Stochastische Analyse von empirischen Zählvergleichen

Klaus Felsenstein
Institut für Statistik und
Wahrscheinlichkeitstheorie
technische universität wien

1 Problemstellung

Den Ausgangspunkt für Entscheidungen und Aussagen bilden häufig Zählungen. Die prinzipielle Vorgangsweise beim Vergleich von Zählungen aus methodischer Sicht soll hier vorgestellt werden. Es existieren statistische Verfahren, die auch bei kleiner Stichprobenzahl durchführbar sind. Dass statistische Testverfahren nur bei hoher Probenzahl aussagekräftige Aussagen liefern, gilt zwar für epidemiologische Untersuchungen bzw. für allgemein gültige Aussagen über die Grundgesamtheit. Für individuelle Vergleiche, die nur für die konkrete Fragestellung Erkenntnisse bringen sollen, gibt es statistische Tests, die unabhängig von der Größe der Stichprobe optimal sind und außerdem vergleichsweise einfach durchzuführen sind, wenn die Zählungen eher kleinere Werte ergeben.

Für Messungen existieren in der Stochastik verschiedene geläufige Vergleichsverfahren (wie t-Tests, Anova-Tabellen u.v.m.), die auf Grund des zentralen Grenzverteilungssatzes auch auf umfangreiche diskrete Zählungen angewandt werden können. Ein auf großer Anzahl basierendes Verfahren stößt in der Praxis oft auf die Schwierigkeit, gleich bleibende Voraussetzungen (Unabhängigkeit oder unveränderte Verteilungsannahmen) garantieren zu können. Hauptsächlich dann, wenn nur eine geringe Anzahl von Beobachtungen vorliegt, sollten Vergleichsverfahren angewandt werden, die die tatsächliche Verteilung des Zählprozesses ins Kalkül ziehen.

Die Basis für allgemeine Zählvorgänge bildet der Poissonprozess. Vergleicht man zwei Poisson-verteilte Zählungen, ergibt nicht die absolute Differenz eine optimale Aussage über den Unterschied der beiden Werte. Ein Umweg über den Anteil an der Summe beider Zahlen führt zu einer besseren Abgrenzung durch absolute Häufigkeiten. Diese Prozedur ist optimal im Sinne der Trennschärfe von statistischen Testprozeduren. Der Vergleich zweier einfacher Zählungen wird dann auf eine Methode zum Vergleich von Vektoren von Häufigkeiten erweitert. Das Verfahren ist im Aufbau ähnlich dem exakten Test nach Fisher für alternativ verteilte Merkmale, der zunächst vorgestellt wird.

2 Häufigkeiten in einer Grundgesamtheit

Ein Vergleichsverfahren für absolute Häufigkeiten bei mehreren Stichproben ist der exakte Test nach Fisher. Dafür werden die Verteilungen von Häufigkeiten mit identischer Klasseneinteilung als Versuchsergebnis mit Gesamtanzahl N dargestellt. Das Versuchsergebnis liegt dann dann in einer Aufteilung der Häufigkeiten auf verschiedenen Klassen als Histogramm mit absoluten Häufigkeiten vor.

Für den Vergleich mehrerer Häufigkeitszählungen bei identischer Klasseneinteilung wird eine *Kontingenztafel* erstellt. Das Merkmal X wird unter Klassen A_i und Y unter Klassen B_j gezählt.

X \ Y	$B_1 \quad B_2 \ldots B_l$	
A_1	$H_{11} \quad H_{12} \ldots H_{1l}$	$H_{1\bullet}$
$\vdots$	$\ldots\ldots\ldots$	$\vdots$
A_k	$H_{k1} \quad H_{k2} \ldots H_{kl}$	$H_{k\bullet}$
	$H_{\bullet 1} \quad H_{\bullet 2} \ldots H_{\bullet l}$	N

Die Randhäufigkeiten für $X \in A_i$ und $Y \in B_j$ sind

$$H_{i\bullet} = \sum_{j=1}^{l} H_{ij} \qquad\qquad H_{\bullet j} = \sum_{i=1}^{k} H_{ij} \ .$$

Die Summe aller Häufigkeiten bezeichnet N. Der einfachste solcher Häufigkeitsvergleiche wird in der Vierfeldertafel bei dichotomer Aufteilung auftreten. Bei zwei Gruppen wird eine Binomalverteilung verglichen. Die Zulassung medizinischer Behandlungen beruht meist auf einem solchen Entscheidungsverfahren, eine solche Vierfeldertafel sei beispielsweise

Vierfeldertafel:

Behandlung \ Resultat	Gesundheit verbessert	nicht verbessert	
Medikament	95	19	114
Placebo	27	11	38
	122	30	N = 152

