Wörter, Texte und Frequenzen

Frequenzvergleiche

Wintersemester 2021/2022

Andreas Blombach, Philipp Heinrich

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Theorie, Hypothese, Überprüfung (Wiederholung)

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Grundprinzip

- am Anfang: **Fragestellung** – z.B. soll irgendeine **Anfangsbeobachtung** erklärt werden
- **Theoriebildung** (unter Einbezug verfügbarer Literatur)
- Ableitung überprüfbarer **Hypothesen** aus dieser Theorie
- Identifizierung der **Variablen** und geeigneter Messmethoden
- **Datensammlung** zur Überprüfung der Hypothese(n): Messen der Variablen
- **Datenanalyse**
- **Ablehnung** oder **Bestätigung** der Hypothese(n)
- ggf. Anpassung der Theorie oder neue Theorie

Hypothesen

- Aussagen/Annahmen, die empirisch prüfbar sind
- häufig konkrete Voraussagen, die sich aus einer Theorie ergeben
- nicht verifizierbar (an Popper und den Falsifikationismus denken!)
- aber: Ergebnisse können Hypothesen **stützen**

Hypothesen prüfen (1)

- Hypothese: Wort X wird in Korpus A häufiger verwendet als in Korpus B.
- Brauchen wir Statistik, um diese Hypothese zu prüfen?

Hypothesen prüfen (1)

- Hypothese: Wort X wird in Korpus A häufiger verwendet als in Korpus B.
- Brauchen wir Statistik, um diese Hypothese zu prüfen?
- Nein: Da es uns einzig und allein um Korpus A und Korpus B geht (die uns komplett vorliegen), haben wir es mit **Grundgesamtheiten** zu tun, nicht mit **Stichproben**. Wir müssen also nur die jeweiligen Worthäufigkeiten in den beiden Korpora ermitteln und vergleichen.
- Beispiel: Wort X kommt in Korpus A 50mal vor, in Korpus B dagegen nur 10mal.

Hypothesen prüfen (2)

- Beispiel: Wort X kommt in Korpus A 50mal vor, in Korpus B dagegen nur 10mal.
- Dabei müssen wir natürlich darauf achten, dass wir, wenn die Korpora nicht zufällig gleich groß sein sollten, relative Häufigkeiten vergleichen, nicht absolute.

- Beispiel: Korpus A enthält 15 Mio. Wörter, Korpus B dagegen 6 Mio.
  - Relative Häufigkeit in Korpus A: 3,33 pMW
  - Relative Häufigkeit in Korpus B: 1,67 pMW

  ⇒ Das Wort kommt in Korpus A ungefähr doppelt so häufig vor.

Hypothesen prüfen (3)

- Was ist nun aber, wenn unsere Korpora für größere Sprachausschnitte stehen sollen – wenn es also darum geht, von den Korpora (als Stichproben) auf die Sprachausschnitte zu schließen, aus denen sie stammen (Grundgesamtheiten)?
- Das ist etwa dann der Fall, wenn unsere Hypothese so aussieht: Wort X kommt in Textsorte Y nicht genauso oft vor wie in Textsorte Z.

Hypothesen prüfen (4)

- Wir müssten dann zunächst überlegen, wie sich Y und Z definieren, ein- und abgrenzen lassen, um die Frage zu klären, um welche Grundgesamtheiten es überhaupt geht.
- Anschließend müssten wir eigentlich Zufallsstichproben aus den Grundgesamtheiten ziehen – das ist praktisch aber meist nicht durchführbar (schon allein, weil die Größe der Grundgesamtheit oft unbekannt und die Auswahl zugänglicher Texte stark eingeschränkt ist) …
- … weshalb wir schließlich Korpora verwenden oder selbst zusammenstellen, die möglichst repräsentativ für Y und Z sein sollen und als Stichproben fungieren.
- Beim Vergleich der Worthäufigkeiten in den beiden Korpora ist nun ein statistischer Test sinnvoll.

