrscheinlichen und der äußere die an den Rändern der Glockenkurve liegenden unwahrscheinlichen Werte. In der obigen Grafik sind die Wertebereiche durch unterschiedlich gefärbte Balken markiert: dunkel die inneren, hell die äußeren.

Den ersten und den letzten Wert des inneren Bereichs bezeichnen wir als unteren und oberen *Grenzwert.* Diese Grenzwerte werden folgendermaßen bestimmt: Zunächst wird eine Wahrscheinlichkeit α festgelegt. Die Wahrscheinlichkeit dafür, dass ein Wert in äußeren Bereich fällt, soll höchstens gleich dieser Wahrscheinlichkeit α sein.

Die untere Grenze u wollen wir so festsetzen, dass die Summe der Wahrscheinlichkeiten aller Werte unterhalb dieser Grenze kleiner als $\alpha/2$ ist, und dass durch Hinzunahme des unteren Grenzwerts diese Schranke erreicht bzw. überschritten wird. Und dasselbe fordern wir für die obere Grenze o: Die Wahrscheinlichkeit, dass ein Wert oberhalb dieser Schranke liegt, ist kleiner als $\alpha/2$. Erst zusammen mit dem oberen Grenzwert wird diese Bedingung nicht mehr eingehalten.

> Unter der Nullhypothese, also wenn die Unabhängigkeitsannahme gilt, ist die Wahrscheinlichkeit für Grenzwertüberschreitung höchstens gleich α. Anders herum ausgedrückt: Unter der Unabhängigkeitsannahme liegt der Wert a mit der Wahrscheinlichkeit von wenigstens $1-\alpha$ im Intervall von u bis o, also: $u \le a \le o$.

Die äußeren Werte bilden den *Ablehnungsbereich*. Jetzt können wir die Regel für den Test knapp formulieren:

> Liegt der Wert a im Ablehnungsbereich, dann wird die Nullhypothese mit einer Irrtumswahrscheinlichkeit von α abgelehnt und die Abweichung vom Schätzwert wird als signifikant auf dem Niveau α eingestuft. Andernfalls wird die Nullhypothese beibehalten.

Für $\alpha = 5\%$ und wenn der Wert im Ablehnungsbereich liegt, nennen wir die Abweichungen vom Schätzwert „leicht signifikant“. Für $\alpha = 1\%$ ist sie „signifikant“ und für $\alpha = 1‰$ „hoch signifikant“. In der obigen Grafik ist der Ablehnungsbereich für das Signifikanzniveau von 5% durch die helleren Balken markiert.

Übung 1. Überprüfen Sie, ob die Vorliebe der Frauen für die Automarke Ford signifikant ist.

Ein Experiment zu relativen Häufigkeiten und Wahrscheinlichkeiten

Für den Rest der Lektion will ich die Zurückhaltung bezüglich des Einsatzes mathematischer Symbolik aufgeben[3]. Ich mache sonst mir und Ihnen das Leben zu schwer. Also: Nur Mut!

Zur Vorbereitung wird ein kleines Experiment gemacht. Es dient der Erkundung dessen, was wir unter einer Wahrscheinlichkeit verstehen wollen. Jeder Teilnehmer bekommt fünf Halmasteine in die Hand gedrückt, drei rote und zwei grüne. (Mit anderen Farben geht das natürlich auch.) Dann soll er die Steine durch seine Hände einschließen, als wolle er einen Schneeball

---

[3] Für das vertiefte Studium sollte man auf ein gutes Grundlagenbuch zurückgreifen. Seit Jahrzehnten ist mir das Buch „Wahrscheinlichkeitsrechnung und mathematische Statistik“ von Marek Fisz (Berlin 1976) ein verlässlicher Ratgeber. Auch der aktuelle Duden „Rechnen und Mathematik“ bietet gute Hilfe. (Meine Ausgabe ist aus den Jahr 2000.) Wer die Herleitungen selbst durchführen möchte, sollte sich mit den Eigenschaften der Binomialkoeffizienten vertraut machen und insbesondere deren Darstellung mittels Fakultät $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ kennen und nutzen können.