Ein asymptotisches Verfahren für die Feststellung von Unterschieden ist der χ^2 -Test, wobei die Häufigkeiten H_{ij} mit dem Produkt der Randhäufigkeiten $H_{i\bullet}H_{\bullet j}$ verglichen werden. Liegt kein Effekt auf X durch Y vor, sollte näherungsweise

$$\frac{H_{ij}}{N} \simeq \frac{H_{i\bullet}}{N} \frac{H_{\bullet j}}{N}$$

gelten. Die Summe der normierten Quadratabstände ist dann approximativ χ^2 verteilt. Der Test ist also nur für große Stichproben mit identischer Klasseneinteilung brauchbar.

Für kleine Probenanzahl wird ein *exakter Test* auf die Wirksamkeit des Faktors Y angewandt. Als Prüfverteilung wird die bedingte Verteilung der Häufigkeiten unter den Randhäufigkeiten herangezogen. Sind die Randsummen $H_{i\bullet}$ und $H_{\bullet j}$ fest, folgen die Zellenwerte der Vierfeldertafel einer Hypergeometrischen Verteilung,

$$\mathbf{P}[H_{11} = K | H_{i\bullet}, H_{\bullet j}] \ = \ \frac{\binom{H_{1\bullet}}{K}\binom{H_{2\bullet}}{H_{21}}}{\binom{N}{H_{\bullet 1}}} \ .$$

Für die Prüfung der Hypothese, dass kein Effekt vorliegt, wird diese Hypergeometrische Wahrscheinlichkeit für das beobachtete K und alle anderen Werte J für $H_{11} = J$, die der Hypothese noch eher widersprechen als der beobachtete Wert K, berechnet und summiert. Unterschreitet diese Summe der Wahrscheinlichkeiten das Niveau des Tests (üblicherweise 5%), wird auf einen Effekt geschlossen. Wäre beispielsweise obige Kontingenztafel

Behandlung \ Resultat	Gesundheit verbessert	nicht verbessert	
Medikament	94	20	114
Placebo	28	10	38
	122	30	N = 152

ausgefallen, dann würde das einem Effekt des Medikaments eher widersprechen.
Das Prinzip des exakten Tests lässt sich auf größere Kontingenztafeln verallgemeinern. Die bedingte Verteilung unter festen Zeilen- und Spaltensummen wird bei mehr als zwei Klassen des Merkmals X zu einer *Poly-Hypergeometrischen* Verteilung. Die Prüfverteilung ist dann eine multivariate Verteilung und die Anzahl der Kombinationen von Häufigkeiten, die der Hypothese eher widersprechen, wird sehr schnell extrem hoch. Die Durchführbarkeit des Tests leidet darunter. Es wird also verbreitet der χ^2-Test ersatzweise angewandt, obwohl in vielen Programmpaketen der exakte Test enthalten ist.

Tatsächlich ist aber der exakte Test optimal. Man kann zeigen, dass es keinen anderen ausgewogenen (unverfälschten) Test mit besserer Operationscharakteristik gibt. Das bedeutet, die Wahrscheinlichkeit für einen Fehler 2. Art ist bei dieser Entscheidungsprozedur minimal. Dieser Fehler tritt ein, wenn ein vorhandener Effekt nicht erkannt wird.

In vielen Situationen wird aber verlangt, die Zählergebnisse direkt und ohne Zuordnung zu Klassen zu vergleichen. Bevor ein solcher Test entwickelt wird, ist eine grundsätzliche Analyse eines Zählvorgangs hilfreich.

3 Poissonprozess

Absolute Zählungen von Ereignissen folgen unter recht allgemeinen Bedingungen einer Poissonverteilung P_λ mit Auftrittsrate λ. Die Anzahl der Ereignisse in einem Zeitintervall $[0, t]$ sei ein Zählprozess N_t. Die Indexmenge kann natürlich eine andere als die Zeit sein, etwa eine Länge oder Fläche, an der an Punkten Ereignisse eintreten können. Die Indexmenge kann prinzipiell beliebig sein, solange sich die Additivitätseigenschaften der Zählung äquivalent zu jenen einer Zählung innerhalb von Zeiteinheiten verhalten.
Es seien folgende für eine allgemeine Zählung nahe liegende Anforderung erfüllt.

- $N_0 \equiv 0$, in einem Intervall ohne Inhalt kann auch kein Ereignis auftreten.
- Die Ergebnisse der Zählungen in disjunkten Intervallen sind unabhängig haben außerdem in gleichlangen Zeiträumen dieselbe Verteilung. Also die Zuwächse sind unabhängig und stationär.