Überprüfung von Hypothesen mit statistischen Tests

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Vorgehen (1)

- Hypothese muss festgelegt sein:
  Wort X kommt in Y und Z verschieden häufig vor: F(X,Y) ≠ F(X,Z)
- andere mögliche Hypothesen:
  - X kommt in Y häufiger vor als in Z: F(X,Y) > F(X,Z)
  - X kommt in Y seltener vor als in Z: F(X,Y) < F(X,Z)
  - Dies sind **gerichtete** Hypothesen (es geht nicht nur um einen Unterschied, sondern um die **Art** des Unterschieds). Auf solche Hypothesen gehen wir hier nicht weiter ein.

Vorgehen (2)

- dann: geeigneten statistischen Test auswählen und Signifikanzschwelle α festlegen (z.B. 0,05 = 5%)
- Für unser Beispiel kommen z.B. der Chi-Quadrat-Test, der G-Test (ein Likelihood-Ratio-Test) oder der exakte Fisher-Test in Frage.
  - All diese Tests bauen darauf auf, dass die gemessenen Häufigkeiten mit den Häufigkeiten verglichen werden, die man erwarten würde, wenn es keinen Zusammenhang zwischen den Variablen gibt (in unserem Fall: wenn das Wort in Y und Z gleich häufig vorkommt).
- Test durchführen, p-Wert ermitteln und mit Signifikanzschwelle vergleichen. Wenn $p < \alpha$, ist das Ergebnis signifikant.

Um die Ecke gedacht: Null- und Alternativhypothese

- eigentliche Hypothese (= Alternativhypothese) lässt sich nicht verifizieren („Wort X kommt in Textsorte Y nicht genauso oft vor wie in Textsorte Z.“)
- daher: Versuch, das Gegenteil (das logische Komplement) der Hypothese (= Nullhypothese) zu falsifizieren („Wort X kommt in Textsorte Y genauso oft vor wie in Textsorte Z.“)
- Problem: Falsifikation ist strenggenommen natürlich ebenfalls nicht möglich …
- aber: Wir können die Wahrscheinlichkeit berechnen, dass eine Stichprobe mindestens so extrem ausfällt wie unsere, falls die Nullhypothese zutrifft (weil wir für die Nullhypothese die **Stichprobenverteilung** kennen).

Stichprobenverteilung

- Zieht man aus einer Grundgesamtheit *alle* möglichen Stichproben und betrachtet dann, wie ein bestimmter Parameter (z.B. der Mittelwert oder eine Teststatistik, also das Ergebnis eines statistischen Tests) über diese Stichproben verteilt ist, erhält man die Stichprobenverteilung (für diesen Parameter).
- Da wir mit unserer Nullhypothese annehmen, dass es keine Unterschiede in der (relativen) Häufigkeit des Wortes X in den Textsorten Y und Z gibt (dass das Wort also gleichverteilt ist), kennen wir die Grundgesamtheit (der Nullhypothese) und wissen damit auch, wie die Stichprobenverteilung aussehen muss, wenn die Nullhypothese gilt.
- Fällt der tatsächlich ermittelte Wert der Teststatistik in den sehr unwahrscheinlichen Bereich der Stichprobenverteilung, wird die Nullhypothese abgelehnt.

Statistische Signifikanz

- *p*-Wert: Wahrscheinlichkeit, bei Gültigkeit der Nullhypothese Daten zu erhalten, die mindestens so extrem sind wie die tatsächlich gemessenen
- Statistische Signifikanz: *p*-Wert liegt unter der vorher festgelegten Signifikanzschwelle α (z.B. 0,05 = 5%)
  - Nullhypothese wird als zu unwahrscheinlich abgelehnt (aber nicht falsifiziert)
  - Alternativhypothese wird angenommen (aber nicht verifiziert)
- „statistisch signifikant“ ≠ „relevant“, „wichtig“, „großer Unterschied“ o.ä.

Vorsicht mit dem *p*-Wert!

