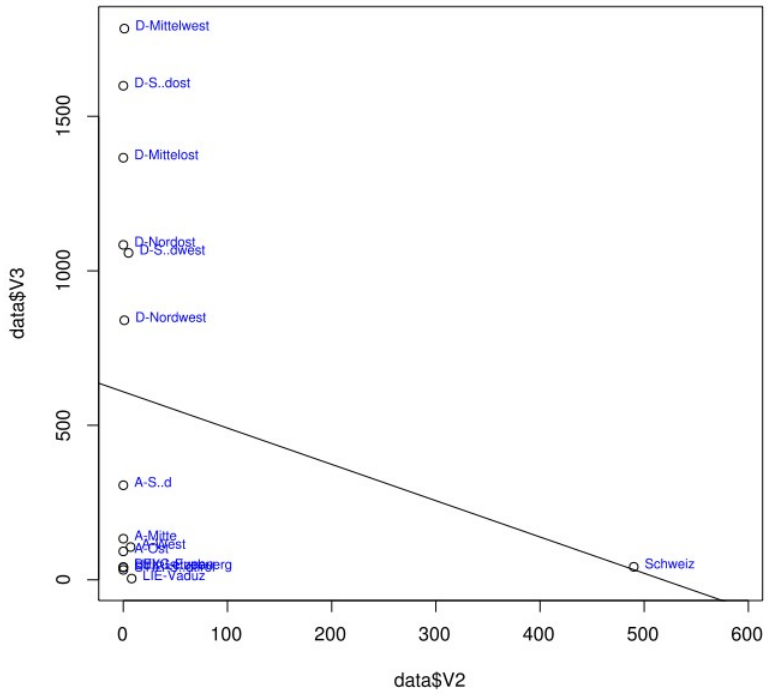
Beispielanalyse für „parkieren“ vs. „parken“

Die statistische Auswertung folgt weitgehend:
http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=4683

Die Ausgabe der automatischen statistischen Auswertung auf der Website (http://kitt.ifi.uzh.ch/kitt/cores/stats2.html) beginnt folgendermassen:

V2: parkieren.txt → Die Belege von „parkieren“ im ganzen Korpus (Export aus CQPWeb)
V3: parken.txt → Die Belege von „parken“ im gesamten Korpus

Dann folgt der Plot der Regionen:



Auf der Y-Achse (data$V3) werden die Belege für „parken“ hochgezählt, auf der X-Achse (data$V2) die Belege für „parkieren“. Dann werden die Regionen in den Plot gesetzt anhand der Belegzahlen für die Varianten (V2, V3). Die schwarze Linie zeigt den Durchschnitt (die Regressionslinie) durch die Daten (je zerstreuter die Regions-Punkte um die Linie verteilt sind, desto heterogener ist die Verteilung der Belegzahlen; wenn alle Punkte auf der Linie liegen, dann gibt es praktisch keine Varianz; Punkte, die nahe beieinander liegen, haben ähnliche Zählungen bzgl. der Varianten usw.). Was man hier schön sieht, ist dass die Schweiz weit weg von den anderen Regionen liegt. „Schweiz“ ist ganz rechts, weil praktisch nur „parkieren“ (V2) vorkommt, die restlichen Regionen sind ganz links, weil (fast) nur „parken“ (V3) in ihnen vorkommt. Der Plot ist also eine erste Übersicht über die Daten.

Als nächstes kommt der eigentliche Chi-Quadrattest:

Pearson's Chi-squared test

- data:mtrx
- X-squared = **8011.733** , df = **14** , p-value = < 2.2e-16

Relevant ist hier p-value =< 2.2e-16

Wenn p-value < 0.05 ist, dann sagt der Chi-Quadrattest, dass die Heterogenität der Verteilung der Belege statistisch gesehen signifikant ist. Das ist hier der Fall, weil 2.2e-16 heisst „0, Komma, 15 Nullen, dann 22“ (also 2.2e-16 ist sehr, sehr viel kleiner als 0.05). Eigentlich könnte man hier aufhören, da man ja hat, was man wollte ;) Aber wir wollen es ja genauer wissen.

Jetzt kommen die vier Tabellen.

Chi square posthoc

	V2	V3	V4	V5	V6
A-Mitte	0	133	133	0	100
A-Ost	0	92	92	0	100
D-Nordost	0	1084	1084	0	100
A-Südost	0	306	306	0	100
BELG-Eupen	0	41	41	0	100
STIR-Südtirol	0	32	32	0	100
D-Südost	0	1599	1599	0	100
D-Mittelost	0	1366	1366	0	100
LUX-Letzebuerg	0	40	40	0	100
D-Nordwest	1	840	841	0	100
D-Mittelwest	1	1784	1785	0	100
D-Südwest	5	1058	1063	0	100
A-West	7	106	113	6	94
LIE-Vaduz	8	4	12	67	33
Schweiz	490	42	532	92	8

Observierte Werte Phänomen 1 (V2), Phänomen 2 (V3), Summe der beiden (V4).
V5 ist V2 in Prozente, V6 ist V3 in Prozente. Summe von V2=512, Summe von V3=8527.