- Für 'kleine' Zeiteinheiten ist die Wahrscheinlichkeit für ein Ereignis proportional zur Länge des Zeitintervalls:
$$\mathbf{P}[N_h = 1] \;=\; \lambda\ h \;+\; o(h)\ .$$
- Es wird für 'kleine' Zeiteinheiten sehr unwahrscheinlich, dass mehrere Ereignisse auftreten. Das wird als *Koinzidenzfreiheit* bezeichnet und durch die Bedingung
$$\mathbf{P}[N_h \geq 2] \;=\; o(h)$$
spezifiziert.

Diese Voraussetzungen lassen nur mehr eine Poissonverteilung für den Zählprozess mit Rate λ zu, also $N_t \sim P_{\lambda t}$ und daher sind auch die Zuwächse $N_{s,t} := N_t - N_s$ Poisson-verteilt mit Parameter $\lambda(t-s)$ für Zeitpunkte $t > s$.

Eine Zählung folgt also unter Homogenitäts- und Unabhängigkeitsannahmen einer Poissonverteilung, effektive Änderungen am Modell drücken sich dann in Unterschieden der zu Grunde liegenden Rate λ aus. Mittels eines Verfahrens soll die Frage beantwortet werden, ob die Ergebnisse zweier einfacher Zählungen in unabhängigen Abschnitten dieselbe Modellverteilung besitzen. Wenn X_1 das Resultat der Zählung im Intervall der Länge t_1 und unabhängig davon X_2 entsprechend im Intervall der Länge t_2 das Resultat ist, dann wird bei gleich bleibender Rate
$$X_1\ t_2 \;\simeq\; X_2\ t_1$$
verlangt. Bei gleichlangen Intervallen $t_1 = t_2$ ist der Abstand in einem gewissen Sinn ausschlaggebend, nur stellt nicht direkt die Verteilung von $|X_1 - X_2|$ eine geeignete Entscheidungsgrundlage dar.

Als praktisches Beispiel für eine Zählung unter obigen Bedingungen wählen wir Daten aus der Kriminalstatistik. Die Anzahl der Tötungsdelikte lag im Jahre 2008 in Österreich bei 101 und im Jahre 2009 bei 141. Es erscheint gerechtfertigt, von einer signifikanten Verschlechterung der Sicherheit auszugehen.

4 Test für Zählvergleiche

Als prinzipielle Möglichkeit, Mittelwerte zu vergleichen, wird oft der t-Test eingesetzt. Dabei wird der standardisierte Abstand der empirischen Durchschnittswerte mit dem Quantil einer t-Verteilung als kritischem Wert verglichen. Die Anwendung dieses gebräuchlichen Tests auf Zählungen wird bei großer Anzahl mit der Normalapproximation der Poissonverteilung gerechtfertigt, hat aber prinzipielle und praktische Nachteile. Die Trennung von Mittelwert und Varianz als voneinander unbeeinflusste Parameter, wie sie der Normalverteilung entspricht, versagt bei der Poissonverteilung, wo Mittelwert und Varianz ident mit der Rate sind. Die Voraussetzung für den t-Test sind identische Varianzen, was der zu prüfenden Hypothese entspricht. Insbesondere die notwendige Normalapproximation bringt es bei eher kleineren Zahlen mit sich, dass die Operationscharakteristik des t-Tests unbrauchbar wird. Zudem verlangt die Varianzschätzung aus einer einzelnen Beobachtung eine symbolische Aufteilung der Varianz.

Ein Testverfahren, das dem exakten Test von Fisher vergleichbar ist, besitzt die für diese Situation optimale Operationscharakteristik unter unverfälschten Tests. Auch hier wird unter der Summe der beiden Beobachtungen bedingt. Es sind $X = x$ und $Y = y$ zwei unabhängige Poissonverteilungen, für die die Hypothese identischer Raten $H_0 : \lambda_X = \lambda_Y$ bzw. einseitige Hypothesen wie $H_0' : \lambda_X \leq \lambda_Y$ geprüft werden sollen.
Die bedingte Wahrscheinlichkeit einer Beobachtung X unter der Summe $S = X + Y$ ist

$$\mathbf{P}[X = k|S = s] \;=\; \frac{\mathbf{P}[X = k \wedge S = s]}{\mathbf{P}[S = s]} \;=\; \frac{p(k;\lambda_X)\; p(s-k;\lambda_Y)}{p(s;\lambda_X + \lambda_Y)} \;,$$

wobei $p(k;\lambda)$ die Punktwahrscheinlichkeit einer Poissonverteilung an der Stelle k mit Rate λ bezeichnet. Unter Gültigkeit der Hypothese $H_0 : \lambda_X = \lambda_Y = \lambda$ und Berücksichtigung, dass die Summe Poisson-verteilt $S \sim P_{2\lambda}$ ist, folgt

$$\mathbf{P}[X = k|S = s] \;=\; \frac{s!}{k!\;(s-k)!}\left(\frac{1}{2}\right)^s .$$