- *p*-Wert gibt lediglich die Wahrscheinlichkeit an, dass der Effekt bei der Größe der vorliegenden Stichprobe (unter Annahme der Nullhypothese) zufällig zustandegekommen ist. Signifikanz gilt nur als Rechtfertigung der Annahme der Alternativhypothese, nicht als ihr Beweis!
- Umgekehrt: Nichtsignifikanz ist kein Beweis der Nullhypothese!
- Signifikanz ≠ bedeutsames Ergebnis! Signifikanz sagt nichts darüber aus, wie **stark** die Daten von der Nullhypothese abweichen (bei einer großen Stichprobe können schon kleine Abweichungen signifikant sein).
- Besonders problematisch: In Fachzeitschriften werden oft nur signifikante Ergebnisse publiziert. Zur Illustration: http://xkcd.com/882/
- Zur Lektüre: http://www.spektrum.de/alias/umstrittene-statistik/wenn-forscher-durch-den-signifikanztest-fallen/1224727

Chi-Quadrat-Test (χ^2-Test)

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Klingt gruselig. Was ist das?

- Gruppe von Hypothesentests: Teststatistik χ^2-verteilt
- Tests für kategoriale Variablen (also nominal- oder ordinalskalierte Variablen)
- Wichtig für uns:
  - Anpassungstest/Verteilungstest: Entspricht die Verteilung einer Variable der erwarteten Verteilung? (Kommen z.B. alle Ausprägungen gleich häufig vor?)
    - Beispiel: In meiner Stichprobe der Größe 100 sind 44 Probanden männlich, 56 Probanden weiblich. Kann ich angesichts dieser Stichprobe davon ausgehen, dass es in der Grundgesamtheit gleich viele Männer und Frauen gibt?
  - Unabhängigkeitstest: Wird die Verteilung einer kategorialen Variable von einer anderen kategorialen Variable beeinflusst?
    - Beispiel: unabhängige Variable „hasst Zahlen“ (Ausprägungen: „ja“ und „nein“), abhängige Variable „verzweifelt an Statistik“ (Ausprägungen: „ja“ und „nein“)

Voraussetzungen

- unabhängige Beobachtungen
- Faustregel: mind. 80% der erwarteten Häufigkeiten ≥ 5 (ansonsten z.B. exakter Fisher-Test)
- alle erwarteten Häufigkeiten > 1

Kontingenztafeln/Kreuztabellen (1)

- beim Unabhängigkeitstest betrachten wir die Häufigkeiten der Kombinationen der Ausprägungen der unabhängigen Variable mit denen der abhängigen
- lässt sich tabellarisch darstellen:

Variable	verzweifelt an Statistik		Summe
hasst Zahlen	Ausprägung	ja	nein	
	ja	33	15	48
	nein	24	28	52
Summe		57	43	n = 100

Kontingenztafeln/Kreuztabellen (2)

- Beim Vergleich von Worthäufigkeiten gehen wir genauso vor und betrachten die gemessenen Häufigkeiten und ihre Kombinationen.
- tabellarisch:

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	50	10	60
	¬Wort X	14999950	5999990	20999940
Summe		15000000	6000000	n = 21000000

Kontingenztafeln/Kreuztabellen (3)

- Wir vergleichen diese Werte dann mit denen, die wir erwarten würden, wenn beide Variablen unabhängig voneinander wären (falls also die Nullhypothese gilt).
- erwarteter Zellenwert = Zeilensumme * Spaltensumme / n

Variable	Korpus			Summe
Wort	Ausprägung	Y	Z	
	Wort X	42,9	17,1	60
	¬Wort X	14999957,1	5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Wie genau kommen diese erwarteten Werte zustande? Und warum?

