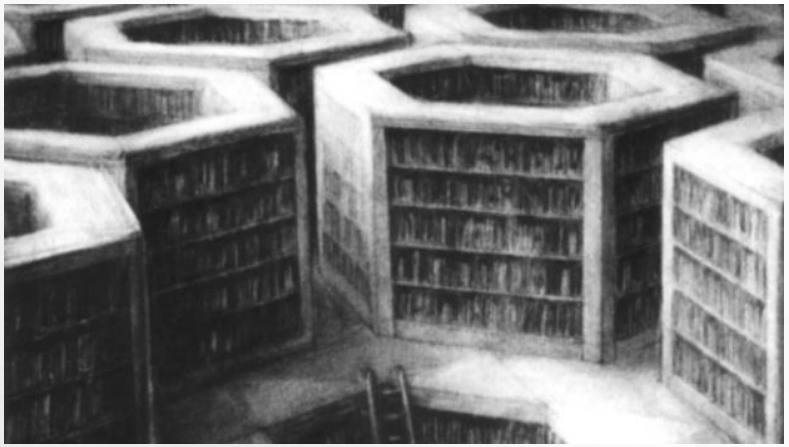
INSTITUT FÜR DEUTSCHE SPRACHE



Alexander Koplenig, Abteilung: Lexik / Projekt: Empirische Methoden

STATISTISCHE SIGNIFIKANZTESTS IN DER KORPUSLINGUISTIK

Mitglied der Leibniz-Gemeinschaft

INFERENZSTATISTIK

- Schließen von Eigenschaften einer Stichprobe auf entsprechende Eigenschaften der Grundgesamtheit
- Mittels wahrscheinlichkeitstheoretischer Überlegungen werden Stichprobenergebnisse (z.B. deskriptive Parameter oder Zusammenhänge) verallgemeinert

GRUNDGESAMTHEIT

- entspricht allen durch eine empirische Fragestellung umgrenzten Einheiten
- Beispiel: Personen, Gruppen, Sätze, Wörter, Argumentstrukturmuster, Korpora
- Um kohärente Aussagen über eine Grundgesamtheit machen zu können, muss man diese zunächst genau definieren

STICHPROBE

Stichprobe als Teilgesamtheit

Um verlässliche Aussagen treffen zu können, sollte die Stichprobe ein möglichst getreues Abbild der Grundgesamtheit sein

Möglichst keine grundsätzlichen Unterschiede zur Grundgesamtheit

STICHPROBE

- Optimal: Zufallsstichprobe
  - Stichprobe unterscheidet sich gegenüber der Grundgesamtheit nur zufällig
  - mit Hilfe theoretischer Überlegungen kann bestimmt werden, mit welcher Wahrscheinlichkeit bestimmte Verzerrungen auftreten
  - quantifizierbar durch Konfidenzintervall

INDUKTIVE STATISTIK

- Angenommen, man würde die Stichprobenziehung *m*-mal wiederholen, dann entstünden *m* unterschiedliche Stichproben
- Für jede dieser *m* Stichproben könnte man einen Mittelwert berechnen
- Diese Maßzahl ist dann das Ergebnis eines Zufallsprozesses

INDUKTIVE STATISTIK

- Über das Gesetz der großen Zahlen weiß man, dass der Erwartungswert von X gleich μ ist [$E(x)=\mu$]
- Über den Hauptsatz der Statistik weiß man, dass die Stichprobenmittelwerte normalverteilt sind [$X \sim N(\mu,\sigma^2)$]

INDUKTIVE STATISTIK

- Mithilfe dieser Information kann man einen Wertebereich bestimmen, in den ein Stichprobenmittelwert mit bestimmter Wahrscheinlichkeit fallen wird (**Konfidenzintervall)**
- Interpretation: Der wahre Mittelwert μ der Variable *X* liegt mit einer Wahrscheinlichkeit 1- α innerhalb der spezifizierten Grenzen (α: Irrtumswahrscheinlichkeit, z.B. 5%)

TESTTHEORIE

- Ausgehend von einer Nullhypothese H_0 über die Lage eines Parameters wird geprüft, ob sich die vorliegenden Daten noch im *Bereich des Wahrscheinlichen* befinden
- Liegt der Parameter außerhalb des Konfidenzintervalls, wird die Nullhypothese H_0 zugunsten der Alternativhypothese H_1 verworfen

TESTTHEORIE

- Aussage: Es besteht eine Wahrscheinlichkeit von mindestens 1-α (z.B. 95%), dass H_0 keine Gültigkeit besitzt
- = Signifikanztest
- signifikantes Ergebnis: empirische Daten stehen in *signifikantem* Gegensatz zur Nullhypothese