formen, und die Steine gut durchschütteln. Blind wählt er sich dann zwei Steine aus und stellt danach fest, wie viele davon rot sind: entweder keiner, einer oder alle zwei. Die Zahl der roten Steine ist das Stichprobenergebnis *a*. Für jede solche Stichprobe kann man sich eine Vierfeldertafel vorstellen:

	Merkmal		
	rot	grün	Zeilensumme
Auswahl	*a*	*b*	*a*+*b* = 2
Rest	*c*	*d*	*c*+*d* = 3
Spaltensumme	*a*+*c* = 3	*b*+*d* = 2	*a*+*b*+*c*+*d* = 5

Nun verschafft sich jeder Teilnehmer durch wiederholte Versuche 20 Stichprobenergebnisse und schreibt auf, in wie vielen der Stichproben er keinen roten Stein, einen roten Stein oder gar zwei rote Steine gefunden hat und erhält so die *absoluten Häufigkeiten* H_0, H_1 und H_2. Er dividiert diese Zahlen jeweils durch die Anzahl der Versuche, also durch 20, und kommt so zu den *relativen Häufigkeiten* h_0, h_1 und h_2. Der arithmetische Mittelwert *m* für die Anzahl der roten Steine je Versuch ist offenbar gleich

$$m = 0\times h_0 + 1\times h_1 + 2\times h_2.$$

Ein Vergleich dieses arithmetischen Mittelwertes mit dem Schätzwert sollte zu der Aussage führen

$$m \approx \tilde{a},$$

wobei $\tilde{a}$ durch $(a+b)(a+c)/N = 2\times 3/5 = 1{,}2$ gegeben ist.

Dann werden die Versuche aller Teilnehmer zusammengefasst: Jeder Teilnehmer meldet seine Werte für die absoluten Häufigkeiten. Diese Zahlen werden durch die Gesamtzahl der Versuche geteilt und wir erhalten wieder die relativen Häufigkeiten und daraus einen neuen arithmetischen Mittelwert *m*. Es sollte jetzt intuitiv klar werden, dass durch die Vergrößerung der Versuchsanzahl der Wert $\tilde{a}$ immer verlässlicher und besser durch *m* angenähert wird.

Die relativen Häufigkeiten sind in der Regel gute Annäherungen an die Wahrscheinlichkeiten und es sollte klar werden, dass sich der Schätzwert $\tilde{a}$ auch über die Formel $0\times p_0 + 1\times p_1 + 2\times p_2$ errechnen lässt.

Übung 2. Berechnen Sie die Wahrscheinlichkeiten p_0, p_1 und p_2 nach den Formeln des vorherigen Abschnitts. Erstellen Sie ein Diagramm der Wahrscheinlichkeitsverteilung als Freihandskizze. Vergleichen Sie die Wahrscheinlichkeiten mit den relativen Häufigkeiten h_0, h_1 und h_2. Ermitteln Sie den Schätzwert $\tilde{a}$ aus den Wahrscheinlichkeiten.

Erwartungswert (Mittelwert) und Streuung (Varianz)

Wir haben es hier durchweg mit *diskreten Zufallsvariablen* zu tun. Eine solche Zufallsvariable *X* kann Werte aus einer endlichen Wertemenge annehmen. Diese Werte wollen wir nummerieren: x_i (i = 1, 2, 3, ..., n). Mit p_i bezeichnen wir die *Wahrscheinlichkeit* dafür, dass $X = x_i$ ist. Der *Erwartungswert* (*Mittelwert*) einer diskreten Zufallsvariablen *X* ist gegeben durch