Diese bedingte Verteilung entspricht demnach einer Binomialverteilung $B_{s,0.5}$. Unter der Nullhypothese sollten die beiden Beobachtungen X und Y im $1-\alpha$ Bereich des Sollwertes $s/2$ liegen. Der Abstand unter der Beobachtung von $X = x$ und $Y = y$ ist

$$|X - \frac{s}{2}| \;=\; \frac{1}{2}|x - y| \;,$$

und die Nullhypothese wird zum Niveau α verworfen, wenn die Wahrscheinlichkeit

$$\mathbf{P}[|X - \frac{s}{2}| \geq \frac{1}{2}|x-y||S = s] \;=\; \mathbf{P}[X \leq x|S = s] \;+\; \mathbf{P}[X \geq y|S = s] \;\geq\; \alpha$$

erfüllt. Der P-Wert dieses Tests ist die Summe der beiden letzten Wahrscheinlichkeiten, die mittels der Binomialverteilung $B_{s,0.5}$ berechnet werden. Unterschreitet dieser P-Wert die Irrtumswahrscheinlichkeit, wird die Annahme gleicher Poissonprozesse abgelehnt. Dieses Prüfverfahren ist sehr einfach durchzuführen. Trotzdem ist es optimal und trennt Nullhypothese und Gegenhypothese bestmöglich. Der Abstand zwischen den beiden Zählungen entscheidet über die Nullhypothese, allerdings mit der Gewichtung dieser Binomialverteilung.

Betrachtet man jetzt nochmals die Steigerung der Tötungsdelikte mit den Zählungen $X = 101$ und $Y = 141$, so ist die Summe der Binomialwahrscheinlichkeiten unter 101 und über 141 mit der Binomialverteilung $B_{242;0.5}$ gleich 0.0102. Der P-Wert ist also zwischen 1% und 5%, die Antwort, ob eine signifikante Veränderung gegeben ist, fällt nicht mehr so eindeutig aus. Bei 5% Irrtumswahrscheinlichkeit ist die Steigerung signifikant während bei 1% Irrtumswahrscheinlichkeit die höhere Anzahl an Tötungsdelikten noch nicht für den Beweis einer signifikanten Steigerung ausreicht.

Zur numerischen Berechnung der Binomialwahrscheinlichkeiten kann bei ausreichend großen Werten auch eine Normalapproximation der Binomialverteilung angewandt werden. Es wird dann

$$\mathbf{P}[X \leq x|S = s] \;\simeq\; \Phi\left(\frac{x + 0.5 - s}{\sqrt{s\;0.25}}\right)$$

näherungsweise eingesetzt. Ab $S \geq 150$ sollte dann der P-Wert durch

$$\mathbf{P}[X \leq x|S = s] \ + \ \mathbf{P}[X \geq y|S = s] \ \simeq \ 2\ \Phi\left(\frac{1 - |x - y|}{\sqrt{s}}\right)$$

approximierbar sein. Der approximierte P-Wert ist im obigen Beispiel gleich 0.012 mit identem Testergebnis.

5 Simultaner Vergleich

Für einen gleichzeitigen Vergleich von verschiedenen Zählungen wird das Prinzip des paarweisen Vergleichs verallgemeinert. Normalverteilte Messreihen werden mit den Methoden der Varianzanalyse behandelt. Wie beim paarweisen Vergleich birgt die Anwendung der *ANOVA* selbst bei großer Anzahl erhebliche Schwachstellen in sich. Kleine einzelne Zählungen können so nicht miteinander verglichen werden. Die Standardmethoden der Varianzanalyse beruhen auf identischen Varianzen der Messreihen, was bei Poissonverteilung der zu testenden Hypothese identischer Raten entspricht. Wenn nur einzelne Messungen (und nicht der Durchschnitt mehrerer Zählungen) einbezogen werden, kann keine empirische Varianz der Messreihe berechnet werden.

Gehen wir von k Zählungen $X_1, X_2, \ldots, X_k$ unter Poissonverteilung $X_i \sim P_{\lambda_i}$ aus, dann lautet die Hypothese, dass kein Effekt auf die Zählungen wirkt, $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_k$. Das Testprinzip lässt sich auf allgemeinere Hypothesen, beispielsweise $H_0 : \lambda_i = \lambda_j \geq \lambda_k$, abwandeln. In der Folge diskutieren wir aber ausschließlich die symmetrische, gleichmäßige Hypothese $H_0 : \lambda_i = \lambda, \forall i$.

