dgfs
Deutsche Gesellschaft
für Sprachwissenschaft

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Tutorium der Sektion CL:

Einführung in die Statistik für Linguisten mit „R“

35. Jahrestagung der DGfS
12. März 2013

Stefan Evert (FAU Erlangen-Nürnberg)
Amir Zeldes (HU Berlin)

Worum geht's?

- Wir sind im Alltagsleben und in der Forschung von Statistik umgeben
- Quantitative Aussagen werden oft akzeptiert, ohne dass man sie wirklich versteht
- Möglichkeiten, anhand der Statistik zu neuen Erkenntnissen zu kommen, werden vernachlässigt
- Als Geisteswissenschaftler schwer einzusteigen (aus eigener Erfahrung …)

Wozu? Ein Beispiel

Osnabrück - "Der Trend zur Gewalt ist ungebrochen. Besonders die Zahl gefährlicher und schwerer Körperverletzungen ist deutlich gestiegen", sagte der Bundesvorsitzende der Gewerkschaft der Polizei (GdP), Konrad Freiberg, der "Neuen Osnabrücker Zeitung". Zwar hätten Niedersachsen, Mecklenburg-Vorpommern und das Saarland ihre Kriminalstatistiken noch nicht vorgelegt, die Tendenz für den Bund sei dennoch eindeutig.
http://www.spiegel.de/panorama/justiz/0,1518,473433,00.html

Wozu? Ein Beispiel

Osnabrück - "Der Trend zur Gewalt ist ungebrochen. Besonders die Zahl gefährlicher und schwerer Körperverletzungen ist deutlich gestiegen", sagte der Bundesvorsitzende der Gewerkschaft der Polizei (GdP), Konrad Freiberg, der "Neuen Osnabrücker Zeitung". Zwar hätten Niedersachsen, Mecklenburg-Vorpommern und das Saarland ihre Kriminalstatistiken noch nicht vorgelegt, die Tendenz für den Bund sei dennoch eindeutig.
http://www.spiegel.de/panorama/justiz/0,1518,473433,00.html

Noch ein Beispiel

Die Arzneimittelausgaben der gesetzlichen Krankenkassen steigen im kommenden Jahr voraussichtlich um 6,6 Prozent auf einen Rekordwert von mehr als 31 Milliarden Euro
http://www.sueddeutsche.de/politik/139/313047/text/

- Sie steigen sicher nicht genau um 6,6%.
- Mit welcher Wahrscheinlichkeit ist es 0,5% mehr?
- Mit welcher Wahrscheinlichkeit gibt es keinen Anstieg?

- ➢ So wie es dasteht, kann es nicht eintreffen
- ➢ Keinerlei Überprüfbarkeit

Professionelle Darstellung solcher Zahlen

Die Ergebnisse für die zwei experimentellen Bedingungen wiesen signikante Differenzen auf. Schüler, die nach der neuen Methode unterrichtet wurden, erreichten signifikant bessere Ergebnisse als die nach der traditionellen Methode unterrichteten ($t = 6,03$, $df = 13$, $p < 0.001$)

- Was ist signifikant?
- Was bedeutet t, df, und p?
- Ist das toll?

Zwischenfazit

- Seinem eigenen Gefühl kann man bei der Beurteilung von Zahlenreihen nicht trauen
- Viele empirische Fragen sind ohne Statistik schlicht nicht zu beantworten
- Ernstzunehmende Aussagen über Zahlen sind ohne Ausbildung unverständlich:
  - der eigenen Anschauung nicht zu trauen
  - sich Mittel und Wege anzueignen, in Ihrer eigenen Forschung Zahlen angemessen zu deuten
  - fremde Zahlen zu verstehen und zu beurteilen

Statistik in der Linguistik

- Was ist der Unterschied zwischen gesprochener und geschriebener Sprache?
- unterschiedlichen Textsorten? Geschlechtern? …
- Wie ähnlich werden bestimmte Wörter oder Konstruktionen gebraucht? (Was ist Gebrauch?)
- Wie produktiv sind Wortbildungsmuster im Vergleich? Ab wann ist ein Ausdruck lexikalisiert?
- Was fällt Deutschlernern besonders schwer?
- Kann man Bedeutung mit distributionellen Kriterien empirisch erschließen?
- …

Was ist R?

- Wir werden in diesem Tutorium auch mit echten Daten arbeiten und statistische Tests anwenden
- Der einzig sinnvolle Weg ist, ein professionelles Statistikprogramm zu lernen
- Hier verwenden wir R: http://cran.r-project.org/
- Text- und kommandozeilenbasiert
- Look & Feel unterscheidet sich ziemlich stark von den üblichen GUI-Programmen (SPSS, Excel, ...)
- Der Einstieg in die Arbeit mit R erscheint schwieriger

Warum R?

- Praktischer Grund: R ist frei (SPSS bspw. sehr teuer)
- Unheimlich gute graphische Möglichkeiten
- Extrem flexibel, alles mit allem kombinierbar
- Erweiterbar durch tausende von Paketen
- Läuft gleichermaßen unter Windows/Linux/Mac
- Erobert sich in der Wissenschaft mehr und mehr eine beherrschende Stellung
- Für Linguistik besonders beliebt (Module zur Verarbeitung linguistischer Daten)

Ablauf des Tutoriums

10:00-12:00 Einleitung, Häufigkeitsvergleich

Mittagspause

13:00-14:30 erste Schritte mit R + Übung 1

Kaffee-Pause

15:00-16:10 Konfidenzintervalle + Übung 2:

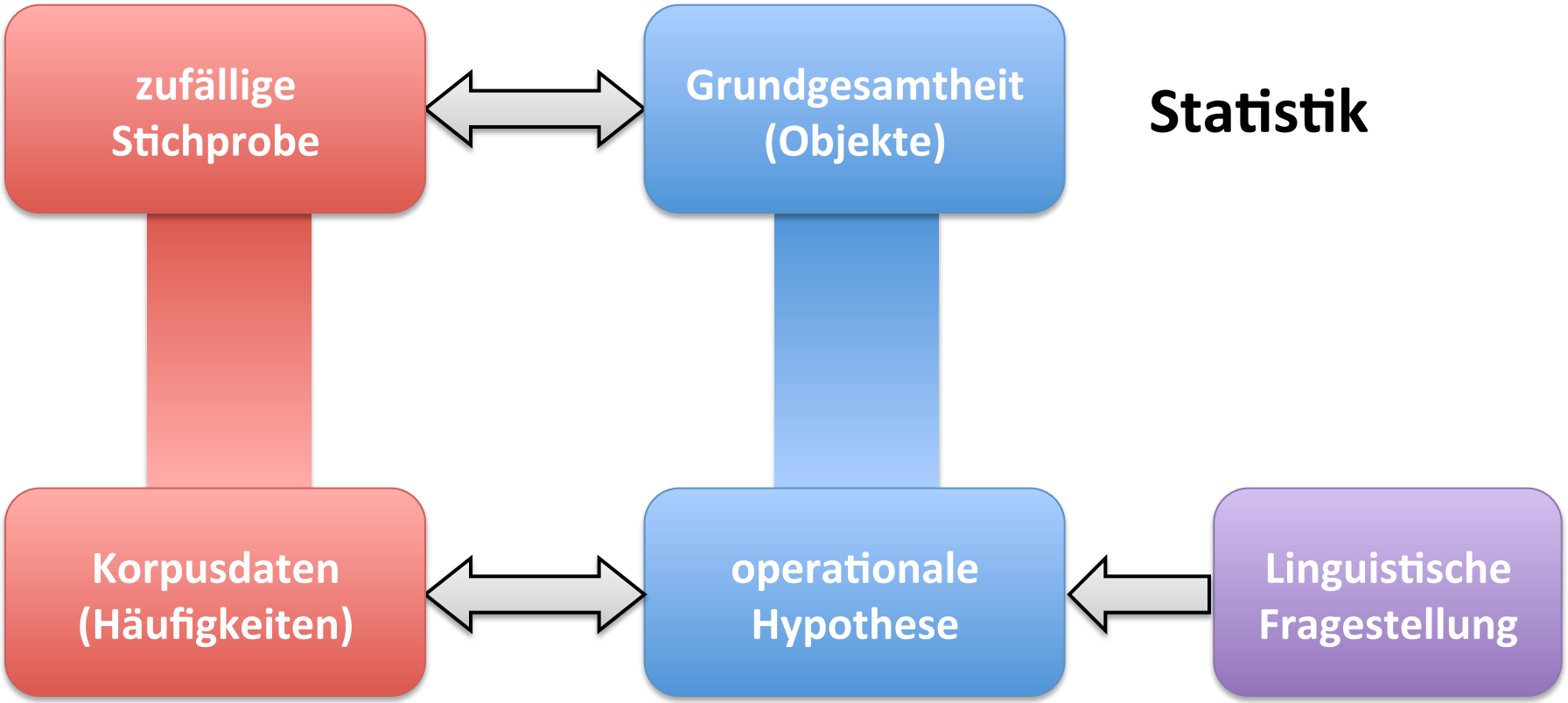
Kurze Pause

16:25-18:00 Kreuztafeln und Assoziation + Übung 3

Haben alle Teilnehmer R installiert?
http://www.r-project.org/

HÄUFIGKEITSVERGLEICH

Quantitative Korpusstudien



Ein Fallbeispiel

Linguistische Fragestellung

- Nominalkomposita bei DaF-Lernern (L2) und deutschen Muttersprachlern (L1)
- Vermutungen
  - L2 bilden weniger Komposita als L1
  - Abhängigkeit von der L1 des Lerners
- Quantitative Studie auf Basis des Lernerkorpus Falko (HU Berlin, Reznicek et al. 2010; s. Zeldes, erscheint für weitere Einzelheiten)

Ein Fallbeispiel

operationale Hypothese

- Operationalisierung:
  Die Häufigkeit von Nominalkomposita ist bei L2-Sprechern geringer als bei L1-Sprechern
- Was bedeutet „Häufigkeit“?
  - Anzahl von Nominalkomposita pro Text?
  - durchschnittliche Anzahl von NK pro Satz?
  - relative Häufigkeit von Nominalkomposita
    = Anteil von NK unter allen Substantiven

Ein Fallbeispiel

operationale Hypothese

- Operationalisierung:
  Die Häufigkeit von Nominalkomposita ist bei L2-Sprechern geringer als bei L1-Sprechern
- Was bedeutet „Häufigkeit“?
  - Anzahl von Nominalkomposita pro Text?
  - durchschnittliche Anzahl von NK pro Satz?
  - relative Häufigkeit von Nominalkomposita
    = Anteil von NK unter allen Substantiven
- In Bezug auf welche Texte? schriftl./mündl.?

zufällige Stichprobe

Ein Fallbeispiel

Korpusdaten (Häufigkeiten)

- Wir benötigen zwei Stichproben von Nomina
  1. aus Texten von deutschen Muttersprachlern
  2. aus Texten von DaF-Lernern
- Was ist eine zufällige Stichprobe?
  - zusammenhängender Text ≠ Zufallsstichprobe
  - Stichproben müssen repräsentativ (für die jeweilige Sprachvarietät) und vergleichbar sein
  - hier: Material aus Lernerkorpus Falko (deutsche Essays von L1 und L2 zu gleichen Themen)

zufällige Stichprobe

Ein Fallbeispiel

Korpusdaten (Häufigkeiten)

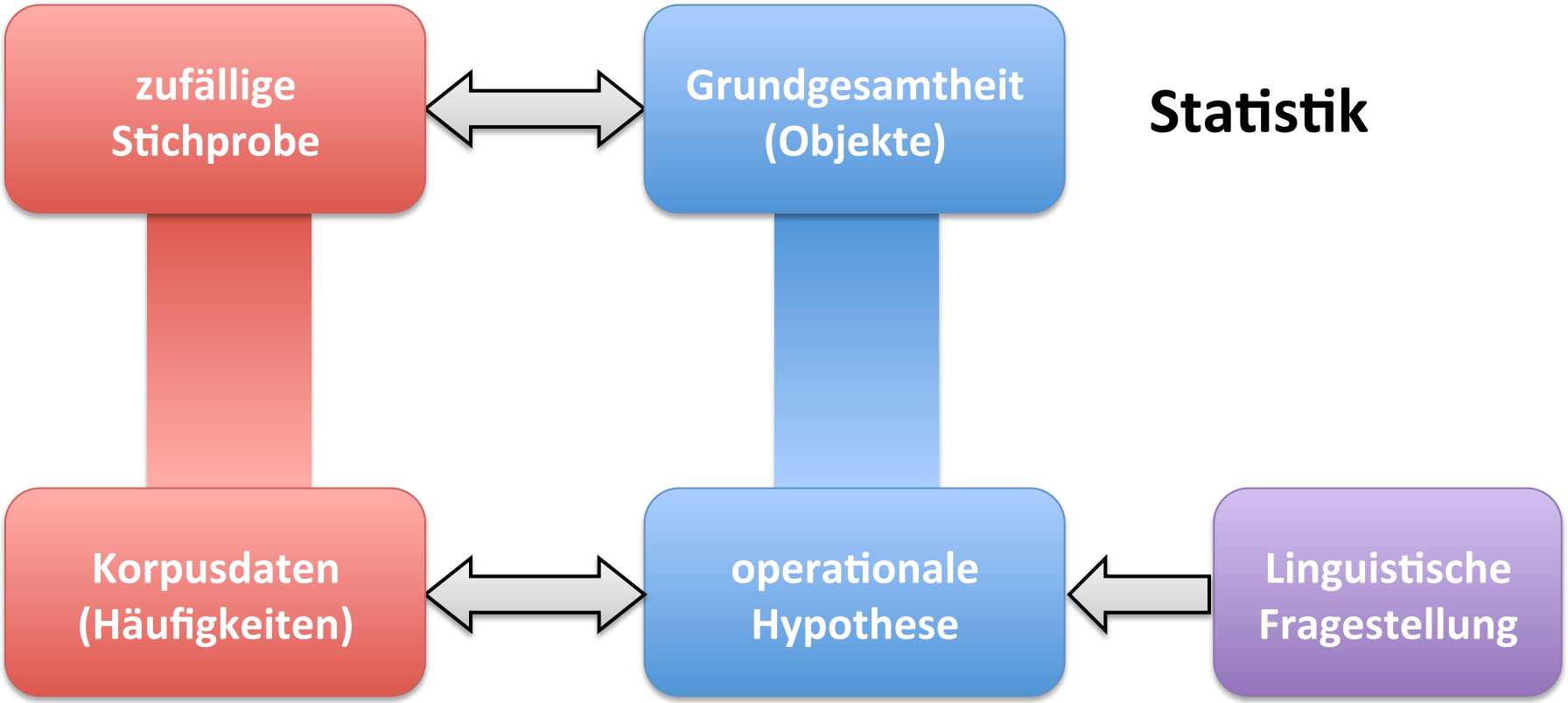
- Stichprobe von unterschiedlichen Wörtern oder einzelnen Vorkommen im Text?
  - Token ⇔ Gebrauchshäufigkeit
    vs. Typen ⇔ Vokabulargröße (Zusammenhang mit Produktivität)
  - üblicherweise Stichprobe von Token
  - Produktivitätsmessung erfordert komplexere statistische Methoden (→ Zipfsches Gesetz)

Grundgesamtheit?