- Wenn die Nullhypothese gilt („Wort X kommt in Textsorte Y genauso oft vor wie in Textsorte Z.“), erwarten wir die gleiche relative Häufigkeit in beiden Korpora.
- Das heißt: Alle Vorkommen des Wortes X müssten gleichmäßig über die beiden Korpora verteilt sein. Etwas über 71% davon müssten im Korpus enthalten sein, das Texte der Textsorte Y enthält, etwas unter 29% im anderen – weil das erste Korpus entsprechend größer ist (beide Korpora zusammen enthalten 21 Mio. Wörter; 15 Mio. davon sind ungefähr 71%).
- Wenn das Wort X insgesamt also 60mal vorkommt, müssten 42,9 dieser Vorkommen auf das erste Korpus entfallen – knapp unter 71%. 15 Mio. / 21 Mio. * 60 ≈ 42,9
- Ergo: erwarteter Zellenwert = Spaltensumme / n * Zeilensumme

Formel (yay!)

$$\chi^2 = \sum_{i=1}^{n} \frac{(beobachtet_i - erwartet_i)^2}{erwartet_i}$$

Also muss für jede Zelle in der Kontingenztafel (von Zelle 1 bis Zelle n, in unserem Fall von 1 bis 4) folgendes getan werden:

- erwarteten Wert vom beobachteten abziehen (je größer die Abweichung, desto größer der Einfluss auf das Endergebnis)
- Ergebnis quadrieren (dadurch werden alle Werte positiv)
- Ergebnis durch erwarteten Wert teilen (dadurch wird es standardisiert)

So ergeben sich also n Werte (hier: 4), die am Ende summiert werden, sodass sich der Chi-Quadrat-Wert ergibt.

Schritt für Schritt (1)

- Erste Zelle (grün unterlegt): erwarteten Wert (E) vom beobachteten (O, für *observed*) abziehen: 50 − 42,9 = 7,1
- Ergebnis quadrieren: $7{,}1^2 = 50{,}41$
- durch erwarteten Wert teilen: 50,41 / 42,9 ≈ 1,18

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (2)

- Zweite Zelle (grün unterlegt): erwarteten Wert (E) vom beobachteten (O, für *observed*) abziehen: 10 − 17,1 = −7,1
- Ergebnis quadrieren: (−7,1)² = 50,41
- durch erwarteten Wert teilen: 50,41 / 17,1 ≈ 2,95

Variable	Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (3)

- Dritte Zelle (grün unterlegt): erwarteten Wert (E) vom beobachteten (O, für *observed*) abziehen: 14999950 − 14999957,1 = −7,1
- Ergebnis quadrieren: $(-7{,}1)^2 = 50{,}41$
- durch erwarteten Wert teilen: 50,41 / 14999957,1 ≈ 0,00 ($3{,}36 * 10^{-6}$)

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (4)

- Vierte Zelle (grün unterlegt): erwarteten Wert (E) vom beobachteten (O, für *observed*) abziehen: 5999990 − 5999982,9 = 7,1
- Ergebnis quadrieren: $7{,}1^2 = 50{,}41$
- durch erwarteten Wert teilen: 50,41 / 5999982,9 ≈ 0,00 ($8{,}40 * 10^{-6}$)

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (5)