Falls alle Zählungen von derselben Poissonverteilung stammen bzw. dieselbe Rate besitzen, dann sollte jeder Wert annähernd denselben Anteil haben. Ist die Summe $S = X_1 + \ldots + X_k$, dann wird erwartet, dass

$$X_i \ \simeq \ \frac{S}{k}$$

erfüllt ist. Die Abstände zwischen X_i und $\frac{S}{k}$ entscheiden über die Annahme. Unter dem 'Abstand' ist aber ein Maß gemäß einer Wahrscheinlichkeitsverteilung gemeint. Wie im Fall zweier Zählungen ist dieses Maß von der bedingten Verteilung der Zählung unter der Summe bestimmt. Die bedingte Einzelverteilung ist

$$\mathbf{P}[X_i = x_i|S = s] \ = \ \frac{(k-1)^{s-x_i}}{k^s}\binom{s}{x_i}$$

und die gemeinsame Verteilung aller Werte ist

$$\mathbf{P}[X_1 = x_1, \ldots, X_k = x_k|S = s] \ = \ \binom{s}{x_1, \ldots, x_k} \prod_{i=1}^{k} \left(\frac{1}{k}\right)^{x_i} \ = \ \binom{s}{x_1, \ldots, x_k} (\frac{1}{k})^s \ ,$$

mit dem Multinomialkoeffizient $\binom{s}{x_1,\ldots,x_k}$. Die bedingte Verteilung ist somit eine Multinomialverteilung $M_{s;(1/k,\ldots,1/k)}$ mit Gesamtanzahl s und Eintrittswahrscheinlichkeit $1/k$.

Die praktische Durchführung des Tests benötigt nur die Punktwahrscheinlichkeiten dieser Multinomialverteilung, die mit

$$p(x_1,\ldots,x_k|s,k) \;:=\; \mathbf{P}[X_1 = x_1,\ldots,X_k = x_k|S = s]$$

bezeichnet wird. Es seien die konkret beobachteten Abstände vom Sollwert $\frac{s}{k}$

$$d_i \;:=\; |x_i \;-\; \frac{s}{k}|\;.$$

Der P-Wert des Tests ist dann

$$\begin{aligned} &1 \;-\; \mathbf{P}[|X_1 - \frac{s}{k}| \le d_1,\ldots,|X_k - \frac{s}{k}| \le d_k|S = s] \\ =\;& 1 \;-\; \mathbf{P}[\frac{s}{k} - d_i \le X_i \le \frac{s}{k} + d_i;\; i = 1,\ldots,k|S = s]. \\ =\;& 1 \;-\; \sum_{m_1,\ldots,m_k} p(m_1,\ldots,m_k|s,k), \end{aligned}$$

wobei die Summe über die Indizes m_i im Bereich $\frac{s}{k} - d_i \le m_i \le \frac{s}{k} + d_i$ gebildet wird.

6 Anwendung und Veranschaulichung

Das Unterscheidungsverfahren für Zählungen lässt sich bei kleinerer Dimension und geringeren Zählungen durch die Bestimmung von gewichteten Abständen von Punkten durchführen. Angenommene Ergebnisse der Zählung, die vom Sollwert gleich weit oder weiter entfernt sind, werden zusammengefasst und das Gewicht dieser Menge bestimmt die konkrete Entscheidung. Dabei ist aber unter *Abstand* nicht prinzipiell der geometrische euklidische Abstand zu verstehen. Die Punktwahrscheinlichkeiten der Multinomialverteilung ergeben die Punkte mit Mindestabstand bezogen auf das Zählergebnis. Diese Punkte bilden eine Umgebung des Sollwerts. Alle Punkte, die außerhalb (im geometrischen Sinn) dieser Umgebung liegen und im ganzzahligen Raster sind, können auch keine größere Multinomialwahrscheinlichkeit besitzen. Für drei Zählwerte lässt sich die Durchführung des Tests mittels einer zweidimensionalen Graphik einfach illustrieren.