Grundgesamtheit (Objekte)

- Was genau ist die „Grundgesamtheit“?
  - Wir interessieren uns für Eigenschaften von Sprechern (L1 und L2)
- Extensionaler Sprachbegriff:
  Sprache als Menge von Äußerungen
  - alle tatsächlichen und denkbaren Äußerungen bzw. Texte aus der relevanten Sprachvarietät
  - Objekte = Token (hier: Substantive im Text)
- Statistik trifft Aussagen über Grundgesamtheit!

Korpusstudie: Nominalkomposita



Vereinfachung

- Annahme: Wir kennen bereits die Häufigkeit von Nominalkomposita bei L1
  - aus bereits publizierten Untersuchungen
  - Ergebnis: 16% Nominalkomposita (hypothetisch!)
- Nur noch eine Stichprobe erforderlich
  - Substantive aus Texten von DaF-Lernern
  - wichtig: gleiche Textsorte, Domäne, usw.

Nullhypothese H_0

- Präzise quantitative Formulierung der operationalen Hypothese erforderlich
- Was ist die einfachste mögliche Hypothese?
  - L2 bilden < 16% Nominalkomposita?
  - L2 bilden ≠ 16% Nominalkomposita?
  - L2 bilden auch genau 16% Nominalkomposita?

Nullhypothese H_0

- Präzise quantitative Formulierung der operationalen Hypothese erforderlich
- Was ist die einfachste mögliche Hypothese?
  - L2 bilden < 16% Nominalkomposita
  - L2 bilden ≠ 16% Nominalkomposita
  - L2 bilden auch genau 16% Nominalkomposita
- Nullhypothese H_0 soll widerlegt werden!
  - statistische Verfahren können Hypothesen nur ablehnen, nicht bestätigen

Nullhypothese H_0

- Mathematische Formulierung von H_0:
  Die Häufigkeit μ von Nominalkomposita bei L2 beträgt genau 16%
- In Formeln:

$$H_0 : \ \mu = 16\% = .16$$

Stichprobe

- Zufallstichprobe von $n = 100$ Substantiven
  - Wie erstellt man eine zufällige Stichprobe aus Texten von DaF-Lernern?
  - Erinnerung: Stichprobe muss repräsentativ sein
- Ergebnis: $k = 12$ Komposita
  - Erwartung unter H_0: $k_0 = 16$ Komposita
  - weniger Komposita als erwartet → H_0 widerlegt?
- Zweite Stichprobe: $k = 17$ Komposita
  - Ablehnung von H_0 wäre voreilig gewesen!

Zufallsschwankungen

- Anzahl von Komposita in Stichprobe unterliegt Zufallsschwankungen
  - weicht i.d.R. vom „tatsächlichen“ Wert ab
  - zufällige Auswahl stellt sicher, dass **im Mittel** die erwartete Anzahl gefunden wird (falls H_0 gilt)

Zufallsschwankungen

- Bedeutung von $k = 12$ in Stichprobe
  - *a)* H_0 stimmt nicht, tatsächliche Häufigkeit geringer
  - *b)* H_0 stimmt, aber Stichprobe enthält zufällig weniger Komposita als erwartet
- Wir können (a) nur dann folgern, wenn (b) sehr unwahrscheinlich ist
  - intuitiv: Risiko, falsches Ergebnis zu publizieren
  - übliche Kriterien: Risiko < 5%, 1% oder 0.1%

Stichprobenverteilung

- Wie groß ist das Risiko, $k = 12$ oder weniger Komposita zu finden, sofern H_0 stimmt?
  → Stichprobenverteilung (unter H_0)
- Wir machen unser Leben zunächst etwas einfacher:
  - stellen wir uns vor, die Häufigkeit von Komposita wäre 50/100 (oder Singular vs. Plural, ...)
  - H_0: $\mu = 0.5$
  - Wie wahrscheinlich sind jetzt 12 Komposita / 100 Nomina?

12 statt 50 von 100?

- Alle Ergebnisse sind prinzipiell möglich (wir ziehen ja zufällige Nomina, wie ein Münzwurf)
- Aber sie haben unterschiedliche Wahrscheinlichkeiten
- Nur eine Folge führt zu 100 Mal NK:

$$P(100*NK) = P(NK) * P(NK) * \ldots * P(NK)$$
$$= P(NK)^{100} = 0.5^{100} = 7.888609e\text{-}31$$

12 statt 50 von 100?

- Auch für 0 NK bzw. 100x Simplex sorgt nur eine Folge:

  $P(100*S) = P(S) * P(S) * \ldots * P(S)$

  $= P(S)^{100} = 0.5^{100} = 7.888609e\text{-}31$

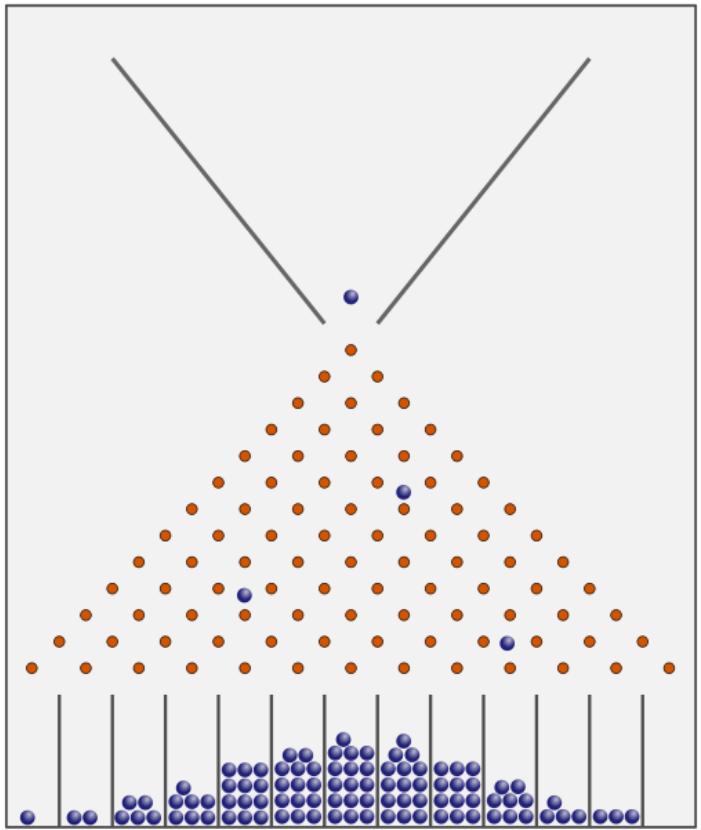
- Viele Kombinationen können zu 50 Mal NK führen (NK,S,NK,NK,S… oder S,NK,NK,...)

12 statt 50 von 100?

- Da alle Kombinationen gleich wahrscheinlich sind, müssen wir nur zählen, wie viele kombinationen zu 12 x NK führen
- Wenn wir unglaublich viele Stichproben ziehen würden...
- würden wir eine Verteilung bekommen
- Bestimmte Ergebnisse werden häufiger vorkommen als andere

Stichprobenverteilung

- Simulation von Münzwürfen: Galtonbrett
  - für Japan-Fans: Pachinko-Maschine



Stichprobenverteilung

- Jetzt nehmen wir an, P(NK) ist nicht 0.5, sondern 16/100. Geht das immer noch so?
  - P(NK)=0.5 ist ähnlich wie ein Münzwurf
  - jeweils 16% Wahrscheinlichkeit für Kompositum (unter H_0: $\mu = 0.16$) entspricht ungefähr Wurf von 6 Augen
  - Auswahl von 100 Token = 100-maliges Würfeln