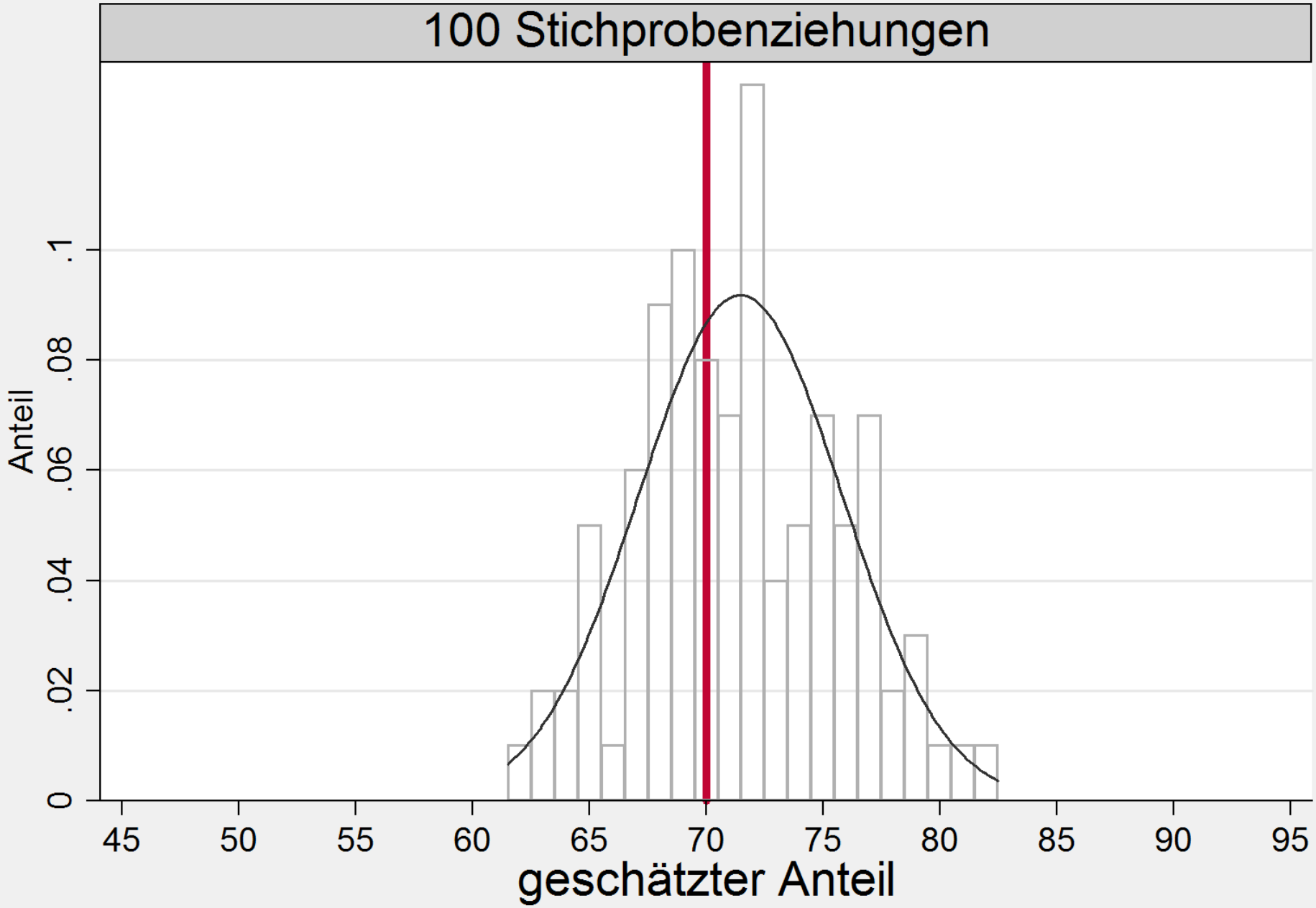
ERLÄUTERUNG: P-WERT

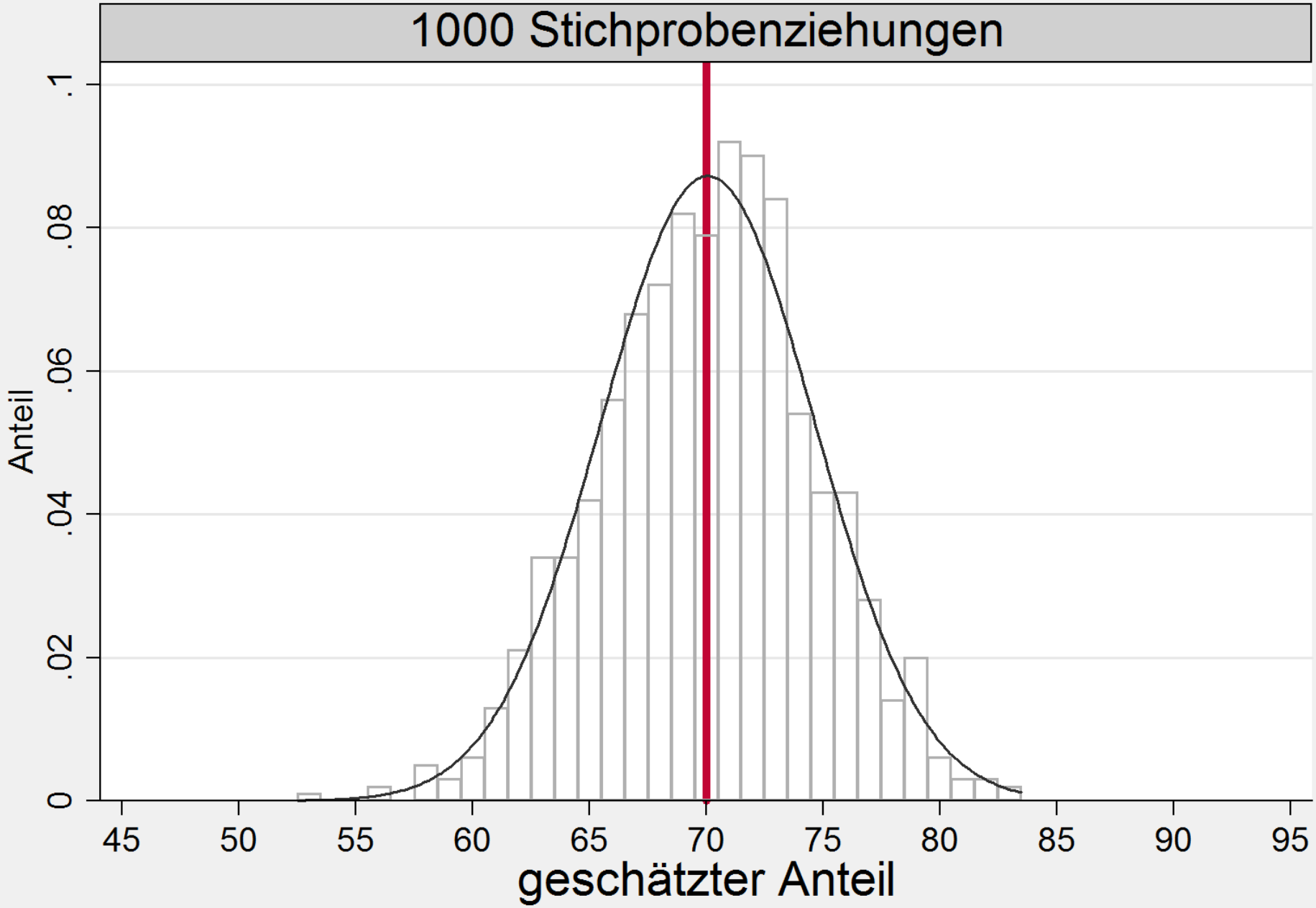
- empirisches Signifikanzniveau
- p-Wert entspricht dem Signifkanzniveau, bei dem die Nullhypothese gerade nicht abgelehnt werden würde
- Beispiel p = 0,0011
- Interpretation: Unter der Voraussetzung, dass die Nullhypothese zutrifft (z.B. kein Zusammenhang) beträgt die Wahrscheinlichkeit, dass sich der mit der Stichprobe errechnete Wert ergibt, 0,11%