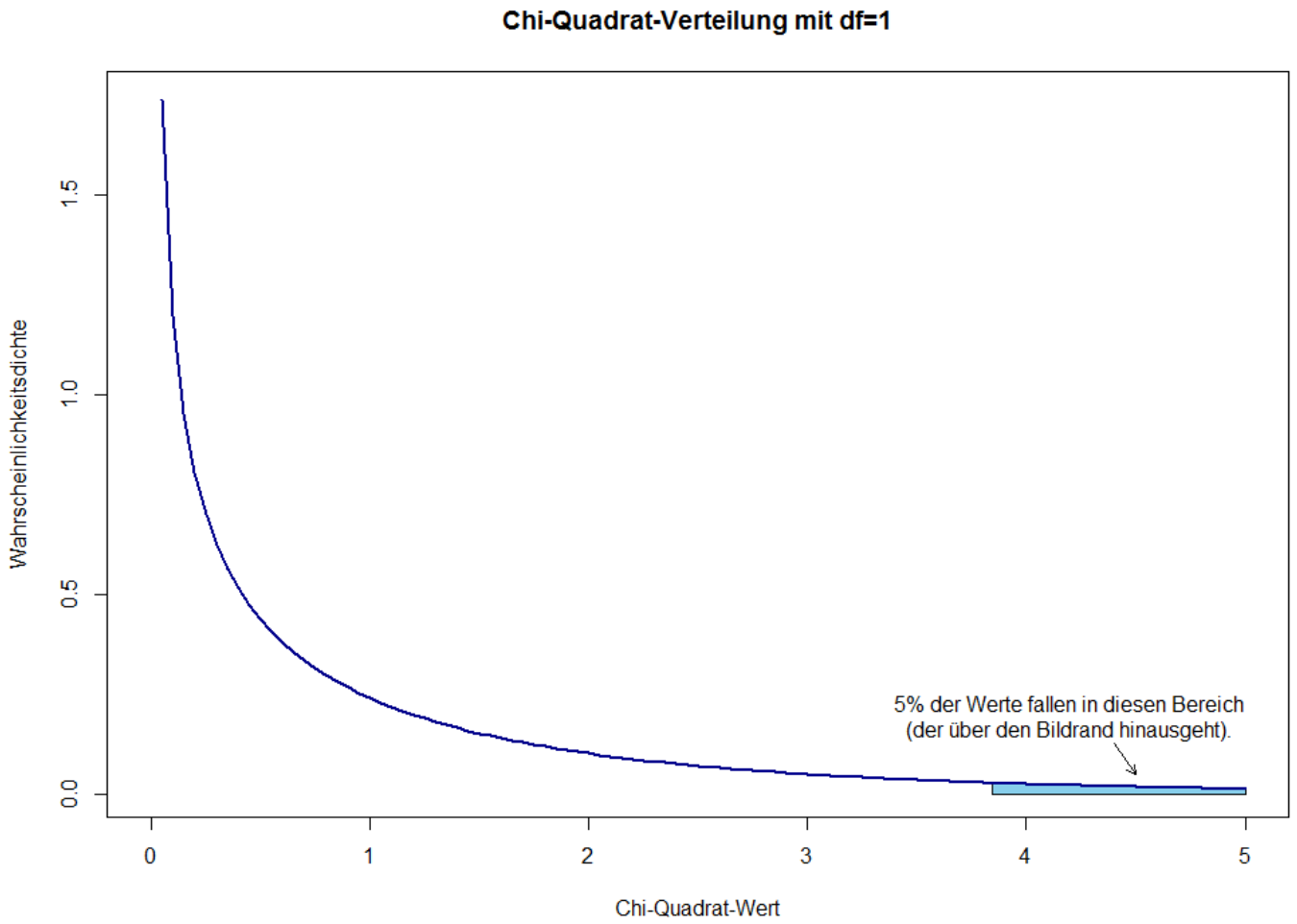
- Einzelergebnisse summieren:
  1,18 + 2,95 + 0,00 + 0,00 = 4,13
- Rechnet man sauber und rundet nicht zwischendurch (angefangen bei den erwarteten Werten), ergibt sich ein Ergebnis von 4,17.

Variable		Korpus	
Wort	Ausprägung	Y	Z
	Wort X	1,19	2,98
	¬Wort X	0,00	0,00

Schritt für Schritt (6)

- Um festzustellen, ob dieser Wert (4,17) nun statistisch signifikant ist, vergleicht man ihn mit dem Wert, über dem – bei Gültigkeit der Nullhypothese – in der Stichprobenverteilung nur ein bestimmter Prozentsatz der Werte liegt (bei α=0,05 z.B. 5%).
- Diesen Schwellenwert kann man z.B. mit geeigneter Statistik-Software errechnen oder einer Tabelle entnehmen: https://de.wikipedia.org/wiki/Chi-Quadrat-Test#Tabelle_der_Quantile_der_Chi-Quadrat-Verteilung
  Für α=0,05 und einen Freiheitsgrad (gilt generell für 2x2-Kontingenztafeln, also Tabellen mit 2 Spalten und 2 Zeilen) ergibt sich ein Schwellenwert von 3,84. Da unser ermittelter Wert von 4,17 darüber liegt, ist er signifikant ($p < .05$).

