Exkurs:
Zur statistischen Absicherung mit dem Chi-Quadrat-Test

Von Karl-Helge Deutrich

Dieser Exkurs ist für Leser gedacht, die mit statistischen Argumentationsweisen gewöhnlich nicht arbeiten. Es wird knapp referiert, mit welchen theoretischen Annahmen und mit welchen Rechenoperationen gearbeitet werden kann, um festzustellen, ob eine bestimmte Hypothese durch empirisch gewonnene Meßwerte zurückgewiesen werden muß oder nicht zurückgewiesen werden kann.

1. Aus Texten, die unter gewissen außersprachlichen Bedingungen und Verhaltensmerkmalen entstanden sind, sind Stichproben gezogen worden. Die Texte der einen Stichprobe sind unter außersprachlichen Bedingungen entstanden, die zum Redekonstellationstyp UNT gehören, eine weitere Stichprobe gehört der außersprachlichen Provenienz nach zum Redekonstellationstyp ERZ, eine dritte zum Redekonstellationstyp DIS.

Es soll festgestellt werden, ob hinsichtlich der Verteilung der Wortart „Substantiv" die Texte der drei Stichproben aus drei verschiedenen Textsorten im oben definierten Sinne stammen, oder ob eine einzige Textsorte angenommen werden muß.

Mit dem sogenannte Chi-Quadrat-Test wird überprüft, ob bei den beobachteten Häufigkeiten der Substantivverteilung die Aussage gerechtfertigt ist, daß ein signifikanter, d. h. systematischer und nicht nur zufälliger Unterschied zwischen den einzelnen Stichproben besteht. Hierzu die Tabelle, in der die beobachteten Häufigkeiten in den einzelnen Stichproben eingetragen sind:

Wörter aller Wortarten \ RK-Typ		UNT	ERZ	DIS	Randsumme
Wortart Substantiv	beob.	a) 766	b) 1673	c) 3303	5742
	erw.	(1217,1)	(1628,1)	(2897,9)	
Wörter anderer Wortarten	beob.	d) 6269	e) 7738	f) 13448	27455
	erw.	(5817,9)	(7782,9)	(13853,1)	
Randsumme		7035	9411	16751	33197

Als sogenannte Nullhypothese wird die Hypothese getestet, daß die drei Stichproben, die übrigens Zufallsstichproben sind, aus einer Textsorte (statistisch gesehen aus derselben Population) entnommen sind, und daß demnach kein Unterschied zwischen den Meßwerten beobachtbar ist, der nicht auf der zufälligen Streuung bei Stichproben (Stichprobenvariabilität) beruht.
Die Alternativhypothese zu dieser Nullhypothese lautet, daß ein systematischer Unterschied hinsichtlich der Substantivverteilung zwischen den Stichproben besteht, und daß somit die drei Stichproben aus drei verschiedenen Textsorten stammen.
Wenn wir annehmen, daß die Nullhypothese zutrifft, müßte in jeder Stichprobe der gleiche Anteil von Substantiven an allen Wörtern zu messen sein, und dann gehören die drei Stichproben der gleichen Textsorte an.
Im gesamten Stichprobenmaterial sind 33197 Wörter enthalten (vgl. Tabelle). Der Anteil der Substantive an allen Wörtern beträgt 5742 oder 17,3%.
Falls die drei Stichproben aus der gleichen Textsorte gezogen worden sind, dann müßte der relative Anteil der Substantive am Wortmaterial jeder der drei Stichproben in etwa dem Anteil entsprechen, der für das gesamte Stichprobenmaterial errechnet worden ist.
Unter der Annahme, daß die Nullhypothese gilt, müßte dann in jedem Feld der Tabelle, das Anteile der Substantive wiedergibt, in etwa 17,3% erscheinen. Für die Stichprobe UNT wären demnach bei Gültigkeit der Nullhypothese 17,3% von 7035 Wörtern, oder 1217,1 Wörter als Substantive zu erwarten.
Wenn man diese Rechenoperation für alle Felder der Tabelle anwendet, erhält man die erwarteten Häufigkeiten in der Verteilung des Wortmaterials, für den Fall daß alle drei Stichproben aus der gleichen Population stammen. Die erwarteten Häufigkeiten sind in der Tabelle in Klammern unter die tatsächlich beobachteten Häufigkeiten gesetzt.
Beim Chi-Quadrat-Text wird geprüft, ob die Verteilung der Häufigkeiten in den Feldern signifikant verschieden ist von der Verteilung der Häufigkeiten in den Zeilen-Randsummen. Die Nullhypothese fordert, daß diese Verteilungen identisch sind.
Die erwarteten Häufigkeiten sind als Werte nicht in jeder Stichprobe gleich groß, da der Umfang der einzelnen Stichproben, gemessen am gesamten Wortmaterial, unterschiedlich ist. Die Werte der erwarteten Häufigkeiten spiegeln in jeder Stichprobe den Anteil wider, der in der