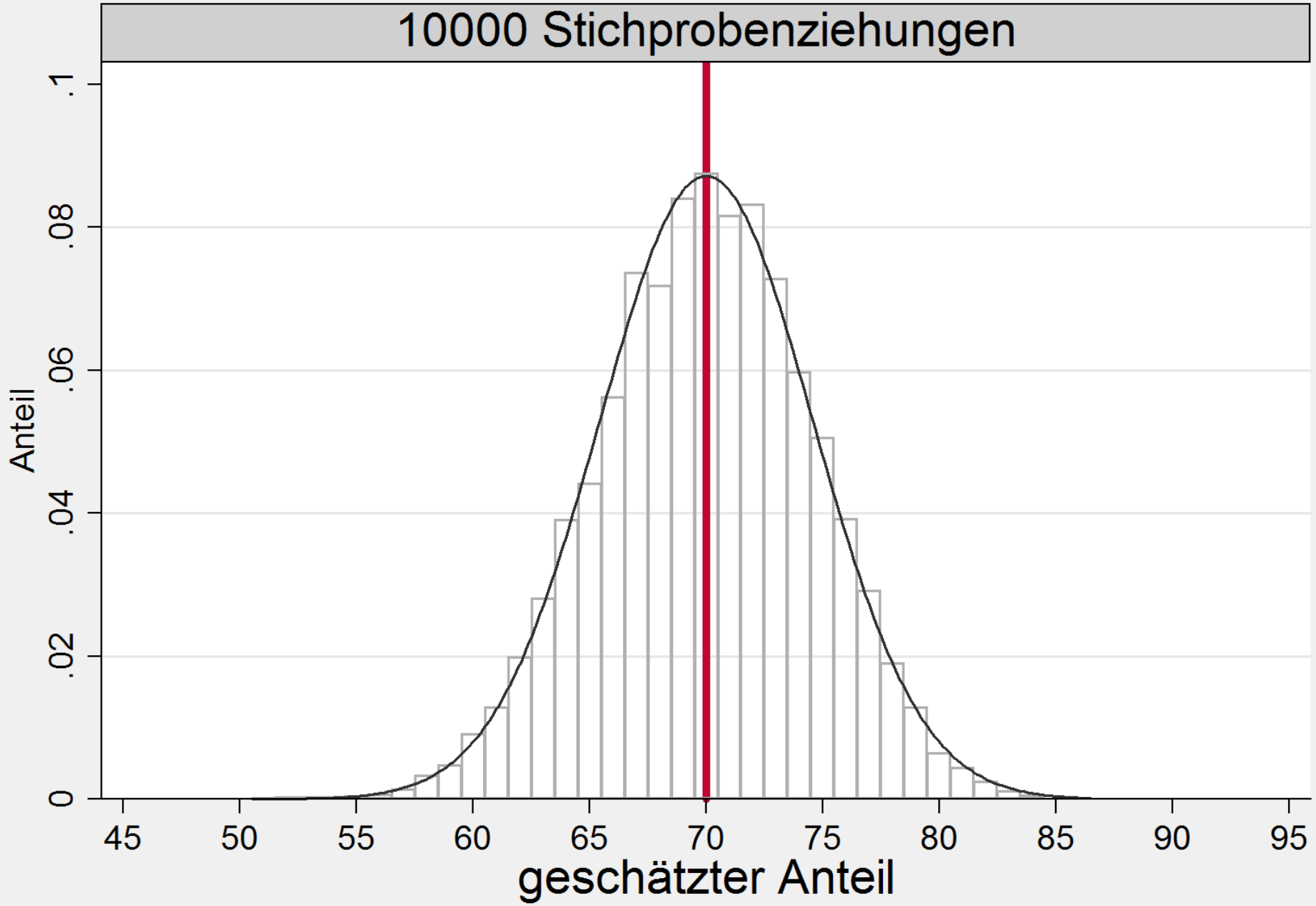
BEISPIEL

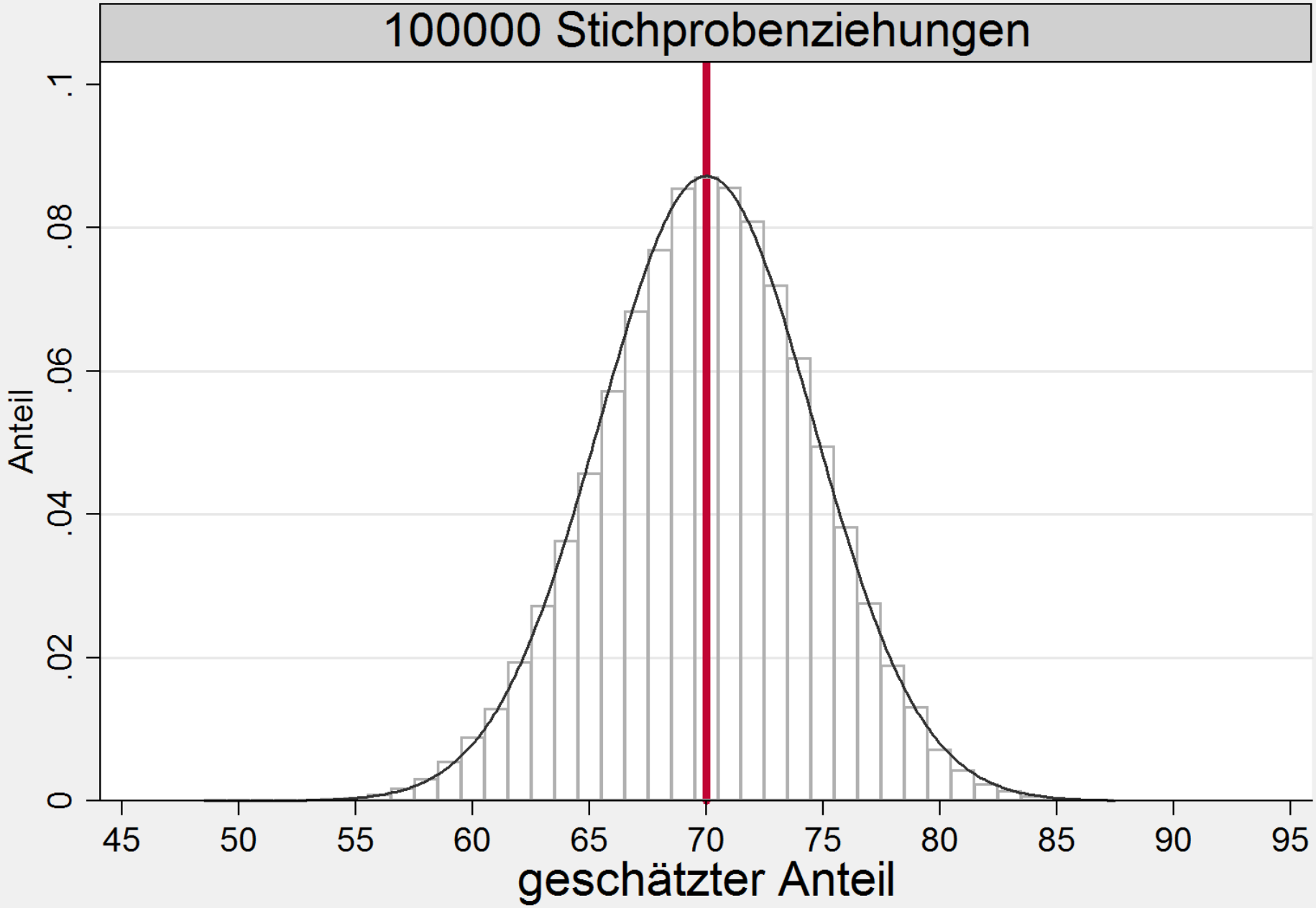
- Annahmen:
  - Grundgesamtheit besteht aus 1 Millionen Elementen
  - 70% der Elemente haben eine bestimmte Eigenschaft
- Stichprobenziehung: 100 Elemente
- Ergebnis: 68 von 100 haben die Eigenschaft