$$\mu = E[X] = \sum_{i=1}^{n} x_i\, p_i\,.$$

Sei *g* eine reelle Funktion und *X* eine Zufallsvariable, dann definiert $g(X)$ ebenfalls eine Zufallsvariable. Ihr Erwartungswert ist gegeben durch

$$E[g(X)] = \sum_{i=1}^{n} g(x_i)\, p_i\,.$$

Unter der *Streuung* oder *Varianz* σ^2 einer Zufallsvariablen X wird der Erwartungswert der quadratischen Abweichung vom Mittelwert verstanden:

$$\sigma^2 = E[(X - \mu)^2]$$

Mit dem bisher ausgebreiteten Formelwerk lässt sich leicht herleiten, dass sich die Streuung auch so ausdrücken lässt

$$\sigma^2 = E[X^2] - \mu^2.$$

Die Wurzel aus der Streuung ist die *Standardabweichung* σ. Mit ihr lässt sich gut darstellen, auf welche Bereiche sich die Werte der Zufallsvariablen konzentrieren: Der Abstand der Realisierungen einer Zufallszahlen liegen mit hoher Wahrscheinlichkeit in einem nicht zu großen Abstand vom Mittelwert. Diese Aussage lässt sich relativ einfach präzisieren. Und weil es der Veranschaulichung der Bedeutung der Parameter μ und σ dient, sei die Herleitung einmal ausgeführt[4].

Die Streuung der diskreten Zufallsvariablen ist definiert durch

$$\sigma^2 = E[(X - \mu)^2] = \sum_{i=1}^{n} (x_i - \mu)^2 p_i .$$

Wir teilen die Summe in zwei Teilsummen auf. In der ersten sind nur die Summanden anzutreffen, für die $(x_i-\mu)^2 \leq 9\sigma^2$ gilt. In der zweiten Teilsumme sind die übrigen Summanden zu finden, also jene, für die bzw. $9\sigma^2 < (x_i-\mu)^2$ bzw. $3\sigma < |x_i-\mu|$ gilt. Wir lassen die erste Teilsumme weg und ersetzen in der zweiten die Werte $(x_i-\mu)^2$ jeweils durch $9\sigma^2$. Damit erhalten wir diese Abschätzung

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p_i > 9\sigma^2 \sum p_i$$

Bei dem zweiten Summenzeichen habe ich den Indexbereich einmal weggelassen, um die Sache nicht mit Formalitäten zu überlasten. Nach dem Bisherigen ist klar, dass nur über die Indizes i summiert wird, für die $3\sigma < |x_i-\mu|$ gilt. Diese Summe ist ganz offensichtlich gleich der Wahrscheinlichkeit, dass die Zufallsvariable $|X-\mu|$ größer als 3σ ist. Die Ungleichung liefert diese Erkenntnis: Die Wahrscheinlichkeit, dass eine Zufallsvariable um mehr als 3σ vom Mittelwert abweicht, ist kleiner als 1/9. Also haben wir diese

> *Veranschaulichung:* In wenigstens 89% der Fälle liegen die Realisierungen der Zufallszahlen im Intervall von μ-3σ bis μ+3σ, und in weniger als 11% der Fälle außerhalb.

Diese Aussage gilt ganz allgemein, also unabhängig von der konkreten Verteilung der Zufallsvariablen. In vielen praxisrelevanten Fällen hat man eine noch stärkere Konzentration der Zufallsvariablen auf den Mittelwert.

Das Testkriterium

Wir setzen jetzt die Gültigkeit der Nullhypothese voraus. Die tatsächlich ermittelte Größe a weicht von ihrem Erwartungswert μ nicht allzu stark ab. Die Abweichung liegt in den allermeisten Fällen im Bereich von nur wenigen Vielfachen der Standardabweichung σ. Als Prüfgröße bietet sich der Quotient $|\Delta|/\sigma = |a - \mu|/\sigma$ an. Große Werte der Prüfgröße sind unwahrscheinlich. Das wollen wir jetzt weiter präzisieren.