Gesamtstichprobe für die Substantive bzw. alle anderen Wortarten errechnet worden ist (vgl. die Tabelle).
Um nun die Nullhypothese (= der Anteil der Substantive ist für die drei Stichproben gleich groß) zu prüfen, wird die statistische Wahrscheinlichkeit berechnet, mit der ein Zufallsunterschied zwischen den beobachteten (d. h. gemessenen) und den erwarteten Häufigkeiten angesetzt werden darf.
Die Wahrscheinlichkeiten für die Chi-Quadrat-Verteilung sind in bereits tabellierter Form in Statistik-Lehrbüchern einzusehen.
Für den eigentlichen Test müssen noch die Freiheitsgrade bestimmt werden: Bei unserer Tabelle mit 6 Feldern wird das Produkt aus Zeilenanzahl minus 1 und Spaltenanzahl minus 1 als Freiheitsgrad angesetzt, d. h. wir berücksichtigen zwei Freiheitsgrade.
2. Für die Berechnung der Teststatistik muß für jedes Feld der 6-Felder-Tabelle die erwartete Häufigkeit errechnet werden. Wenn die Nullhypothese zutrifft, müssen die beobachtete und die erwartete Häufigkeit in jedem Feld annähernd gleich sein. Ist die Differenz zwischen beobachteter und erwarteter Häufigkeit groß, dann wächst die Wahrscheinlichkeit, daß die Nullhypothese nicht zutrifft.
Wir gehen von der Differenz zwischen beobachteter und erwarteter Häufigkeit für jedes Feld aus, und quadrieren diese Differenz. Bei steigender Differenz wächst dann das Quadrat der Differenz schneller als die Differenz selbst. Relativiert wird diese quadrierte Differenz durch Division mit der erwarteten Häufigkeit des in Frage kommenden Feldes.
Der Wert für Chi-Quadrat errechnet sich aus der Summe der quadrierten Differenz von beobachteter und erwarteter Häufigkeit in jedem Feld, dividiert durch den erwarteten Wert in dem betreffenden Feld; als Formel:

$$\text{Chi}^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$\sum$ bedeutet das Symbol für Summierung, f_o stellt die beobachtete und f_e die erwartete Häufigkeit für jedes Feld in der Tabelle dar.

Für jedes Feld wird $\frac{(f_o - f_e)^2}{f_e}$ berechnet.

Feld	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
a	766	1217,1	–451,1	203491,2	167,194
b	1673	1628,1	44,9	2016,0	1,238
c	3303	2897,9	405,1	164106,0	56,629
d	6269	5817,9	451,1	203491,2	34,977
e	7738	7782,9	–44,9	2016,0	0,259
f	13448	13853,1	–405,1	164106,0	11,846
Summe	33197	33197			272,143

Unsere Teststatistik ergibt einen Chi-Quadrat-Wert von

272,143.