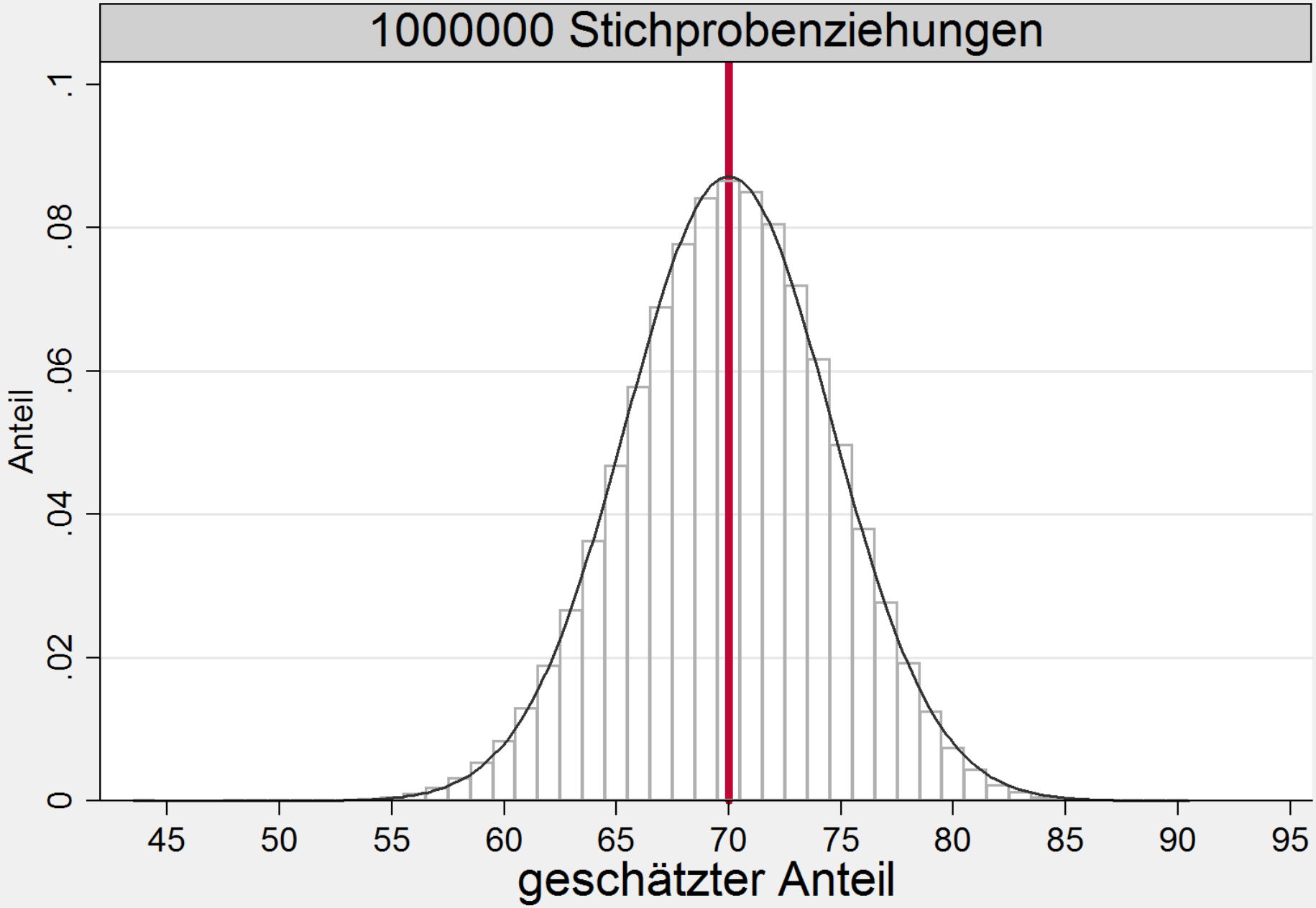
- Wiederholung der Stichprobenziehung


100 Stichprobenziehungen
Anteil
0
.02
.04
.06
.08
.1
45
50
55
60
65
70
75
80
85
90
95
geschätzter Anteil


1000 Stichprobenziehungen
Anteil
0
.02
.04
.06
.08
.1
45
50
55
60
65
70
75
80
85
90
95
geschätzter Anteil


10000 Stichprobenziehungen
Anteil
0
.02
.04
.06
.08
.1
45
50
55
60
65
70
75
80
85
90
95
geschätzter Anteil


100000 Stichprobenziehungen
Anteil
0
.02
.04
.06
.08
.1
45
50
55
60
65
70
75
80
85
90
95
geschätzter Anteil


1000000 Stichprobenziehungen
Anteil
0
.02
.04
.06
.08
.1
45
50
55
60
65
70
75
80
85
90
95
geschätzter Anteil

BEISPIEL

Ergebnisse:

- Mittelwert: 70.00
- Minimum: 44
- Maximum: 90
- In 90% der Stichprobenziehungen ergibt sich ein Wert zwischen 62 und 77

KORPUSLINGUISTIK

- Schmid (2010): Relative Vorkommenshäufigkeit bestimmter sprachlicher Strukturen instanziiert kognitive Verankerung (*From-corpus-to-cognition*-principle):
  - Beispiel: Struktur X tritt in einem Korpus häufiger auf als Struktur Y
  - Folgerungen:
    - Struktur X kommt im Diskurs ein wichtigerer Status zu als Struktur Y
    - Struktur X ist kognitiv stärker verankert

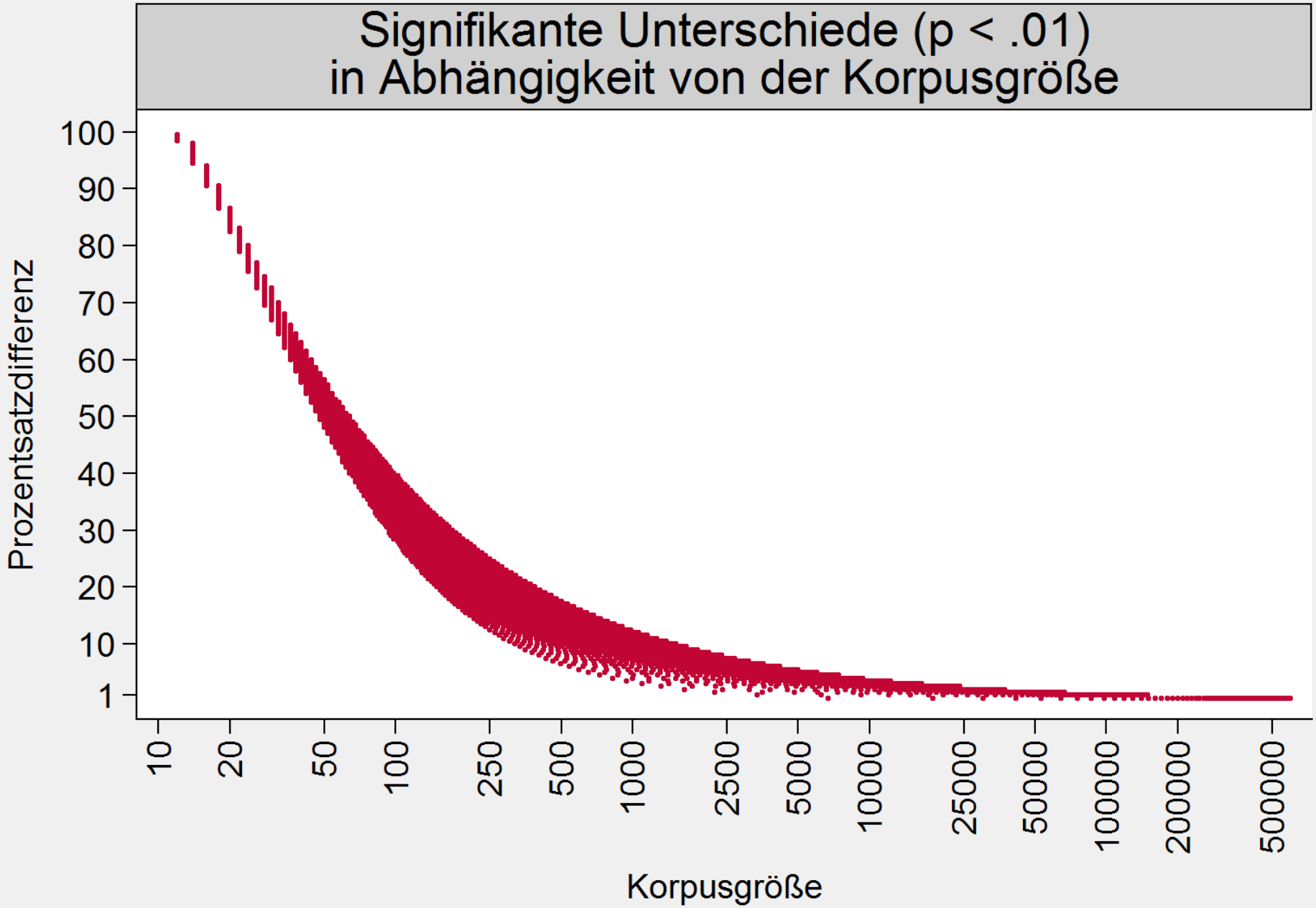
KORPUSLINGUISTIK: ZIELE

- Weniger: Aussagen über die untersuchten Korpora
- Eher: Aussagen über die sprachlichen Varietäten, die das jeweilige Korpus als Ausschnitt repräsentieren soll (Baroni & Evert, 2009, S. 2)
- Weniger: statistische Kennwerte ausrechnen
- Eher: etwas über ein linguistisches Phänomen herausfinden (Evert 2006, S. ii)

INTERPRETATION SIGNIFIKANZTEST

- = Ergebnis „nur“ im statistischen Sinn bedeutend
- ≠ substanzwissenschaftlich bedeutend (Fahrmeir, Künstler, Pigeot, & Tutz, 2001, S. 407)
- **Wichtig**: Ob sich ein Unterschied als statistisch signifikant erweist, hängt neben der Größe des Unterschieds vor allem von der Größe der Stichprobe ab

→ je größer das Korpus, desto eher ist ein Zusammenhang signifikant


Signifikante Unterschiede (p < .01)
in Abhängigkeit von der Korpusgröße
Prozentsatzdifferenz
100
90
80
70
60
50
40
30
20
10
1
10
20
50
100
250
500
1000
2500
5000
10000
25000
50000
100000
200000
500000
Korpusgröße