---

[4] Es handelt sich um einen Spezialfall der Tschebyscheffschen Ungleichung.

Wir führen einen Grenzwert G für den Quotient $|\Delta|/\sigma$ ein und fordern, dass dieser Grenzwert G bei geltender Nullhypothese nur in seltenen Fällen überschritten wird. Wir setzen zunächst fest, was wir unter den „seltenen Fällen“ verstehen wollen: Die Zahl α bezeichne eine gewählte kleine Wahrscheinlichkeit, üblicherweise 5%, 1% oder 1‰. Den Grenzwert $G = G(\alpha)$ wählen wir so, dass die folgende Aussage gilt:

> Unter der Nullhypothese, also wenn die Unabhängigkeitsannahme gilt, ist die Wahrscheinlichkeit für Grenzwertüberschreitung, d.h. $G < |\Delta|/\sigma$, höchstens gleich α. Andersherum ausgedrückt: $|\Delta|/\sigma \le G$ gilt mit der Wahrscheinlichkeit 1-α.

Bei Überschreitung des Grenzwerts wird die Nullhypothese mit einer Irrtumswahrscheinlichkeit von α abgelehnt. Man spricht dann auch von einer deutlichen (signifikanten) Abweichung auf dem Signifikanzniveau von α. Für $\alpha = 5\%$ nennen wir eine den Grenzwert überschreitende Abweichung „leicht signifikant“, für $\alpha = 1\%$ „signifikant“ und für $\alpha = 1‰$ „hoch signifikant“.

Was wir noch benötigen, ist die Funktion G, und die ist für jede Verteilung getrennt zu bestimmen. Glücklicherweise aber ähneln sich in der Praxis die Verteilungen sehr stark. Sie haben meist den glockenförmigen Verlauf wie in unserem Beispiel. Das ist eine Folge des so genannten *zentralen Grenzwertsatzes* der Wahrscheinlichkeitsrechnung[5]. Für alle diese Verteilungen erhalten wir näherungsweise dieselbe Funktion G.

Aus Gründen der Praktikabilität nehmen wir nicht den Quotienten $|\Delta|/\sigma$ sondern dessen Quadrat Δ^2/σ^2 als Prüfgröße. Die folgende Tabelle zeigt die zu den drei üblicherweise gewählten Signifikanzniveaus gehörenden Grenzwerte $G^2(\alpha)$.

Tabelle 4 Grenzwerte für die Prüfgröße

Sprechweise	leicht signifikant	signifikant	hoch signifikant
α	5%	1%	1‰
$G^2(\alpha)$	3,841	6,635	10,828

Und so sieht das *Testkriterium* jetzt aus:

> Überschreitet die Prüfgröße Δ^2/σ^2 den Grenzwert $G^2(\alpha)$, dann wird die Nullhypothese bei einer Irrtumswahrscheinlichkeit von α abgelehnt und die Differenz Δ als signifikant auf dem Niveau α eingestuft. Ansonsten wird die Nullhypothese beibehalten.

Vierfeldertafel-Test

Erwartungswert und die Streuung der hypergeometrischen Verteilung sind[6]

$$\mu = E[X] = n\frac{M}{N} = \frac{(a+b)(a+c)}{N}$$

und

$$\sigma^2 = \frac{nM(N-M)(N-n)}{N^2(N-1)} = \frac{(a+b)(a+c)(b+d)(c+d)}{N^2(N-1)}.$$

---

[5] Siehe zum Beispiel „Wahrscheinlichkeitsrechnung und mathematische Statistik“ von Marek Fisz (DVW, Berlin 1976, S. 237)

[6] Wer sich etwas Übung im Umgang mit kombinatorischen Formeln zulegen will, kann sich dies Formeln selbst herleiten. Das ist zwar etwas langwierig, aber es ist nicht schwer: HypergeometrischeVerteilung.pdf.