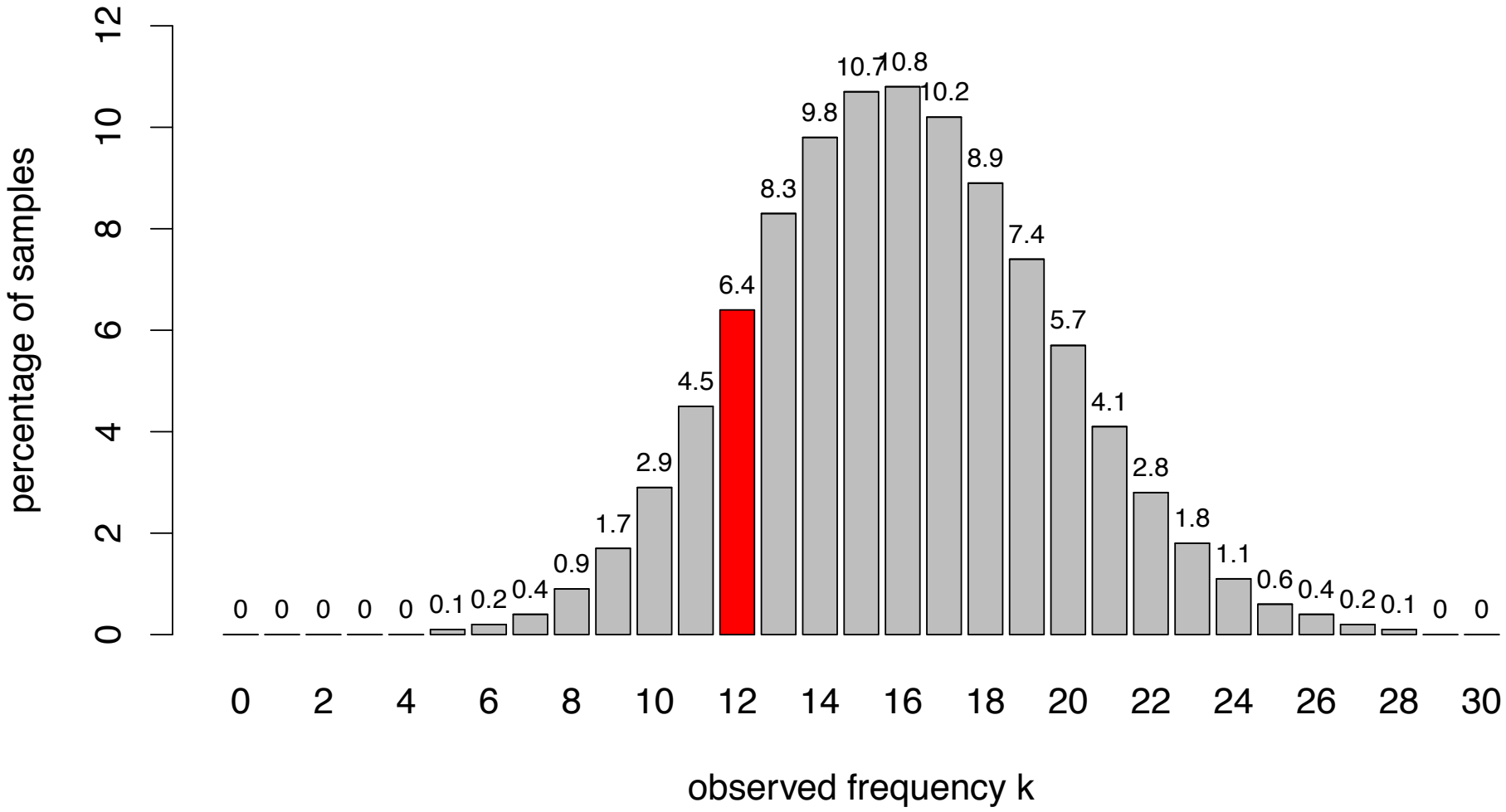
Stichprobenverteilung

- Auch hier gibt es wahrscheinlichere und weniger wahrscheinliche Ergebnisse
- Wahrscheinlichkeit, in 3x Würfeln 2x 6 zu bekommen:

  {6+~6+6, 6+6+~6, ~6+6+6}

- Das kann man auch für 100x Würfeln machen
- Wie oft kommen 12 NK in einer Stichprobe von 100 Nomina vor, wenn P(NK) = 0.16?

Binomialverteilung graphisch

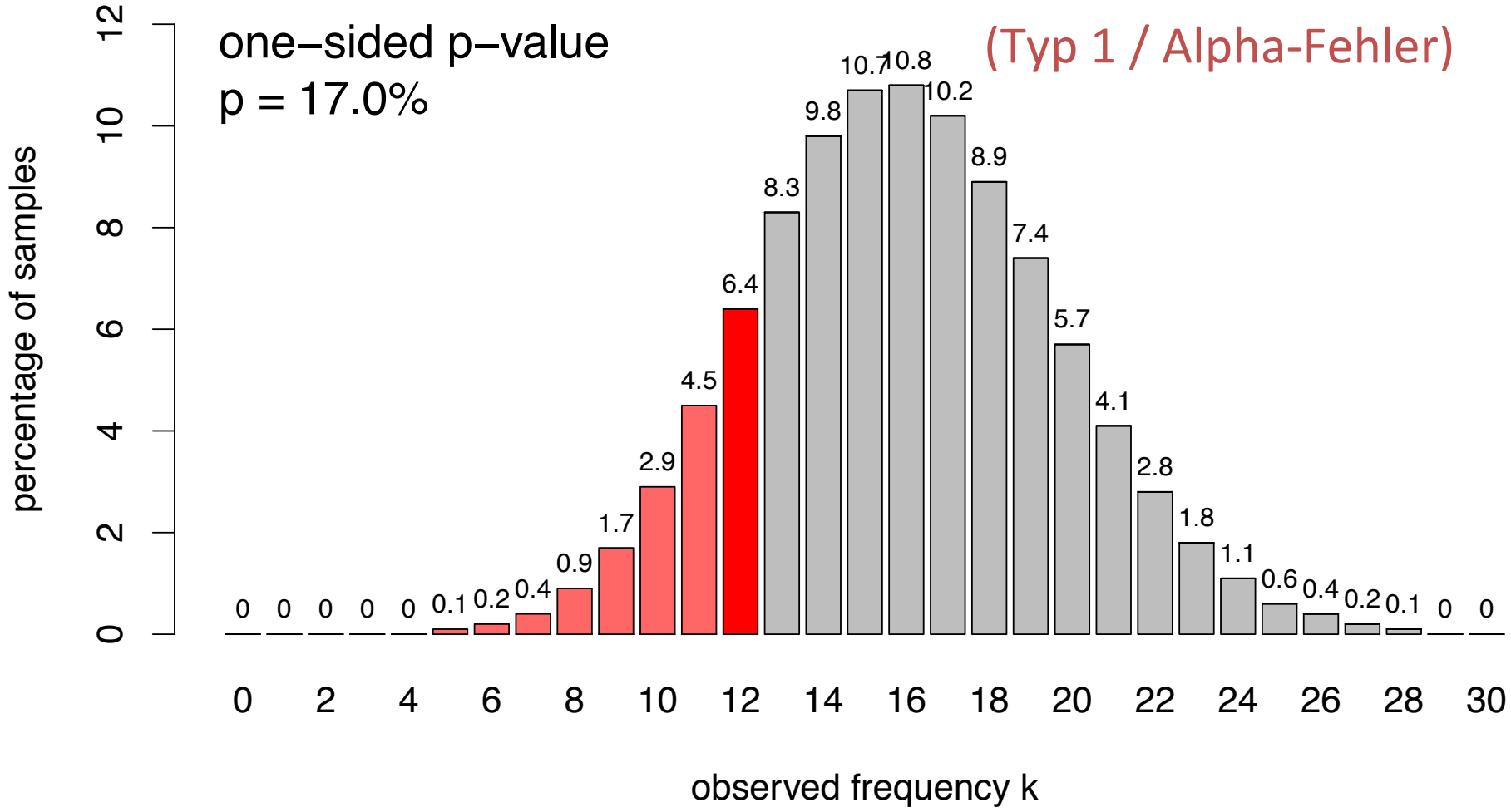


Gibt es wirklich einen Unterschied?