INTERPRETATION SIGNIFIKANZTEST

- Beispiel: Prozentsatzdifferenz. Anteil in Korpus 1 50,0**0**% *vs.* Korpus 2 50,0**5**% ist hochsignifikant wenn man eine Korpusgröße von (K1+K2) ≈ 600.000 zu Grunde legt

→ Ab einer bestimmten Korpusgröße ist jeder (noch so triviale) Unterschied hochsignifikant

KORPORA UND REPRÄSENTATIVITÄT

- Idealfall (Perkuhn, Keibel und Kupietz (2012), S. 46-47): Korpus soll Schlussfolgerungen und Verallgemeinerungen zulassen
- Voraussetzung: Stichprobe repräsentativ für die Grundgesamtheit
- Problem: Sprache und deren Proportionen nicht allgemeingültig definierbar

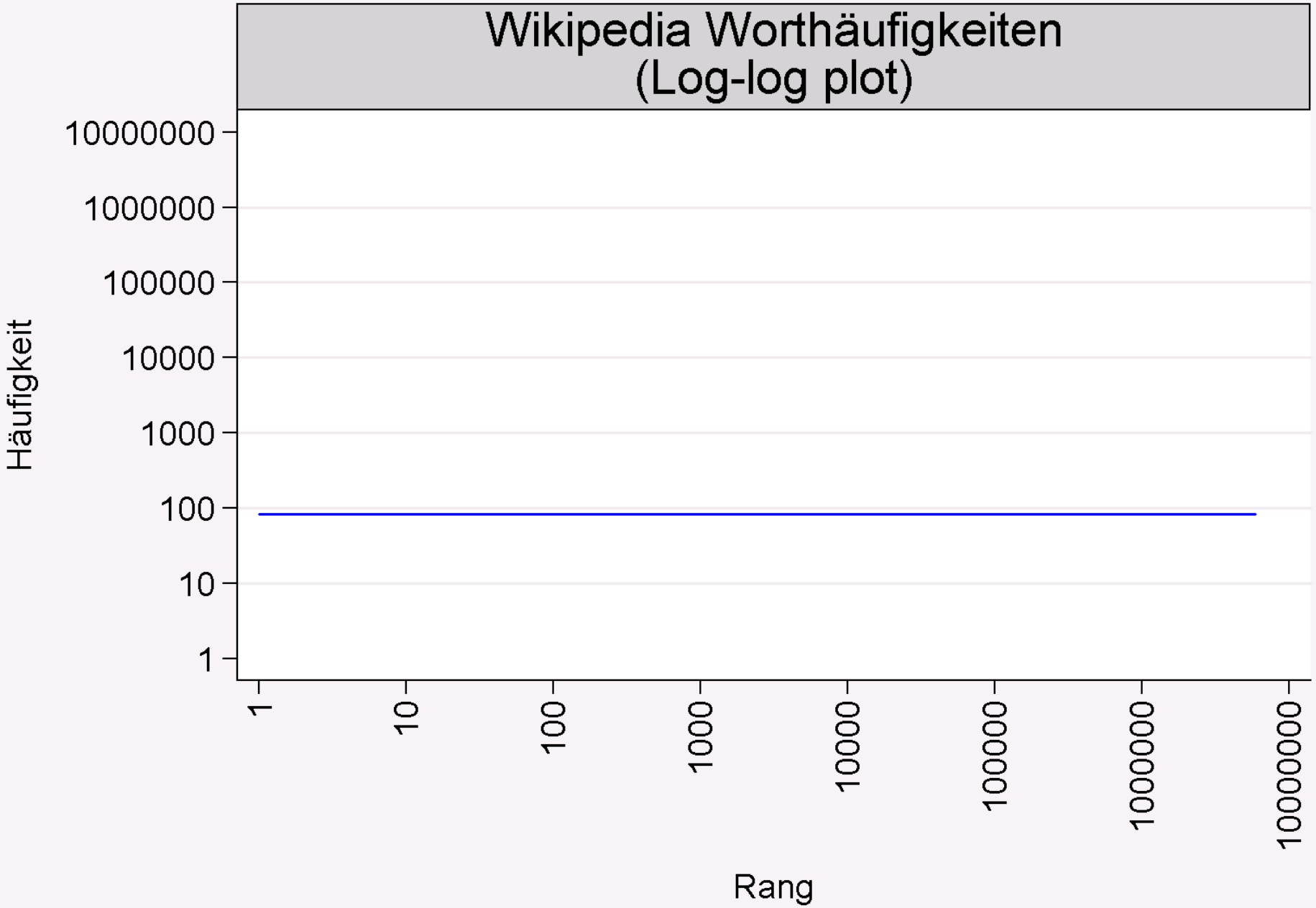
KORPORA UND REPRÄSENTATIVITÄT

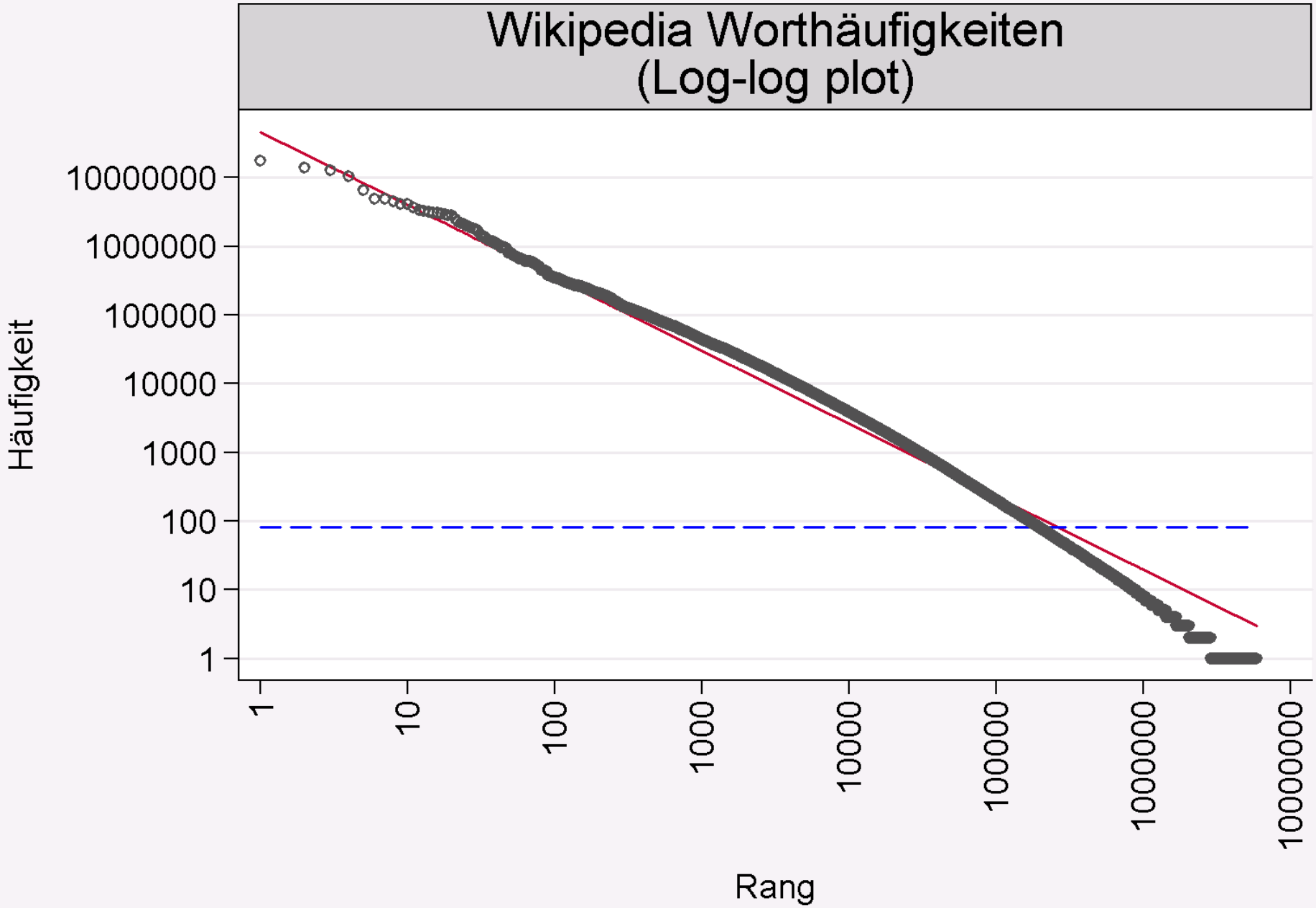
- Vorschlag von Perkuhn und Kollegen: Um „die Repräsentativität zu approximieren oder zumindest einschätzbar zu machen, wird bei der Korpuskomposition meist die Verteilung bezüglich solcher Dimensionen kontrolliert, die man (i) für die Sprachdomäne **intuitiv als relevant** erachtet (bzw. die voraussichtlich Auswirkungen auf Befunde haben werden) und (ii) **mit vertretbarem Aufwand in Erfahrung** bringen kann.“