Der Erwartungswert μ ist gleich unserem bereits früher ermitteltem Schätzwert $\tilde{a}$. Es ist also $\tilde{a} = \mu$

Für die Differenz erhalten wir die Formel $\Delta = a - \tilde{a} = \frac{ad - bc}{N}$. Wir bilden die Prüfgröße Δ^2/σ^2 und erhalten

$$\Delta^2/\sigma^2 = \frac{(N-1)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Diese Prüfgröße wird sehr oft für die Unabhängigkeitsprüfung mittels Vierfeldertafel verwendet. Das *Testkriterium* sieht jetzt folgendermaßen aus:

> Überschreitet die Prüfgröße $\frac{(N-1)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$ den Grenzwert $G^2(\alpha)$, dann wird die Nullhypothese bei einer Irrtumswahrscheinlichkeit von α abgelehnt und die Differenz Δ als signifikant auf dem Niveau α eingestuft. Ansonsten wird die Nullhypothese beibehalten.

Übung 3. Überprüfen Sie anhand des Testkriteriums die Vorliebe der Frauen für die Automarke Ford.

Stichprobenumfang und Einflussfaktoren sind relevant

Ob eine Differenz *groß* ist oder nicht, sagt einem am besten die relative Abweichung $\Delta/\tilde{a}$. Ob die Differenz *deutlich* ist, zeigt die Prüfgröße Δ^2/σ^2: Mit der Prüfgröße steigt die Signifikanz der Abweichung.

Wie hängen Größe und Deutlichkeit einer Abweichung von der Stichprobengröße *N* ab?

Zunächst machen wir uns klar, dass mit einer Vergrößerung von *N* auch die Größen *a, b, c, d* und damit auch *n* wachsen, und zwar in etwa proportional zu *N*, vorausgesetzt, die Statistik gibt die tatsächliche Verteilung der Merkmale halbwegs getreulich wieder.

Daraus folgt, dass die Differenz Δ, der Schätzwert $\tilde{a}$ und auch die die Streuung σ^2 in etwa proportional zu *N* wachsen. Aus diesem Grunde wächst auch die Prüfgröße Δ^2/σ^2 proportional zu *N*, wohingegen die relative Abweichung $\Delta/\tilde{a}$ in etwa gleich bleibt. Also: Der Stichprobenumfang hat auf die Größe einer Abweichung kaum Einfluss, wohingegen die Deutlichkeit mit dem Stichprobenumfang zunimmt.

Zwei Fälle verdienen besondere Aufmerksamkeit. Bei recht kleinen Stichproben ($N \leq 1000$) werden systematische Abweichungen nicht deutlich genug erkannt. Die meisten der in den Zeitungen veröffentlichten Statistiken fallen in diese Kategorie. Sie haben einen viel zu kleinen Stichprobenumfang, als dass interessante Effekte überhaupt erkennbar sein könnten. Meist werden irgendwelche Zufälligkeiten im Zusammenhang mit der Stichprobenentnahme überinterpretiert und zu Gesetzmäßigkeiten hochgejubelt. Beispiele dafür bieten die Aral-Studie und alle möglichen Meinungsumfragen.

Das andere Extrem ist die sehr kleine systematische Abweichung, die bei einer sehr großen Stichprobe ($N \approx 1$ Mio.) deutlich herauskommt, also hoch-signifikant ist. Dabei können Einflussfaktoren eine Bedeutung erlangen, die man eigentlich glaubte, vernachlässigen zu können. Beispiele dafür sind in der Astrologie-Studie von Gunter Sachs zu finden.

Jedenfalls das können wir den bisherigen Überlegungen entnehmen: Eine veröffentlichte Statistik ohne erkennbaren Stichprobenumfang und ohne kritische Würdigung von Einflussfakto-

ren ist wertlos. Werden uns dann noch irgendwelche Schlussfolgerungen aus dieser Statistik nahe gebracht, liegt der Verdacht eines Manipulationsversuchs nahe.