- Wir wissen nun, wie wahrscheinlich 12xNK sind (ca. 6,4% der Kombinationen unter H_0)
- Ist das jenseits von einem plausiblen Ergebnis, wenn H_0 stimmt?
- Wir können die Grenze zwischen plausibel und unplausibel beliebig verschieben
- Das ändert nur die Wahrscheinlichkeit, dass wir überreagieren
- Was ist die Wahrscheinlichkeit, dass diese Grenze falsch ist?

Binomialverteilung graphisch

Signifikanzwert = p-value
= Risiko falscher Ablehnung
(Typ 1 / Alpha-Fehler)



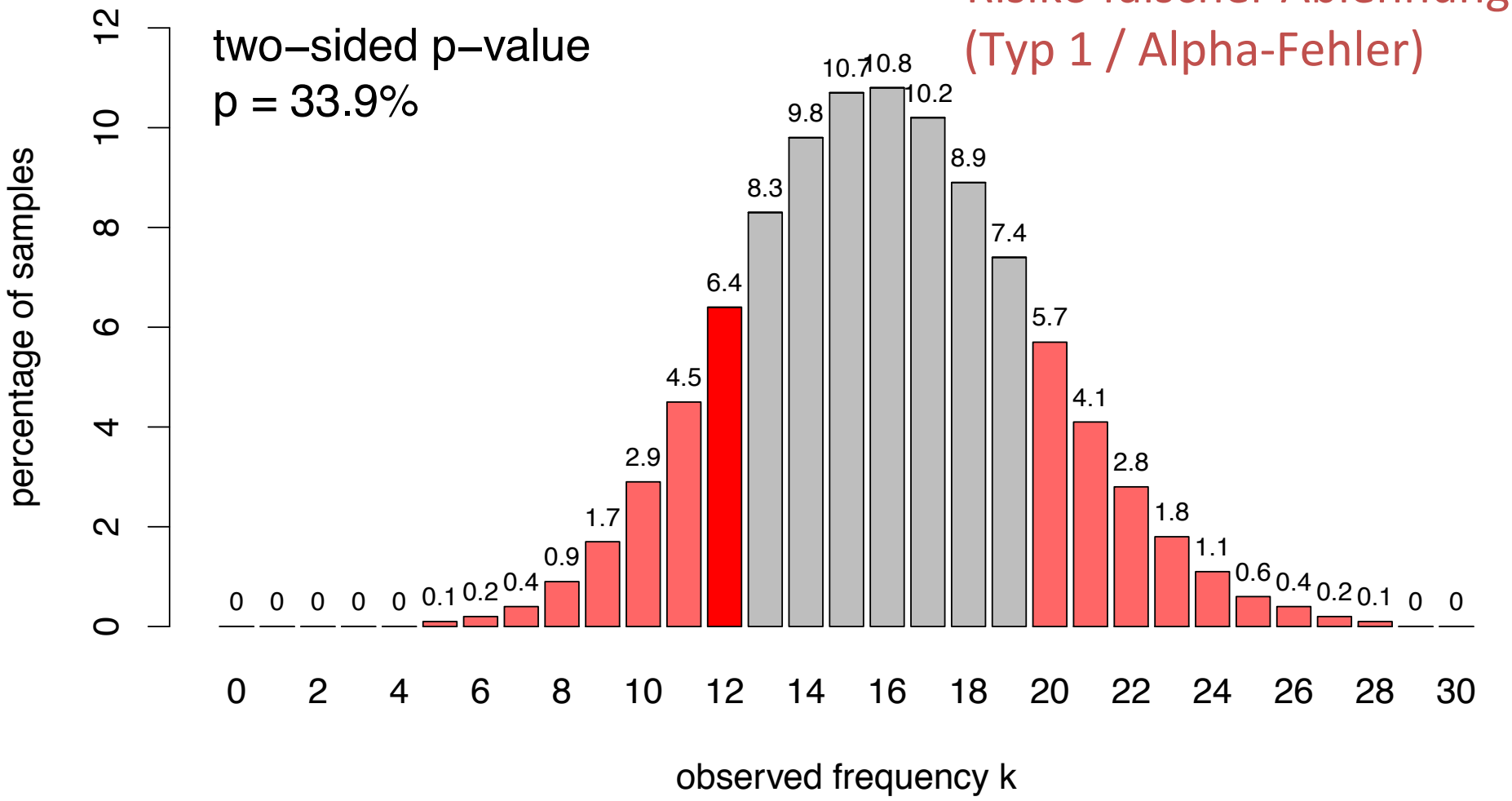
Einseitig oder zweiseitig?

- Wir haben gerade gezeigt, dass wir in 17% der Fälle einen Fehler machen würden, auch wenn H_0 stimmt
- Wir haben aber eine Alternative nicht berücksichtigt
- Was passiert, wenn Lerner in Wirklichkeit **mehr** Komposita produzieren?
- Keine Möglichkeit der Ablehnung aufgrund von zu vielen NK

Binomialverteilung graphisch

zweiseitiger Test (empfohlen)

Signifikanzwert = p-value
= Risiko falscher Ablehnung
(Typ 1 / Alpha-Fehler)



Signifikanzniveau

- Wann darf H_0 abgelehnt werden?
  - Signifikanzwert *p* beziffert das Fehlerrisiko
  - Ermessensfrage: welches Risiko ist akzeptabel?
- Übliche Signifikanzniveaus
  - $p < .05$ (5%) *
  - $p < .01$ (1%) **
  - $p < .001$ (0.1%) ***

Eine kleine Stärkung, bevor wir uns mit R beschäftigen …

MITTAGSPAUSE

ERSTE SCHRITTE MIT R

Erste Schritte in R

- Starten Sie jetzt R:
  - Windows/Mac: einfaches R-GUI
    - unter Windows Rgui.exe
    - unter Mac Os X R bzw. R64
  - Linux, SunOS, …: Befehl R auf Kommandozeile
  - empfohlenes GUI: RStudio (Windows, Mac, Linux)

- Das Programm gibt einen Prompt aus und wartet auf Ihre Eingabe:

  >

Erste Schritte in R

- Sie geben einen Befehl ein bzw. stellen eine Anfrage:

```
> 2+2
```

- Das Programm antwortet Ihnen und wartet auf weitere Befehle:

```
> 2+2
[1] 4
>
```

- Ignorieren Sie fürs erste die Markierung `[1]` vor der Antwort … ihre Bedeutung wird bald klar werden.

R als Taschenrechner

- Die Bedeutung der Symbole `+ - * /`
  ist leicht zu erkennen: plus, minus, mal, geteilt
- Der Ausdruck: `2^3`
  bedeutet $2^3 = 2 \cdot 2 \cdot 2$.
  - Vorsicht: Auf manchen Tastaturen müssen Sie die Taste ^ zweimal drücken, damit Sie etwas sehen (dead key)!
- Alle wichtigen mathematischen Funktionen sind schon vorhanden:
  `sqrt(4)` $= \sqrt{4} = 2$
  `log2(256)` $= \log_2 256 = 8$ (weil $2^8 = 256$)
  ...