	V2	V3
A-Mitte	7.5335767	125.4664233
A-Ost	5.2111959	86.7888041
D-Nordost	61.4014825	1022.5985175
A-Südost	17.3328908	288.6671092
BELG-Eupen	2.3223808	38.6776192
STIR-Südtirol	1.8125899	30.1874101
D-Südost	90.5728510	1508.4271490
D-Mittelost	77.3749309	1288.6250691
LUX-Letzebuerg	2.2657374	37.7342626
D-Nordwest	47.6371280	793.3628720
D-Mittelwest	101.1085297	1683.8914703
D-Südwest	60.2119704	1002.7880296
A-West	6.4007080	106.5992920
LIE-Vaduz	0.6797212	11.3202788
Schweiz	30.1343069	501.8656931

Erwartete Werte Phänomen 1 (V2), Phänomen 2 (V3)

	V2	V3
A-Mitte	-7.533576723	0.452350332
A-Ost	-5.211195929	0.312903989
D-Nordost	-61.401482465	3.686825264
A-Südost	-17.332890807	1.040745877
BELG-Eupen	-2.322380794	0.139446343
STIR-Südtirol	-1.812589888	0.108836170
D-Südost	-90.572850979	5.438407377
D-Mittelost	-77.374930855	4.645944013
LUX-Letzebuerg	-2.265737360	0.136045213
D-Nordwest	-45.658120030	2.741521925
D-Mittelwest	-99.118420067	5.951522350
D-Südwest	-50.627170183	3.039886377
A-West	0.056111113	-0.003369167
LIE-Vaduz	78.835971208	-4.733671544
Schweiz	7017.797238636	-421.380577716

Unterschied Observiert-Erwartet
Hat ein Unterschied ein negatives Vorzeichen '-', liegt der observierte unter dem erwartetem Wert. Ist kein Vorzeichen vorhanden (implizit '+'), liegt der observierte über dem erwarteten Wert.

	V2	V3
A-Mitte	-2.85	2.85
A-Ost	-2.36	2.36
D-Nordost	-8.60	8.60
A-Südost	-4.36	4.36
BELG-Eupen	-1.57	1.57
STIR-Südtirol	-1.39	1.39
D-Südost	-10.80	10.80
D-Mittelost	-9.83	9.83
LUX-Letzebuerg	-1.55	1.55
D-Nordwest	-7.31	7.31
D-Mittelwest	-11.44	11.44
D-Südwest	-7.80	7.80
A-West	0.25	-0.25
LIE-Vaduz	9.15	-9.15
Schweiz	88.91	-88.91

Standardisierte Pearson residuals
Liegt ein Wert über 2 oder 3 (absolut), kann (als Daumenregel) von einem signifikanten Unterschied gesprochen werden. Die Vorzeichen ('-' negativ; kein, resp. implizit '+' positiv) zeigen wieder die Richtung des Unterschieds an.

Die erste listet die Anzahl Belege für V2 („parkieren“) und V3 („parken“) pro Region. V5 und V6 ist dasselbe als Prozentsätze. Also bei Schweiz sind 92% der Belege in V2 („parkieren“) etc.

Die zweite Tabelle (Erwartete Werte) gibt an, was die Anzahl Belege in den Regionen sein müssten, wenn die Belege gleichmässig verteilt wären (wie das berechnet wird, führt hier aber zu weit). (Achtung: Wenn hier mehr als ca. 20% der erwarteten Werte unter 5 liegen, gilt der Chi-Quadrattest als

nicht mehr zuverlässig.)

Die dritte Tabelle (Unterschied Observiert-Erwartet) zeigt den Unterschied zwischen den Werten in der ersten und zweiten Tabelle (es ist nicht eine reine Subtraktion der Werte in den Tabellen; die Details führen hier wieder zu weit). Wenn man hier grosse Zahlen sieht, weiss man, dass diese Regionen „auffällig“ bzgl. ihrer Belegverteilung sind (Die Schweiz hat geradezu astronomisch viele Belege für V2 (parkieren) und viel zu wenige für V3 (parken) im Vergleich zu den anderen Regionen; A-West dagegen verhält sich sehr „brav“).

Die letzte Tabelle ist nochmals dasselbe in Grün; es ist eine Art Normalisierung der Tabelle 3, die es erlaubt, auf die statistische Signifikanz der „Auffälligkeit“ einer Region zu schliessen (wenn ein Wert ungefähr über 3 (absolut; also Wert < -3 oder Wert >3) liegt, dann geht man von signifikanter Abweichung aus). Wieder sticht die Schweiz heraus; A-West ist unauffällig.

Insgesamt dienen die vier Tabellen dazu herauszufinden, welche Region(en) für die Heterogenität der Belegverteilung verantwortlich ist. Es ist denkbar, dass besonders eine Region auffällig ist (wie hier die Schweiz). Es kann aber auch sein, dass alle oder mehrere Regionen ziemlich unterschiedliche Verteilungen aufzeigen.