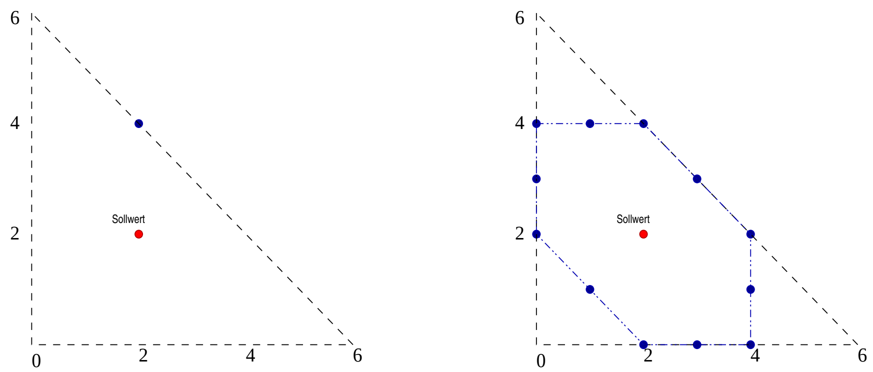
Als Beispiel wählen wir den Vergleich von Fehlerraten verschiedener Produktionen. Hier sind kleinere Zählwerte wünschenswert. Bei einer Prüfung der Produkte aus den Produktionen A, B, C ergeben sich aus gleichgroßen Stichproben bei

Produkt A: 2 Fehler B: 4 Fehler C: 0 Fehler .

Die Hypothese, dass die Fehlerraten der drei Produktionen gleich sind, soll mit dem exakten Poissonvergleichstest geprüft werden.

Die Summe ist daher $S = 6$ und alle ganzzahligen Ergebnisse mit Summe 6 bilden folgendes Dreieck. Als *Sollwert* bei perfekter identischer Verteilung ergibt sich der Punkt (2/2/2). Die Graphik zeigt die geometrische Umgebung des Sollwerts von Punkten, die denselben Abstand wie die Beobachtung (2/4/0) haben oder auf einer Verbindungsstrecke liegen. Da nur Punkte mit Summe 6 einbezogen werden, genügt natürlich die zweidimensionale Darstellung.

Figur 1: Sollwert, Beobachtung und Umgebungspunkte



Die Multinomialwahrscheinlichkeit unter gültiger Nullhypothese ist bei gegebener Dimension k und Summe $S = s$ immer proportional zum Multinomialkoeffizienten. Für die Dimension $k = 3$ ist das Gewicht eines Punktes $(x, y, s - x - y)$ äquivalent (also proportional) zu

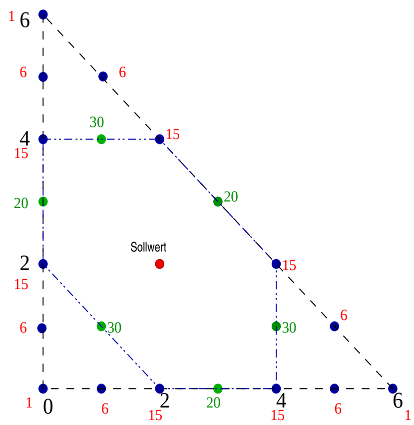
$$\binom{s}{x, y} = \binom{s}{x} \binom{s - x}{y} .$$

Für die beobachteten Werte ist bei $x = 2$ und $y = 4$ dieses Gewicht

$$\binom{6}{2, 4} = 15 .$$

Die Gewichte aller Punkte aus der Umgebung oder außerhalb sind in folgender Graphik neben den Punkten angegeben. Punkte mit Gewicht 15 oder weniger werden einbezogen.

Figur 2: Gewichte außerhalb der Umgebung



Die dunkel markierten Punkte (mit Gewichten) zählen zum Bereich, der zur Berechnung des P-Wertes herangezogen wird. Der P-Wert berechnet sich aus dem Verhältnis der Summe der Gewichte dieser Punkte zur Summe aller Gewichte. In diesem Beispiel beträgt der P-Wert 0.177 und liegt somit deutlich über dem üblicherweise verwendeten Signifikanzniveau von 5%. Die Beobachtungen widersprechen nicht deutlich genug der Annahme, dass die Fehlerraten in den drei Produktionen gleich sind. Die Nullhypothese wird nicht verworfen.
Angenommen, die Fehleruntersuchung bei den drei Produktionen führt zu folgendem Ergebnis.

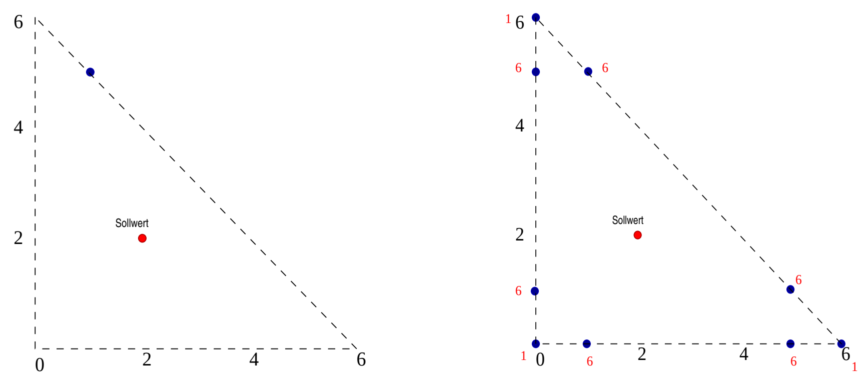
Produkt A: 1 Fehler B: 5 Fehler C: 0 Fehler .