Kleine Übung

- Berechnen Sie:

$\frac{1}{5} - 2$	`1/5 - 2`
$\frac{1}{5-2}$	`1/(5-2)`
$(2+3)^2$	`(2+3)^2`
$2 \cdot 3^2$	`2 * 3^2`
2^{5-2}	`2^(5-2)`
$\sin(3.1415)$	`sin(3.1415)`

Ergebnisse in Variablen merken

- Berechnungen wiederholen sich oft → Sie wollen mit Zwischenergebnissen weiterrechnen
- Dafür gibt es Variablen:
  - `a <- 3` erzeugt eine Variable („Behälter“) namens `a`
  - Dieser Behälter enthält nun den Zahlenwert `3`
  - Wenn Sie einfach nur `a` auf der Kommandozeile eingeben, bekommen Sie den Inhalt der Variablen angezeigt
  - Die Zeichen `<-` gehören zusammen (←) und dienen dazu, einer Variablen einen Wert zuzuweisen
  - Sie können Variablen auch zu neuen verknüpfen, z.B. `c <- a + b`
- Was gibt R aus, wenn Sie der Variablen `Mio` die Zahl `1000000` zuweisen und dann ihren Wert anzeigen lassen? Was bedeutet das?

Mehr zu Variablen

- Es ist meist sinnvoll, den Variablen sprechendere Namen zu geben als `a`, `b` oder `c`:
  - `Reaktionszeit.Mittelwert` oder
  - `alter_standardabweichung`
  - Regeln: Trennzeichen `.` und `_` ; Umlaute vermeiden

    (Hier geht's fürs erste weiter mit `a`, `b`, `c`, das ist kürzer)
- Sie können den Wert einer Variablen verändern. Ganz wichtig ist das in folgendem Beispiel:

```
> a <- 5
> a <- a + 1
> a
[1] 6
```

Eine praktische Kleinigkeit

- Mit der Zeit werden Ihre Befehle länger werden und es ist mühsam, sie immer wieder neu einzutippen
- Mit der Pfeil-Hoch-Taste (↑) bekommen Sie den vorherigen Befehl angezeigt
  - Sie können Ihn wiederverwenden und bearbeiten
  - Sie können auch weiter zurückblättern durch alle Befehle, die Sie bisher eingegeben haben
- Mit der Tabulator-Taste (⇆) können Sie Variablen- und Dateinamen automatisch vervollständigen

Vergleichsoperationen

- Sie können mit R auch Zahlenwerte vergleichen:

```
> 5 < 10
[1] TRUE
> 2 + 2 == 5
[1] FALSE
```

- Das Ergebnis lässt sich als Wahrheitswert in einer Variablen speichern:

```
> a <- (-1 < 0)
> a
[1] TRUE
```

Die Klammern sind hier eigentlich nicht erforderlich

Vergleichsoperatoren

Operator	Bedeutung
`>`	größer als
`>=`	größer oder gleich
`<`	kleiner als
`<=`	kleiner oder gleich
`==`	ist gleich
`!=`	ist nicht gleich

- Ist √1000 größer oder kleiner als 33? Weisen Sie der Variable `groesser` im ersten Fall den Wert `TRUE` zu, sonst den Wert `FALSE`.

Zeichenketten

- R kann auch mit Zeichenketten umgehen
  (→ wichtig für uns Korpuslinguisten ;-)
- Zeichenketten werden in einfachen oder doppelten Anführungszeichen geschrieben
- Können beliebigen Variablen zugewiesen werden

```
> a <- "Hallo"
> b <- 'Welt'
> paste(a, b)  # miteinander verketten
[1] "Hallo Welt"
```

Ich bin ein Kommentar

Vektoren

- Sie können auch mehrere Zahlen in *einer* Variable speichern, indem Sie sie mit `c()` aneinanderhängen

```
> noten <- c(2.0, 1.7, 3.3, 1.0, 2.7)
> noten
[1] 2.0 1.7 3.3 1.0 2.7
```

- Mit solchen „Vektoren“ lassen sich viele statistische Berechnungen sehr einfach durchführen

```
> sum(noten)        # Summe
> length(noten)     # Anzahl
> mean(noten)       # Mittelwert
> sd(noten)         # Standardabweichung
```

Vektoren

- Ein zweites Beispiel: Zahlenreihen

```
> n <- 1:100   # „:“ erzeugt eine Zahlenreihe
> quadrate <- n^2
> quadrate
[1]     1     4     9    16    25    36    49
[8]    64    81   100   121   144   169   196
...
```

  - jetzt sollte klar sein, was `[1]` und `[8]` bedeuten

- Direkter Zugriff auf einzelne Elemente eines Vektors:

```
> quadrate[5]
> quadrate[1:10]
> quadrate[7] <- 42   # geschummelt!
```

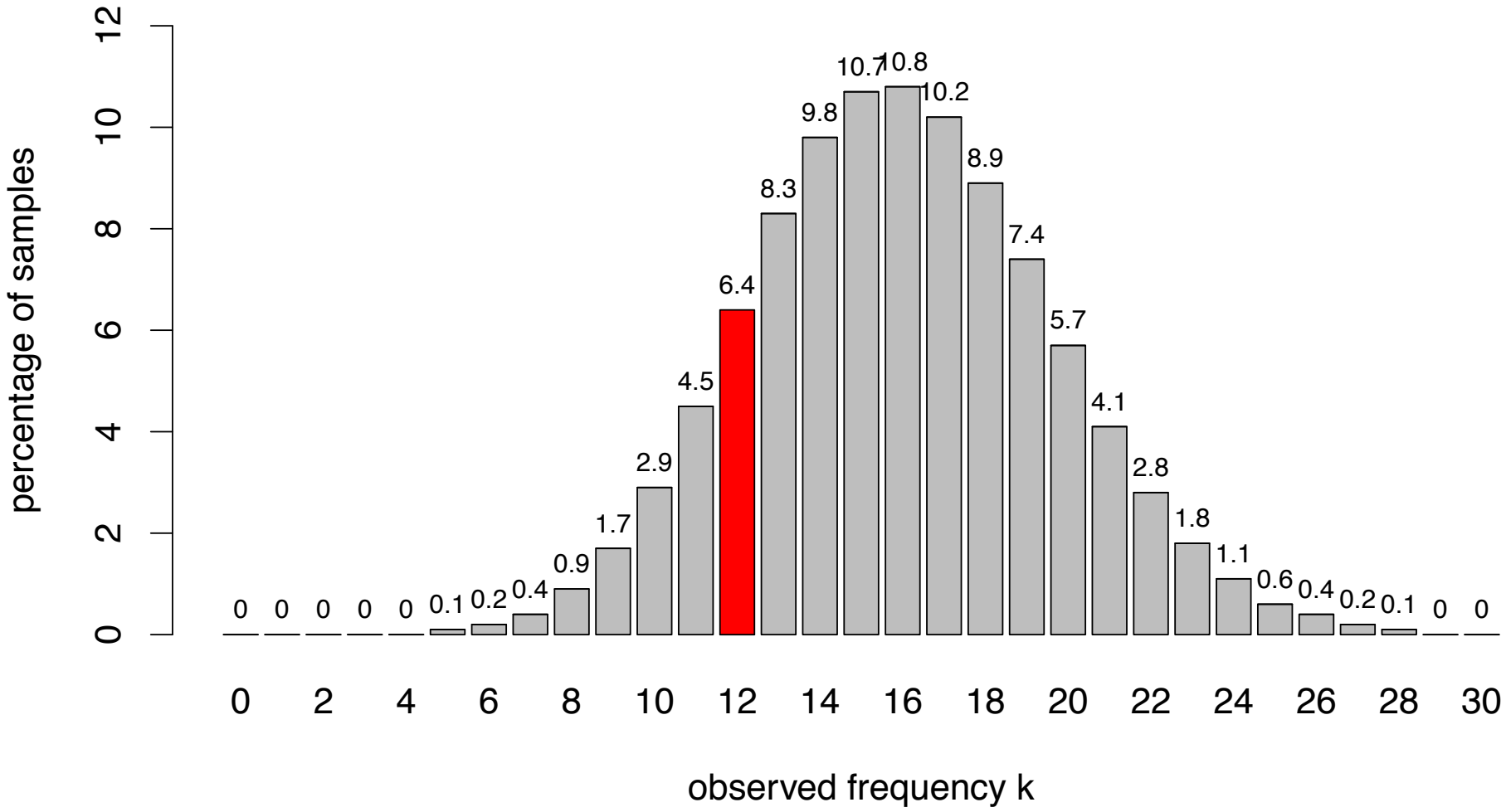
Wahrscheinlichkeiten in R

- Für den Binomialtest wollten wir wissen:
  - Was ist die Wahrscheinlichkeit von 12/100 NK falls die Nullhypothese $\mu = 16/100$ stimmt?
- R bietet vordefinierte Funktion für die Binomial- und viele andere statistische Verteilungen

```
> dbinom(12, 100, 16/100)
[1] 0.06417714
```