Als nächstes folgt die Gegenüberstellung einzelner Regionen ggü. dem Rest der Regionen:

Chi Square Test zwischen jeweils einer Kategorie und der Summe der restlichen

	p-value
A-Mitte	0.00786036743435074
A-Ost	0.0327015667562983
D-Nordost	1.4655738062638e-17
A-Südost	2.28449414000886e-05
BELG-Eupen	0.217196006309492
STIR-Südtirol	0.314623415840427
D-Südost	6.55790428966079e-27
D-Mittelost	1.57286709536831e-22
LUX-Letzebuerg	0.226106696103189
D-Nordwest	4.94505501299398e-13
D-Mittelwest	4.9622314570834e-30
D-Südwest	1.09195851532327e-14
A-West	0.967564972509028
LIE-Vaduz	1.55610871911802e-17
Schweiz	0

Dabei wird ein Chi-Quadrattest gemacht, indem alle Regionen ausser einer zusammengenommen werden und gegen die eine, die nicht genommen wurde, verglichen wird. Die Belege in den zusammengenommenen Regionen werden addiert und deren Summe dann mit der isolierten Region verglichen. Hierbei geht es wieder darum zu sehen, welche der Regionen besonders auffällig sind bzgl. der Belegverteilung auf V2 und V3 im Vergleich zur Gesamtheit der anderen. Je kleiner der p-Wert, desto auffälliger. Wieder sieht man, dass die Schweiz sehr auffällig ist (p-Wert 0) und dass A-West sehr „normal“ ist (p-Wert fast 1).

Im Anschluss folgt der paarweise Chi-Quadrattest von jeweils zwei Regionen:

Der ist eigentlich nicht besonders interessant, da er nichts über das gesamte Korpus aussagt, sondern

Paarweiser Vergleich (Chi Square Test zwischen jeweils 2 Kategorien)

	A-Mitte	A-Ost	D-Nordost	A-Südost	BELG-Eupen	STIR-Südtirol	D-Südost	D-Mittelost	LUX-Letzebuerg	D-Nordwest	D-Mittelwest	D-Südwest	A-West	LIE-Vaduz
A-Ost	NaN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D-Nordost	NaN	NaN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
A-Südost	NaN	NaN	NaN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
BELG-Eupen	NaN	NaN	NaN	NaN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
STIR-Südtirol	NaN	NaN	NaN	NaN	NaN	NA	NA	NA	NA	NA	NA	NA	NA	NA
D-Südost	NaN	NaN	NaN	NaN	NaN	NaN	NA	NA	NA	NA	NA	NA	NA	NA
D-Mittelost	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NA	NA	NA	NA	NA	NA	NA
LUX-Letzebuerg	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NA	NA	NA	NA	NA	NA
D-Nordwest	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	NA	NA	NA	NA	NA
D-Mittelwest	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	NA	NA	NA	NA
D-Südwest	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	1.0e+00	NA	NA	NA
A-West	7.9e-01	1.0e+00	2.6e-12	5.1e-03	1.0e+00	1.0e+00	2.2e-18	1.2e-15	1.0e+00	7.3e-08	1.3e-17	9.7e-06	NA	NA
LIE-Vaduz	1.2e-17	2.5e-12	4.0e-139	7.5e-40	1.3e-05	2.1e-04	9.9e-205	4.7e-175	1.7e-05	6.9e-96	8.0e-203	3.9e-83	1.0e-06	NA
Schweiz	2.3e-100	8.7e-85	9.1e-311	5.9e-147	3.4e-55	3.4e-47	0.0e+00	0.0e+00	2.2e-54	2.7e-260	0.0e+00	1.1e-301	1.0e-83	6.3e-01

eben nur über jeweils zwei Regionen. In der Tabelle werden wiederum die p-Werte des Chi-Quadrattests angezeigt. Man sieht hier, dass z.B. A-West und A-Ost sich nicht unterscheiden (p= 1.0e+00; d.h. 1). Die Schweiz hingegen unterscheidet sich maximal von beispielsweise D-Südost (p=0.0e+00, d.h. 0).

Damit ist die Auswertung bzgl. der Regionen abgeschlossen. Darauf folgend kommt dieselbe Auswertung bezogen auf die Zeitungen und dann bgzl. der Länder. D.h. wir analysieren die Verteilung von „parken“ vs. „parkieren“ nicht in Bezug auf die Regionen, sondern in Bezug auf die Zeitungen und dann die Länder. Man kann sich Zeitungen, Regionen und Länder als verschiedene „Zoom“-Stufen bzgl. der Analyse der Verteilung vorstellen: Die Zeitungen bilden die feingliedrigste Analyseebene, die Länder die grobkörnigste. Wir fokussieren im Projekt aber v.a. die Regionen.

Die Analyse unseres Beispiels anhand der Verteilung in den Regionen hat gezeigt, dass die Varianten „parken“ und „parkieren“ statistisch signifikant heterogen verteilt sind. Wir haben gesehen, dass die Region „Schweiz“ Auslöser dieser Heterogenität ist, wobei „parkieren“ fast ausschliesslich in der Schweiz vorkommt.