  (S.47, meine Hervorhebungen)

ZUFALL UND KORPORA

- Evert (2006): Zufallsstichproben im Bereich von natürlichen Sprachen unrealistisch
  - Wörter kommen nicht zufällig vor (Kilgarriff 2005)
  - Beispiel: Korpus aus allen Wikipedia Artikeln (Stand 2011)
  - fast eine Halbe Milliarde Wortformtoken und fast 6 Millionen Wortformtypen
  - Bei Gleichverteilung würde man erwarten, dass jedes Wort ungefähr (487.107.690/5.903.710) = 83 Mal vorkommt


Wikipedia Worthäufigkeiten
(Log-log plot)
Häufigkeit
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
1000000
10000000
Rang

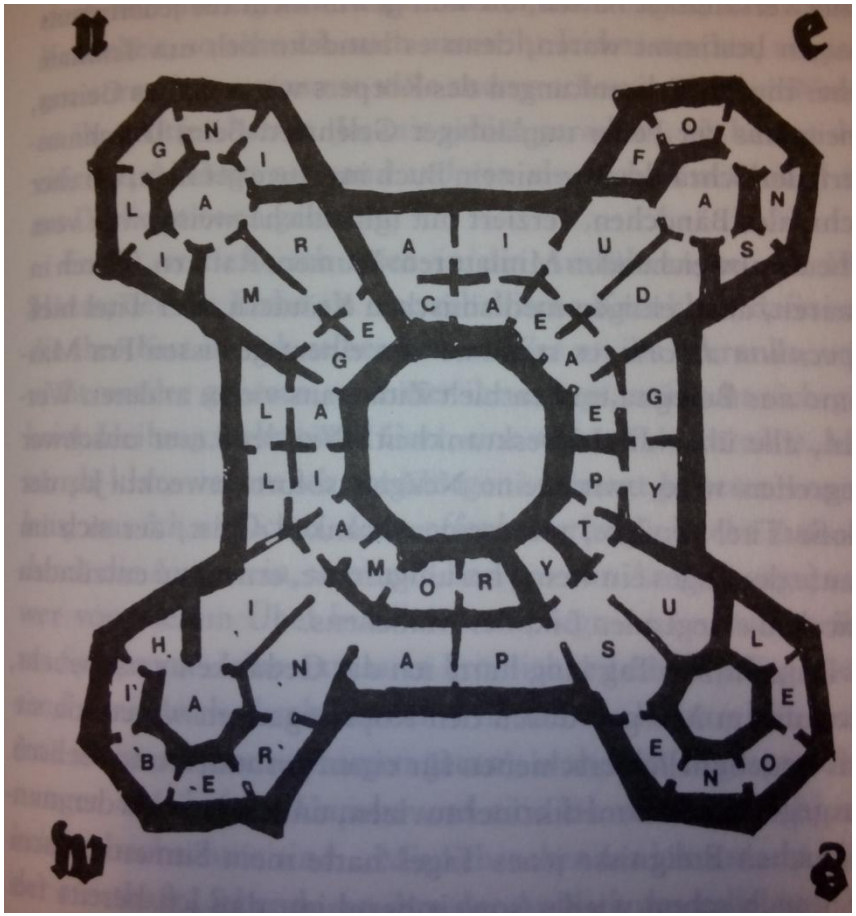

Wikipedia Worthäufigkeiten
(Log-log plot)
Häufigkeit
10000000
1000000
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
1000000
10000000
Rang

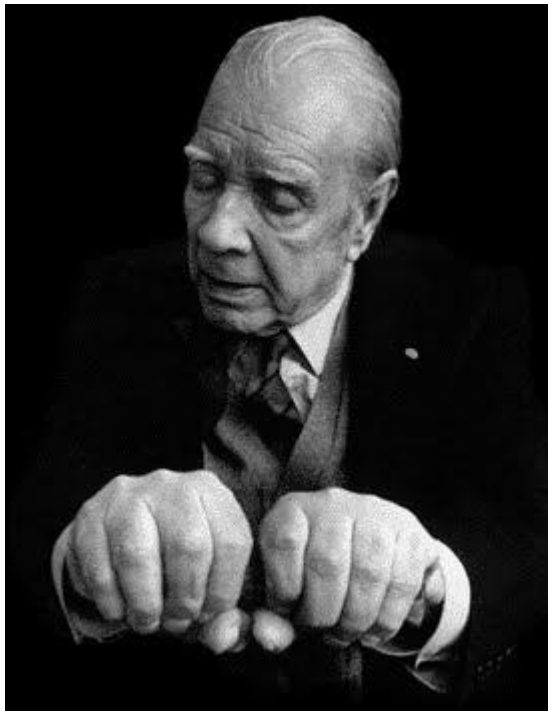
ZUFALL UND KORPORA

- Auch Sätze werden nicht zufällig gebildet
- Frage: Wo ist der Zufall?
- Antwort von Evert (2006): Bibliotheksmetapher
  - Gigantisch (unendlich) große Bibliothek, welche die Sprache in ihrer Gesamtheit repräsentiert
  - Jedes Buch in der Bibliothek stellt ein Fragment dar
  - Korpus als zufällige Auswahl eines Buches von einem der vielen Regale (oder noch besser als einzelne Sätze, Phrasen, Wörter)
- Wie sieht diese Bibliothek aus?

DIE BIBLIOTHEK VON BABEL

Borges (1941): „Das Universum (das andere die Bibliothek nennen) setzt sich aus einer unbegrenzten und vielleicht unendlichen Zahl sechseckiger Galerien zusammen .“





Eco (1995) : „Nehmen wir also an, alle Bücher aller großen Bibliotheken würden mit einem Scanner aufgenommen (…). Das würde heißen, ihr ganzer Inhalt, samt Typographie und Seitenumbruch, würde in das Gedächtnis eines zentralen Computers eingespeist.“

PROBLEME DIESER SICHTWEISE

- Manche Texte prinzipiell nicht in einem Korpus (und auch nicht der Bibliothek):
  - Beispiel: intime Gespräche zwischen Ehepartnern oder Diplomaten/-innen.
  - Bei Aufnahme mit Einverständnis der Beteiligten: Problem der Reaktivität (Diekmann, 2002, S. 520–523)

- Zeitungstexte als Hauptbestandteil vieler Korpora:
  - Redaktionelle Bearbeitung vor Veröffentlichung nach festgelegten Regeln
  - Deshalb: nur bedingt prototypischer Ausschnitt der geschriebenen Sprache (Gries & Berez, noch nicht erschienen, S. 2).