R nimmt die Sache ziemlich genau …

Wahrscheinlichkeiten in R



Binomialtest in R

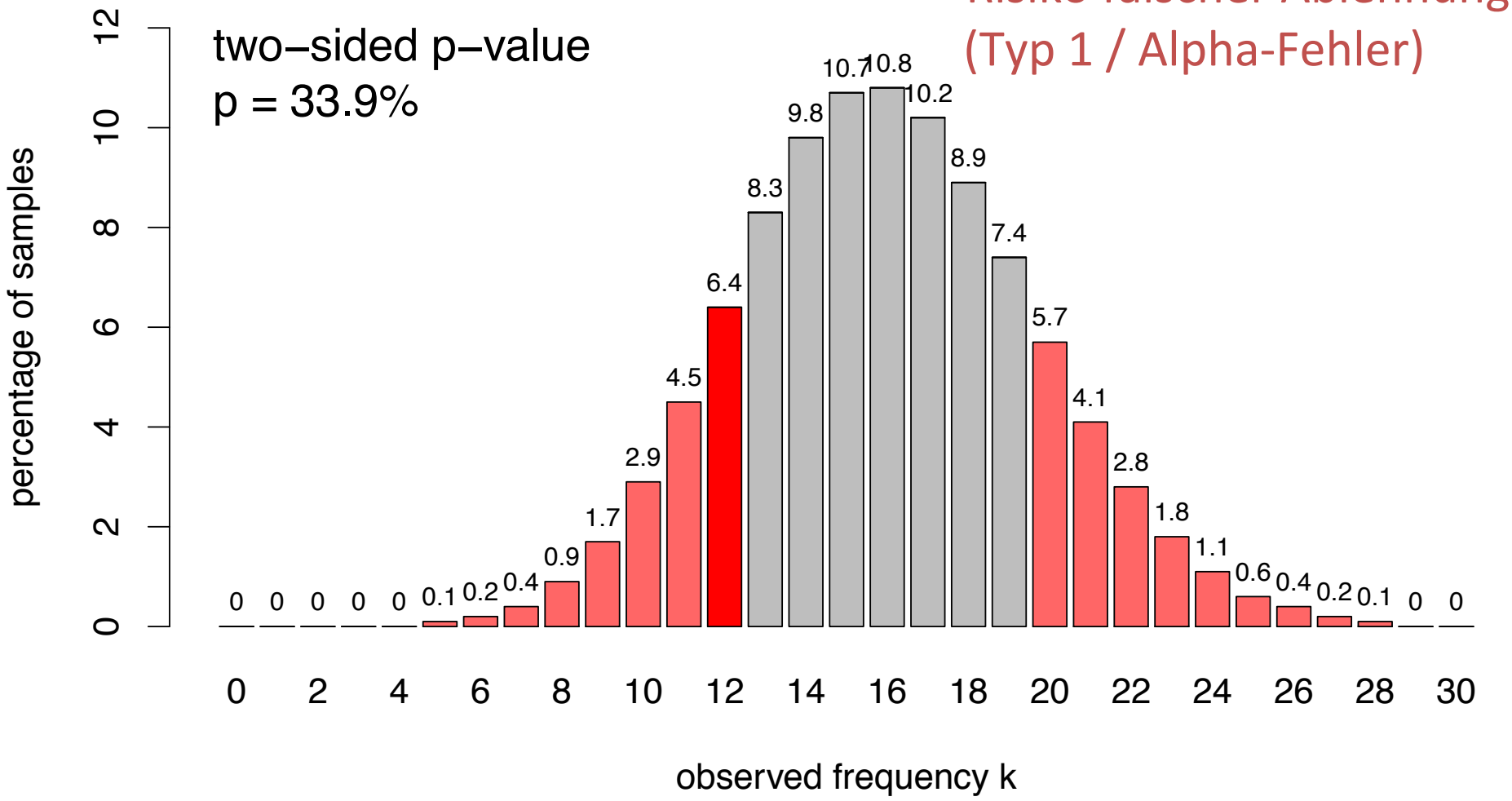
- Die eigentliche Frage war aber anders:
  - Wie groß ist das Risiko, das Ergebnis 12 NK oder ein noch ungewöhnlicheres Ergebnis zu erhalten, falls die Nullhypothese μ = 16/100 tatsächlich stimmt?
  - Passiert das in weniger als 5% aller Fälle?

→ Binomialtest

Binomialtest in R

zweiseitiger Test

Signifikanzwert = p-value
= Risiko falscher Ablehnung
(Typ 1 / Alpha-Fehler)



Binomialtest in R

Nullhypothese

```
> binom.test(12, n = 100, p = 16/100)
```

Stichprobengröße

```
    Exact binomial test

data:  12 and 100
number of successes = 12, number of trials = 100,
    p-value = 0.3392
alternative hypothesis: true probability of success is
    not equal to 0.16
95 percent confidence interval:
 0.0635689 0.2002357
sample estimates:
probability of success
                  0.12
```

Binomialtest in R

```
> binom.test(x, n = 100, p = 16/100)
```

Eine kleine Übung:

- Ab welchem x ist Signifikanz p < 0.05* erreicht?
- Ab welchem x ist Signifikanz p < 0.01** erreicht?

→ ausprobieren

Ein kleiner Unterschied

- Was ist, wenn Lerner tatsächlich weniger Komposita verwenden: im Schnitt 12/100 statt 16/100?
  - dann finden wir in einer Stichprobe von 100 Substantiven üblicherweise ca. 12 NK
- Dieser Unterschied ist nach dem Binomialtest nicht signifikant – kann man ihn also nie erkennen?
  - nur wenn zufälligerweise noch weniger als 12 NK auftreten

Ein kleiner Unterschied

- Bisher standen sog. **Typ I-** oder **Alpha-Fehler** im Mittelpunkt: irrtümliche Ablehnung von H_0
  - Signifikanzwert berechnet Riskio von Alpha-Fehlern
- **Typ II-** oder **Beta-Fehler**:
  - H_0 ist falsch, kann aber nicht abgelehnt werden
  - z.B. tatsächliche Häufigkeit von 12% NK bei L2
- Risiko von Beta-Fehlern ist schwer abzuschätzen
  - hängt von tatsächlichem Durchschnittswert ab
  - je größer der Unterschied, desto kleiner ist das Risiko, einen Beta-Fehler zu begehen

Typ II/Beta-Fehler

- Können Sie das Risiko für Beta-Fehler abschätzen?
  - Nullhypothese: 16% NK
  - tatsächlicher Wert bei L2-Sprechern: 12% NK
- H_0 wird abgelehnt bei ≤ 8 / 100 NK in Stichprobe
  - Wie groß ist die Wahrscheinlichkeit, eine solche Stichprobe zu erhalten?

```
> dbinom(0:8, 100, 12/100)
> sum(dbinom(0:8, 100, 12/100))
[1] 0.1385921
```

- Risiko für Beta-Fehler: 100% – 13.86% = 86.14%

Trennschärfe

- Können wir das Risiko von Beta-Fehlern verringern?
- Was wäre bei einer 5x größeren Stichprobe?
  - tatsächlich im Schnitt 12/100 NK bei L2
  - in Stichprobe von *n* = 500 Substantiven sind also üblicherweise 60 NK zu erwarten

```
> binom.test(60, n = 500, p = 16/100)
…… p = .01452* ……
```

- größere Stichprobe → bessere **Trennschärfe** (*power*)
  → kleinere Differenz von H_0 (**Effektgröße**) genügt, um Beta-Fehler zu vermeiden
  - Passieren jetzt überhaupt keine Beta-Fehler mehr?