Die Summe der Fehler ist gleich geblieben, die Ermittlung des P-Wertes erfolgt analog dem vorigen Beispiel mit gleichem Proportionalitätsfaktor. Die Beobachtung hat jetzt das Gewicht

$$\binom{6}{1,5} = 6$$

mit 9 Punkten, die gleich oder weniger Gewicht haben. Abbildung 3 zeigt den Abstand des Sollwerts vom Beobachtungspunkt und den Umgebungspunkten.

Figur 3: Umgebungspunkte für (1/5/0)



Die Summe der Umgebungspunkte führt auf einen P-Wert von 0.053. Bei einer Irrtumswahrscheinlichkeit von 5% kann die Hypothese noch knapp angenommen werden, bei einem Niveau von 10% wird aber die Annahme identer Fehlerraten in den drei Produktionen deutlich verworfen.

Die Durchführung des Tests bei höherer Dimension oder größeren Werten als im obigen Beispiel benötigt die Berechnung einer sehr großen Anzahl von Multinomialkoeffizienten, die noch dazu hohe Zahlen enthalten. Die aufwendige numerische Berechnung der Kombinationen kann anderseits durch die Anwendung einer Normalapproximation umgangen werden, ohne dass die Genauigkeit nennenswert leidet. Im Abschnitt 4 wird beschrieben, wie für die Binomialwahrscheinlichkeit eine Normalwahrscheinlichkeit verwendet wird. Im hochdimensionalen Fall ist eine solche Approximation einer Multinomialverteilung durch eine multivariate Normalverteilung unbedingt erforderlich, um den Test durchführen zu können.

Prinzipiell ist eine Multinomialverteilung $M_{N;(\theta_1,\ldots,\theta_k)}$ für große N durch eine multivariate

Normalverteilung mit identem Mittelvektor und identer Kovarianzmatrix zu approximieren. Diese Normalverteilung ist $(k-1)$-dimensional, wie die Multinomialverteilung auch, die k Parameter werden nur der Übersichtlichkeit wegen angegeben. Für den Test wird die Multinomialverteilung $M_{s;(1/k,\ldots,1/k)}$ mit Gesamtanzahl s und Eintrittswahrscheinlichkeit $1/k$ angenähert.

Die bedingte Verteilung des Zählvektors $X_1, X_2, \ldots, X_k$ unter der Gesamtanzahl $S = s$ ist asymptotisch normalverteilt

$$(X_1, X_2, \ldots, X_k) \xrightarrow{D} N(\mu, \Sigma)$$

mit dem Sollwert als Mittelvektor

$$\mu = (\frac{s}{k}, \ldots, \frac{s}{k})^\top$$

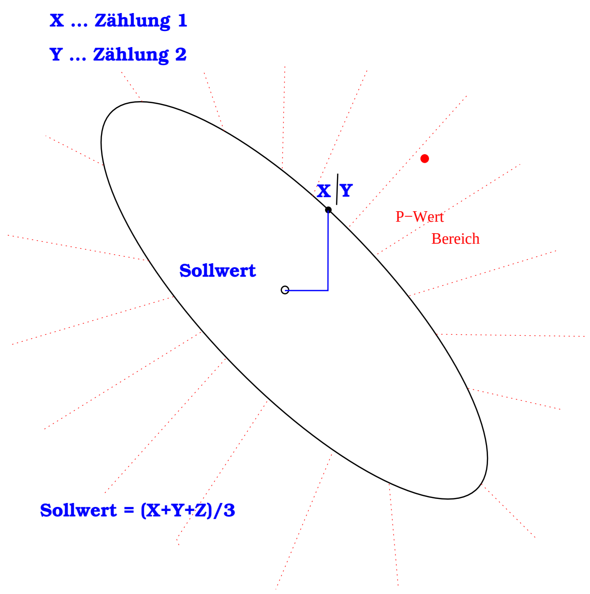
und der Kovarianzmatrix

$$\Sigma = \mathrm{diag}(\frac{s}{k}, \ldots, \frac{s}{k}) - \frac{s}{k^2}\mathbf{1}\mathbf{1}^\top.$$

Dabei bezeichnet $\mathbf{1} = (1, \ldots, 1)^\top$ den Einsvektor, und die k-ten Komponenten bei dem Vektor und der Matrix werden für die $(k-1)$-dimensionale Normalverteilung nicht berücksichtigt. Der Konfidenzbereich einer multivariaten Normalverteilung besteht aus einem Ellipsoid, und der Umgebungsbereich einer Beobachtung ist dann der Bereich außerhalb des Ellipsoids.