BALANCIERTHEIT

- BNC enthält 10% gesprochene Sprache, 90% geschriebene Sprache, davon fast 19% „Imaginative“ auf der Wortebene bzw. 27.10% auf der Satzebene, Rest: „Informative“.
  - Anmerkung: Hier sieht man wieder: Sprache ist nicht zufällig, Problem der „unit of measurement“ (Evert 2006: x)
- Beispiel: Angenommen die Bibliothek enthält nicht nur alles (wirklich alles) an geschriebener Sprache, sondern auch komplette Transkriptionen der gesprochenen Sprache

BALANCIERTHEIT

- Ein Gedankenexperiment: Wir interessieren uns für ein relativ seltenes sprachliches Phänomen
  - hierfür gibt es in der Abteilung „geschriebene Sprache“ unserer Bibliothek (also insgesamt in der geschriebenen Sprache) 580 Belege.
  - In der Abteilung „gesprochene Sprache“ finden sich dagegen 758 Belege
- Die Wege in der Bibliothek sind weit, also können wir aus pragmatischen Gesichtspunkten lediglich 300 der insgesamt (580+758) = 1338 Belege auswählen.

BALANCIERTHEIT

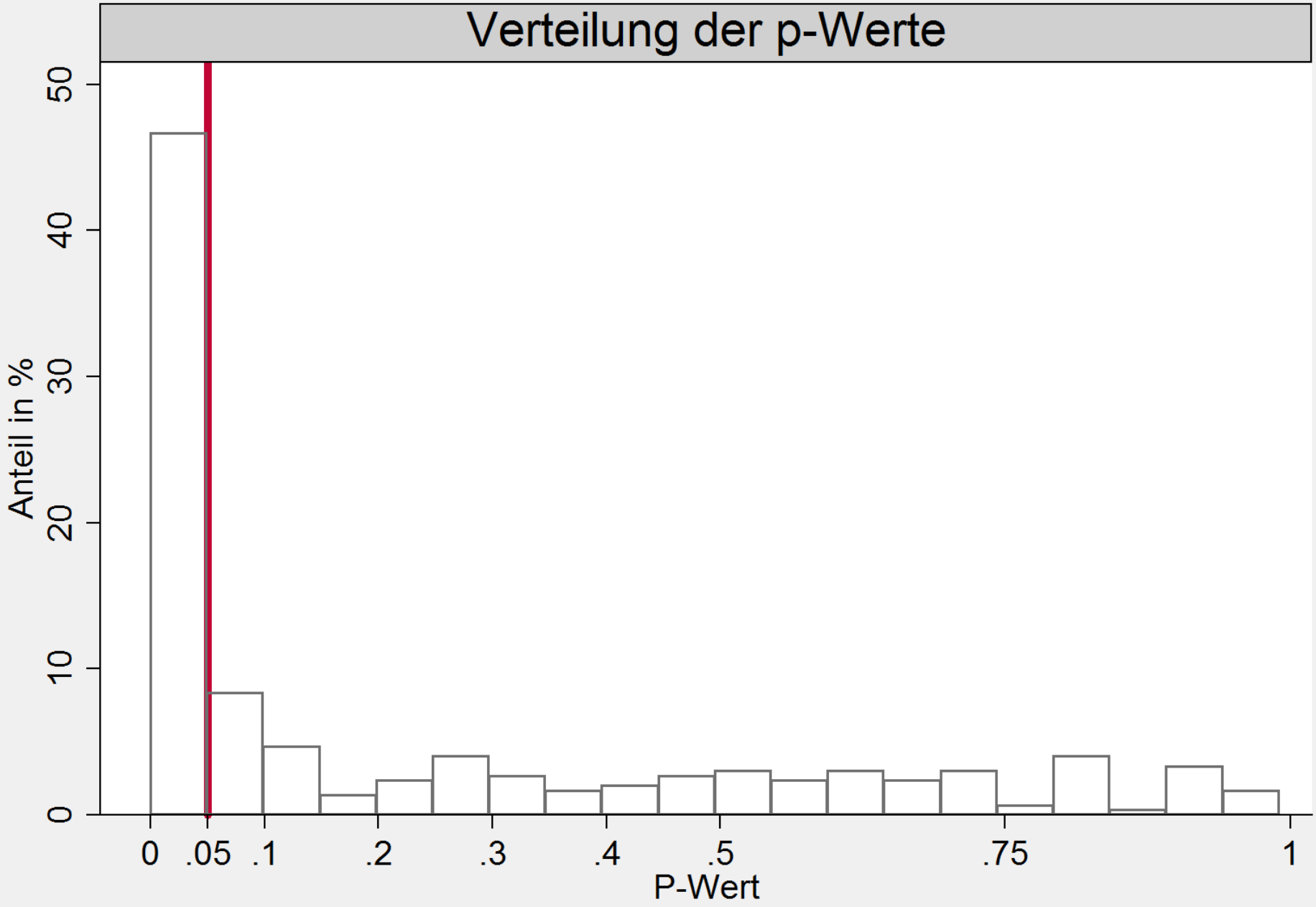
- BNC Sampling (90 % geschriebene Sprache, also 270 Belege / 10 % gesprochene Sprache, also 30 Belege)

	Y = 0	Y = 1	Gesamt
X = 0	47,86	28,13	**37,33**
X = 1	52,14	71,88	**62,67**
Gesamt	**100,00**	**100,00**	**100,00**

- Unterschied hochsignifikant $p < .0005$ ($X^2 = 12.43$)

BALANCIERTHEIT

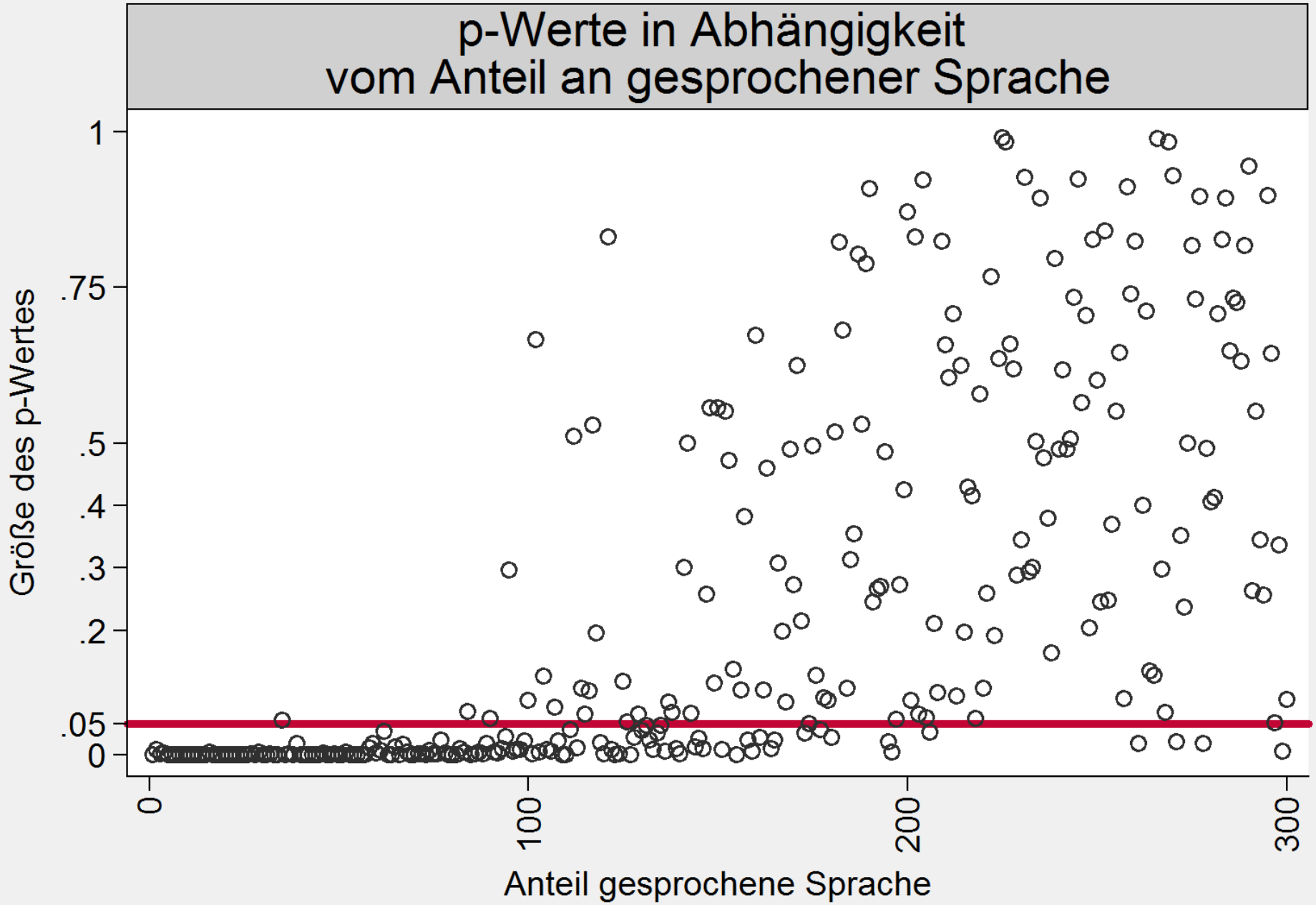
- Simulation: Wiederholung der Stichprobenziehungen
  - *0 Elemente aus der gesprochenen Abteilung, 300 aus der geschriebenen Abteilung*
  - *1 Elemente aus der gesprochenen Abteilung, 299 aus der geschriebenen Abteilung*
  - *2 Elemente aus der gesprochenen Abteilung, 298 aus der geschriebenen Abteilung*
  - *…*
  - *300 Elemente aus der gesprochenen Abteilung, 0 aus der geschriebenen Abteilung*
- Bei jeder Ziehung: Notiere p-Wert