Die Wahrscheinlichkeitsverteilung der Chi-Quadrat-Werte bei sehr vielen Stichproben, unter der Annahme, daß die Nullhypothese gilt, ist in einer Verteilungstabelle (sampling distribution) als Stichprobenvarianz tabelliert. In einer Chi-Quadrat-Verteilungstabelle finden wir unter Berücksichtigung von zwei Freiheitsgraden unter dem Signifikanzniveau von 0,1% (p = 0.001) einen Chi-Quadrat-Wert von 13,815. Das bedeutet: Unter der Annahme, die Nullhypothese wäre korrekt, würden wir einen Wert von Chi-Quadrat in dieser Größe oder größer nur einmal bei 1000 Analysen mit drei Stichproben erhalten; d. h. aber, daß nur äußerst selten die Differenzen von beobachteten und erwarteten Häufigkeiten so hoch oder höher als 13,815 sein werden, wenn es keine Unterschiede zwischen den Textgruppen unter dem Aspekt der Verteilung der Substantive geben würde.

Wir entscheiden uns, daß wir bei einem Ergebnis, das nur eine sehr geringe Wahrscheinlichkeit des Auftretens besitzt, annehmen, die Nullhypothese sei falsch und müsse daher zurückgewiesen werden.

Wir nehmen daher an, daß unser Analyseexperiment mit den drei Stichproben nicht gerade dasjenige ist, das nur einmal bei 1000 Experimentmessungen auftritt, wenn die Nullhypothese stimmt. Wir weisen mit unserer Stichprobenmaßzahl Chi-Quadrat = 272,143 bei 0.1% Fehlerwahrscheinlichkeit die Nullhypothese zurück.

Unser Ergebnis lautet:
Auf Grund des getesteten Stichprobenmaterials kann die Nullhypothese zugunsten der Alternativhypothese zurückgewiesen werden; es gilt demnach:
In den drei Stichproben ist ein systematischer Unterschied hinsichtlich der Substantivverteilung auf Grund des Chi-Quadrat-Tests anzunehmen, da der gemessene Chi-Quadrat-Wert den erwarteten Wert weit übersteigt. Die gemessenen Werte der drei Stichproben deuten hinsichtlich der Substantivverteilung darauf hin, daß die Stichproben aus drei verschiedenen Textsorten stammen.
Da die Stichprobe verhältnismäßig groß ist, ist eine statistische Signifikanz leichter zu erreichen. Im eben diskutierten Fall wurden signifikante Beziehungen zwischen Redekonstellationstyp und Substantivgebrauch aufgedeckt. Zu fragen ist jedoch, wie *stark* diese Beziehungen sind. Für die vorliegenden verteilungsfreien Nominaldaten soll die Stärke der Beziehung zwischen den Variablen mit Pearson's Kontigenzkoeffizient „C" gemessen werden. Der Wert 0 (Null) signalisiert: keine Beziehung; je weiter sich C dem Wert 1 nähert, desto stärker wird die Beziehung. Dabei ist zur Interpretation der Werte zu beachten, daß ein C-Wert von 0.3 für Nicht-Naturwissenschaften bereits beachtenswerte Beziehungen anzeigt.
Pearson's C berechnet sich nach der Formel:

$$C = \sqrt{\frac{Chi^2}{Chi^2 + N}}$$

Im Falle der Substantivdistribution ergibt C:

$$C = \sqrt{\frac{272.143}{33469.143}} = 0.09$$

Die Beziehung ist nur „schwach", was aber nicht verwundern kann, da sich Textsorten sicher nie in solche trennen lassen, die ausschließlich Substantive und solche, die keinerlei Substantive aufweisen. Die für unterschiedliche linguistische Einheiten errechneten Kontingenzkoeffizienten erlauben eine Vergleichbarkeit der Stärke der Beziehungen.