Die folgende Abbildung zeigt die graphische Darstellung des Umgebungsbereichs für eine Beobachtung der Dimension 3. Punkte außerhalb der Ellipse (rot markiert) bilden den Umgebungsbereich mit dem P-Wert als Gewicht.

Figur 4: Annahmebereich des Tests und Umgebung



Da die Summe gegeben ist, wird vom Beobachtungsvektor (X, Y, Z) nur (X, Y) berücksichtigt. Der Abstand des Beobachtungswertes vom Sollwert bestimmt die Entscheidung über die Hypothese identischer Verteilungen der Komponenten. Der Abstand wird aber mit einer durch die Kovarianzmatrix transformierten Distanz (Mahalanobis-Distanz) gemessen.

Mit der multivariaten Verteilungsfunktion, die in vielen Mathematik-Paketen implementiert ist, lässt sich der exakte Poisson-Test mit geringem Aufwand durchführen.

Für eine aktuelle politische Fragestellung wenden wir den exakten Poisson-Test an. Das Problem einer steigenden Kriminalität wird häufig diskutiert. Die Anzahl aller angezeigten Fälle innerhalb eines Jahres ist sicher Poisson-verteilt. Um die Behauptung einer steigenden Kriminalität zu belegen, muss die Hypothese einer gleich bleibenden Rate signifikant verworfen werden. Die österreichische Kriminalstatistik weist folgende Zahlen für alle angezeigten Fälle aus.

Jahr	Angezeigte Fälle
2006	591.597
2007	572.695
2008	594.240
2009	589.495

Der Durchschnittswert der vier Jahre beträgt 587.007, dieser Sollwert und die Beobachtungswerte spannen den Umgebungsraum auf, dessen Gewicht mit der 3-dimensionelen Normalverteilung berechnet wird. Man erhält die Wahrscheinlichkeit 0.0027, das ist der P-Wert. Eine Ablehnung der Hypothese einer gleich bleibenden Rate ist nicht möglich, aus dieser Statistik kann nicht auf Veränderungen in der Anzahl angezeigter Fälle geschlossen werden.

Zusammenfassung und Erweiterungen

Die vorgestellte Methode beinhaltet Tests für Häufigkeiten und Zählungen, die auch bei kleinen Stichproben anwendbar sind und zudem optimal sind. Die Konstruktion basiert auf den Eigenschaften suffizienter und vollständiger Statistiken, die mit dem bedingten Erwartungswert zu optimalen Testverfahren führen. Solche Statistiken existieren auch für andere Verteilungen. So kann man dieses Prinzip beispielsweise auch für geometrische Verteilungen anwenden. Die Zählung erfolgt dann nicht in einer festen Zeiteinheit wie beim Poissonprozess, sondern dauert bis zum Eintritt des ersten Erfolges. Die Negative-Binomialverteilung entspricht dann der Zählung bis zum wiederholten Erfolg (Eintritt eines bestimmten Ereignisses).

Ein weiterer Aspekt in der Analyse von Zählvergleichen ist die Berücksichtigung von Kenntnissen und Erfahrungen über die Struktur der Zählung. Primär bei wenigen Beobachtungen sind Expertenwissen und a-priori Information wertvoll für die Bewertung von Testergebnissen. Dazu werden Methoden von lernenden Systemen im Rahmen der Bayes‘schen Statistik angewandt, deren eigentliche Grundlage bedingte Wahrscheinlichkeitsverteilungen und bedingende

Ereignisfelder darstellen. Im allgemeinen reagieren Unterscheidungsverfahren unter Vorinformation für quantitative Größen sensitiver als Verfahren ohne Vorinformation. Es bestehen viele Möglichkeiten aussagekräftige statistische Verfahren auch bei kleinen Stichproben durchzuführen.

Literatur

Casella, G., Berger R. (1986). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *J. Amer. Statist. Assoc., Vol. 82, 106-111.*

Hedayat, A., Jacroux M., Majumdar D. (1988). Optimal Designs for Comparing Test Treatments with Controls. *Statistical Science, Vol. 3, 462-491.*

Lehmann, E. (2008). *Testing Statistical Hypotheses.* Wiley, New York.

Sachs L., Hedderich J. (2006). *Angewandte Statistik. Methodensammlung mit R.* Springer, Berlin.