Stichprobenverteilung



Effektgröße

- Chi-Quadrat-Wert ist abhängig von der Stichprobengröße und sagt uns daher leider nichts über die Effektgröße
- verschiedene Möglichkeiten:
  - Korrelationskoeffizient φ (Phi) für Vierfeldertabellen
  - Cramérs V (entspricht bei Vierfeldertabellen φ)
  - Quoten- oder Chancenverhältnis (*odds ratio*)

G-Test

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

G-Test (Log-Likelihood-Test)

- generell sinnvoller als Chi-Quadrat-Test, aber noch immer weniger verbreitet (außer z.B. in der Computer- und Korpuslinguistik), da von Hand etwas schwieriger zu berechnen
- Chi-Quadrat-Test ist eigentlich nur eine Approximation des *G*-Tests
- bei sehr kleinen Stichproben ist der exakte Fisher-Test vorzuziehen
- lässt sich ebenso wie der Chi-Quadrat-Test als Anpassungs- oder Unabhängigkeitstest verwenden
- Teststatistik ebenfalls chi-quadrat-verteilt
- Formel:

$$G = 2 \cdot \sum_{i=1}^{n} beobachtet_i \cdot \ln\left(\frac{beobachtet_i}{erwartet_i}\right)$$

Schritt für Schritt (1)

- Erste Zelle (grün unterlegt): beobachteten Wert (O) durch erwarteten (E) teilen: 50 / 42,9 ≈ 1,17
- natürlichen Logarithmus des Ergebnisses berechnen: ln(1,17) ≈ 0,16
- mit beobachtetem Wert multiplizieren: 0,16 * 50 = 8

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (2)

- Zweite Zelle (grün unterlegt): beobachteten Wert (O) durch erwarteten (E) teilen: 10 / 17,1 ≈ 0,58
- natürlichen Logarithmus des Ergebnisses berechnen: ln(0,58) ≈ −0,54
- mit beobachtetem Wert multiplizieren: −0,54 * 10 = −5,4

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (3)

- Dritte Zelle (grün unterlegt): beobachteten Wert (O) durch erwarteten (E) teilen: 14999950 / 14999957,1 ≈ 1,00
- natürlichen Logarithmus des Ergebnisses berechnen: ln(1) = 0
- mit beobachtetem Wert multiplizieren: bleibt 0

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (4)

- Vierte Zelle (grün unterlegt): beobachteten Wert (O) durch erwarteten (E) teilen: 5999990 / 5999982,9 ≈ 1,00
- natürlichen Logarithmus des Ergebnisses berechnen: ln(1) = 0
- mit beobachtetem Wert multiplizieren: bleibt 0

Variable		Korpus		Summe
Wort	Ausprägung	Y	Z	
	Wort X	O: 50 E: 42,9	O: 10 E: 17,1	60
	¬Wort X	O: 14999950 E: 14999957,1	O: 5999990 E: 5999982,9	20999940
Summe		15000000	6000000	n = 21000000

Schritt für Schritt (5)

- Einzelergebnisse summieren und verdoppeln:
  (8 − 5,4 + 0 + 0) * 2 = 5,2
- Rechnet man sauber und rundet nicht zwischendurch (angefangen bei den erwarteten Werten), ergibt sich ein Ergebnis von 4,64.

Variable		Korpus	
Wort	Ausprägung	Y	Z
	Wort X	7,71	−5,4
	¬Wort X	−7,14	7,14

Schritt für Schritt (6)

- Mit dem ermittelten Wert (4,64) verfährt man nun genauso wie mit einem Chi-Quadrat-Wert:
  - Um festzustellen, ob der Wert statistisch signifikant ist, vergleicht man ihn mit dem Wert, über dem – bei Gültigkeit der Nullhypothese – in der Stichprobenverteilung nur ein bestimmter Prozentsatz der Werte liegt.
  - https://de.wikipedia.org/wiki/Chi-Quadrat-Test#Tabelle_der_Quantile_der_Chi-Quadrat-Verteilung
    Für α=0,05 und einen Freiheitsgrad ergibt sich ein Schwellenwert von 3,84. Da unser ermittelter Wert von 4,64 darüber liegt, ist er signifikant ($p < .05$).