3. Im folgenden werden kurz die Ergebnisse anderer getesteter Merkmalskombinationen dargestellt. Die Methode ist dem beschriebenen Verfahren analog und muß nicht mehr dargestellt werden.

Attribuierung von Substantiven
(vgl. für die Zahlen die Tabelle 1b)

gemessener Chi-Quadrat-Wert: 72.19

auf einem Signifikanzniveau
von 0.1% erwarteter Wert
bei 6 Freiheitsgraden: 22.457

$$\text{Kontingenzkoeffizient } C = \sqrt{\frac{72.19}{3985.19}} = 0.13$$

Mit einer Fehlerwahrscheinlichkeit von 0.1% kann die Nullhypothese zurückgewiesen werden, welche besagt, daß die drei Stichproben hinsichtlich der Attribuierung von Substantiven aus einer Grundgesamtheit, d. h. einer Textsorte stammen; die gemessenen Daten sprechen nicht dagegen, daß hinsichtlich der Attribuierung von Substantiven drei verschiedene Textsorten angesetzt werden können.

Form abhängiger Sätze
(vgl. für die Zahlen Tabelle 2)

gemessener Chi-Quadrat-Wert: 70.976

auf einem Signifikanzniveau
von 0.1% erwarteter
Chi-Quadrat-Wert bei
9 Freiheitsgraden: 27.877

$$\text{Kontingenzkoeffizient } C = \sqrt{\frac{70.976}{1134.976}} = 0.25$$

Auf dem 0.1% Signifikanzniveau sprechen die gemessenen Daten nicht dagegen, hinsichtlich der Verteilung verschiedener Formen abhängiger Sätze vier verschiedene Textsorten anzunehmen.

Anteil der Gefügesätze
(vgl. für die Zahlen Tabelle 2b)

gemessener Chi-Quadrat-Wert: 156.141

auf einem Signifikanzniveau von 0.1% erwarteter Chi-Quadrat-Wert bei 2 Freiheitsgraden: 13.815

$$\text{Kontingenzkoeffizient } C = \sqrt{\frac{156.141}{2083.141}} = 0.27$$

Auf dem 0.1% Signifikanzniveau läßt sich die Nullhypothese, welche besagt, daß die drei Stichproben aus einer Grundgesamtheit gezogen worden sind, zurückweisen. Daher gilt die Alternativhypothese: Hinsichtlich der Verteilung der Gefügesätze sprechen die gemessenen Daten nicht dagegen, drei verschiedene Grundgesamtheiten (= Textsorten) anzunehmen.

Verteilung bestimmter Konjunktionen
(vgl. für die Zahlen Tabelle 3)

gemessener Chi-Quadrat-Wert: 69.341

auf einem Signifikanzniveau von 0.1% erwarteter Chi-Quadrat-Wert bei 10 Freiheitsgraden: 29.588

$$\text{Kontingenzkoeffizient } C = \sqrt{\frac{69.341}{546.341}} = 0.36$$

Auf dem 0.1% Signifikanzniveau sprechen die gemessenen Daten nicht dagegen, hinsichtlich der Verteilung der Konjunktionen *daß*, *wenn*, *wie*, *ob*, *als* und anderer Konjunktionen drei verschiedene Textsorten anzunehmen.

Verteilung von Parenthesen auf Funktionstypen
(vgl. für die Zahlen die Tabelle 4)

gemessener Chi-Quadrat-Wert: 410.937

auf einem Signifikanzniveau
von 0.1% erwarteter
Chi-Quadrat-Wert bei
2 Freiheitsgraden: 13.815

$$\text{Kontingenzkoeffizient } C = \sqrt{\frac{410.937}{1463.937}} = 0.53$$

Auf dem 0.1% Signifikanzniveau sprechen die gemessenen Daten nicht dagegen, hinsichtlich der Verteilung der Funktionen von Parenthesen drei verschiedene Textsorten anzunehmen.