Verteilung der p-Werte
Anteil in %
0
10
20
30
40
50
0
.05
.1
.2
.3
.4
.5
.75
1
P-Wert

BALANCIERTHEIT

- Ergebnisse:
  - je nachdem, wieviel gesprochene Sprache man in der Stichprobe zulässt, verändert sich das Ergebnis völlig
  - In 53% aller Mischungen ist der p-Wert größer als 0.05, also nicht signifikant
  - Was nun? Mittlerer p-Wert? Wie?

- **Statistik hilft nicht dabei zu entscheiden, welches Ergebnis am ehesten der „Wahrheit“ entspricht**


p-Werte in Abhängigkeit
vom Anteil an gesprochener Sprache
Größe des p-Wertes
1
.75
.5
.4
.3
.2
.05
0
0
100
200
300
Anteil gesprochene Sprache

FAZIT

„So gesehen ist die Sprechweise von repräsentativen Stichproben bzw. Korpora nicht etwa nur deswegen ungeeignet, weil sie unbestimmt und/oder notwendig zirkulär wäre, sondern sie ist vor allem deswegen als verfehlt zu verwerfen, weil sie außerstande setzt, das im wahrscheinlichkeitstheoretischen Begründungs-zusammenhang statistischen Schließens und Argumentierens vorausgesetzte *Kriterium der Zufälligkeit* zu erkennen und für das Verfahren der Korpusbildung zu fordern. “

(Rieger, **1979**, S. 68, Hervorhebungen im Orginal)

INSTITUT FÜR DEUTSCHE SPRACHE

VIELEN DANK

Mitglied der Leibniz-Gemeinschaft

LITERATUR

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton NJ: Princeton University Press.

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baroni, M., & Evert, S. (2009). Statistical methods for corpus exploitation. In A. Lüdeling & M. Kytö (Hrsg.), *Corpus linguistics: An international handbook* (Bd. 2, S. 777–802). Berlin: De Gruyter Mouton.

Diekmann, A. (2002). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (8. Aufl.). Reinbek: Rowohlt Taschenbuch Verlag.

Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), 177–190.

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2001). *Statistik: der Weg zur Datenanalyse ; mit 34 Tabellen*. Berlin [u.a.]: Springer.

Gaëtanelle Gilquin, & Stefan Th. Gries. (2009). Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26.

Gilquin, G. (2008). What You Think Ain't What You Get: Highly polysemous verbs in mind and language. In G. Desgulier, J.-B. Guignard, & J. R. Lapaire (Hrsg.), *Du fait grammatical au fait cognitif. From Gram to Mind* (Bd. 2). Pessace: Presses Universitaires de Bordeaux.

Gries, S. T., & Berez, A. L. (noch nicht erschienen). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Hrsg.), *Handbook of Linguistic Annotation*. Berlin, New York: Springer. Abgerufen von http://www.linguistics.ucsb.edu/faculty/stgries/research/InProgr_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot.pdf

Hans-Jörg Schmid. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In Dylan Glynn & Kerstin Fischer (Hrsg.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (S. 101–133). Berlin, New York: de Gruyter.

Jann, B. (2005). *Einführung in die Statistik*. München; Wien: Oldenbourg.

Kilgarriff, A. (2005). Language is never ever ever random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263-276.

Lüdeling, A., & Evert, S. (2005). The emergence of productive non-medical -itis. Corpus Evidence and qualitative analysis. In S. Kepser & M. Reis (Hrsg.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin, New York: De Gruyter Mouton.

Perkuhn, R., Keibel, Holger, Kupietz, Marc. (2012). *Korpuslinguistik*. Paderborn: Fink.

Rieger, B. (1979). Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In H. Bergenholtz & B. Schaeder (Hrsg.), *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora* (S. 52–70). Königsstein/ Taunus: Scriptor. Abgerufen von http://www.uni-trier.de/fileadmin/fb2/LDV/Rieger/Publikationen/Aufsaetze/79/rub79.html

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 191–233.

Schmid, H.-J. (2007). Entrenchment, salience and basic levels. In D. Geeraerts & H. Cuyckens (Hrsg.), *The Oxford Handbook of Cognitive Linguistics* (S. 117–138). Oxford: Oxford University Press.

BNC:

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#BNCcompo

Bilder:

1.Folie:

Bibliothek: http://doktorpeng.weebly.com/9/post/2013/03/short-story-friday-4-jorge-luis-borges-die-bibliothek-von-babel.html abgerufen am 19.03.2014

Borges: https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcRweNDNlbSom4CUus1io6BayibyvR-uQTkl3WupTnKrQxqYxALm abgerufen am 19.03.2014

Eco: http://2.bp.blogspot.com/_PtPNUMoshEs/S-iyffCGmAI/AAAAAAAAD_c/b6EoxKxIs4A/s320/cuar01_proust_eco0507.jpg abgerufen am 19.03.2014

29.Folie

JL Borges: http://4.bp.blogspot.com/_AtZ57cO4VaM/TVE3ToO-2bI/AAAAAAAACJA/sKevrqkEpnY/s1600/Borges%252Bbiblioteca%252Btotal%252Bdescontexto.gif abgerufen am 19.03.2014

Bibliothek: eigenes Foto aus Eco: Der Name der Rose

Nicht wissenschaftliche Literatur:

Eco, Umberto (1982) – Der Name der Rose. Deutsche Übersetzung von Burkhart Kroeber.

Eco, Umberto (1990/1995) – Gesammelte Streichholzbriefe. Deutsche Übersetzung von Burkhart Kroeber.

Borges, Jorge Luis (1941): Die Bibliothek von Babel. Online unter: http://mcn.privat.t-online.de/borgbib.htm abgerufen am 19.03.2014