Häufigkeitsvergleich

- Wie kann H_0 formuliert werden, wenn die Kompositahäufigkeit bei L1 nicht bekannt ist?

$$H_0 : \ \mu_1 = \mu_2$$

- Häufigkeitsvergleich von 2 Stichproben
  - keine Annahme über genauen Wert von $\mu_1 = \mu_2$
  - vergleichbare Stichproben aus L1- und L2-Texten, aber Stichprobengröße darf unterschiedlich sein
- R-Befehl: `prop.test()`

Häufigkeitsvergleich

- Stichprobe L2: $k = 52$ NK, $n = 500$
- Stichprobe L1: $k = 76$ NK, $n = 500$

```
> prop.test(c(52, 76), c(500, 500))
```

Two samples: frequency comparison

	Frequency count	Sample size
Sample 1	52	500
Sample 2	76	500

Clear fields

Calculate

95% confidence interval

in automatic format

with 4 significant digits

Häufigkeitsvergleich

- Stichprobe L2: $k = 52$ NK, $n = 500$
- Stichprobe L1: $k = 76$ NK, $n = 500$

```
> prop.test(c(52, 76), c(500, 500))

    2-sample test for equality of proportions with …

data:  c(52, 76) out of c(500, 500)
X-squared = 4.7395, df = 1, p-value = 0.02948
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.091306444 -0.004693556
sample estimates:
prop 1 prop 2
 0.104  0.152
```

R beenden

- Sie dürfen R jetzt kurz verlassen:

```
> q()
Save workspace image? [y/n/c]: n
```

- Falls Sie ein GUI verwenden, wählen Sie den entsprechenden Befehl im Menü aus
  - im Dialogfenster ebenfalls „nicht speichern“ anklicken

ÜBUNG 1

Deutschkenntnisse & Bilingualität

- Fragestellung: Lernen bilinguale Schüler schlechter Deutsch als monolinguale?
  - Operationalisierung: Anzahl Fehler bei Diktat (500 Wörter)
  - zwei Gruppen deutscher Muttersprachler:
    (a) monolingual
    (b) bilingual (Deutsch-Englisch oder Deutsch-Russisch)
  - jeweils Schüler der 4. Klasse an einer Berliner Schule
- Ergebnisse zusammengestellt in tabellarischer Form in der Datei `diktate.txt`

Deutschkenntnisse & Bilingualität

- Datei `diktate.txt`
- 20 Zeilen = Schüler
- 3 Spalten
  - Fallnummer
  - Sprache: MONO / BI
  - Anzahl Fehler in Diktat von 500 Wörtern

Fall	Sprache	Fehler
1	MONO	22
2	MONO	20
3	MONO	10
4	MONO	16
…	…	…
11	BI	21
12	BI	19
13	BI	28
14	BI	28
…	…	…

Deutschkenntnisse & Bilingualität

- Was ist unsere Nullhypothese?
- Wie groß sind die Stichproben?
- Welche Häufigkeiten werden verglichen?

Fall	Sprache	Fehler
1	MONO	22
2	MONO	20
3	MONO	10
4	MONO	16
…	…	…
11	BI	21
12	BI	19
13	BI	28
14	BI	28
…	…	…

Schritt 1: Einlesen der Tabelle

- R kann tabellarische Daten in Textform mit dem Befehl `read.table()` einlesen
  - kann mit geeigneten Parametern an zahlreiche unterschiedliche Formate angepasst werden
  - Voreinstellung für TAB-getrennte Felder: `read.delim()`
  - Voreinstellung für CSV-Format: `read.csv()`, `read.csv2()`
- Wir benötigen die Option `header=TRUE`, da die Tabelle eine Kopfzeile enthält

Schritt 1: Einlesen der Tabelle

- Drei Möglichkeiten zur Auswahl der Datei
  1. Arbeitsverzeichnis wechseln (GUI-Menü) oder R im entsprechenden Verzeichnis starten (Kommandozeile)
  2. Vollen Pfad zur Datei angeben (Verzeichnis + Dateiname), bei GUI oft per Drag & Drop möglich
  3. Interaktive Auswahl mit `file.choose()`
- Einlesen der Tabelle in Variable Diktate

```
> Diktate <- read.table("diktate.txt", header=TRUE)
```

oder

```
> Diktate <- read.table(file.choose(), header=TRUE)
```

```
> Diktate
```

Schritt 2: Zugriff auf Tabellen

- Statistische Auswertungen werden sehr oft auf tabellarischen Daten durchgeführt
- R hat dafür einen eigenen Datentyp `data.frame` (analog zu Zahlen, Zeichenketten, Vektoren, …)
- Zugriff auf Zeilen, Spalten und einzelne Elemente

```
> Diktate[12, "Sprache"]
> Diktate[12, 2]
> Diktate[12, ]      # ganze Zeile
> Diktate$Fehler     # ganze Spalte
```

Schritt 3: Tabellen bearbeiten

- Wir benötigen die gesamte Anzahl der Fehler für jede Gruppe von Schülern (MONO und BI)
- Dazu spalten wir die Tabelle in zwei Teile auf:

```
> MONO <- subset(Diktate, Sprache == "MONO")
> BI <- subset(Diktate, Sprache == "BI")
```

- Jetzt können wir die Werte in der dritten Spalte jeder Teiltabelle aufsummieren:

```
> sum(MONO$Fehler)
> sum(BI$Fehler)
```

  - R-Profis machen es so:

```
rowsum(Diktate$Fehler, Diktate$Sprache)
```

Schritt 4: Häufigkeitsvergleich

- Stichprobe BI: $k = 226$ Fehler, $n = 5000$
- Stichprobe MONO: $k = 185$ Fehler, $n = 5000$

```
> prop.test(c(226, 185), c(5000, 5000))

    2-sample test for equality of proportions with …

data:  c(226, 185) out of c(5000, 5000)
X-squared = 4.0598, df = 1, p-value = 0.04392
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0002197596 0.0161802404
sample estimates:
prop 1 prop 2
0.0452 0.0370
```

Diskussion

- Ergebnis: bilinguale Schüler machen signifikant mehr Fehler als monolinguale Schüler ($p = 0.044 < 0.05^*$)
  - Was bedeutet dieses Ergebnis?
- Wie groß ist der Unterschied eigentlich?
  - Stichprobe BI: 226/5000 Wörter = 4.5% falsch
  - Stichprobe MONO: 185/5000 Wörter = 3.7% falsch
- Machen bilinguale Schüler tatsächlich über 20% mehr Fehler als monolinguale?
  - D.h. beträgt der tatsächliche Unterschied zwischen BI und MONO ebenfalls 0.8 Prozentpunkte (wie in Stichproben)?

KAFFEEPAUSE

KONFIDENZINTERVALLE

Konfidenzintervalle

- Was ist, wenn gar keine Hypothese vorliegt?
  - Fragestellung: „Wie häufig bilden L2-Sprecher Komposita?“
- Häufigkeitsschätzung auf Basis einer Stichprobe von $n = 1000$ Substantiven
  - Ergebnis: $k = 120$ NK unter $n = 1000$ Substantiven
  - direkter Schätzwert = Punktschätzer:

$$\hat{\mu} = \frac{k}{n} = \frac{120}{1000} = 0.12 = 12\%$$

  - entspricht Schätzwerten in der Diskussion von Übung 1
- Wie zuverlässig ist dieser Schätzwert?

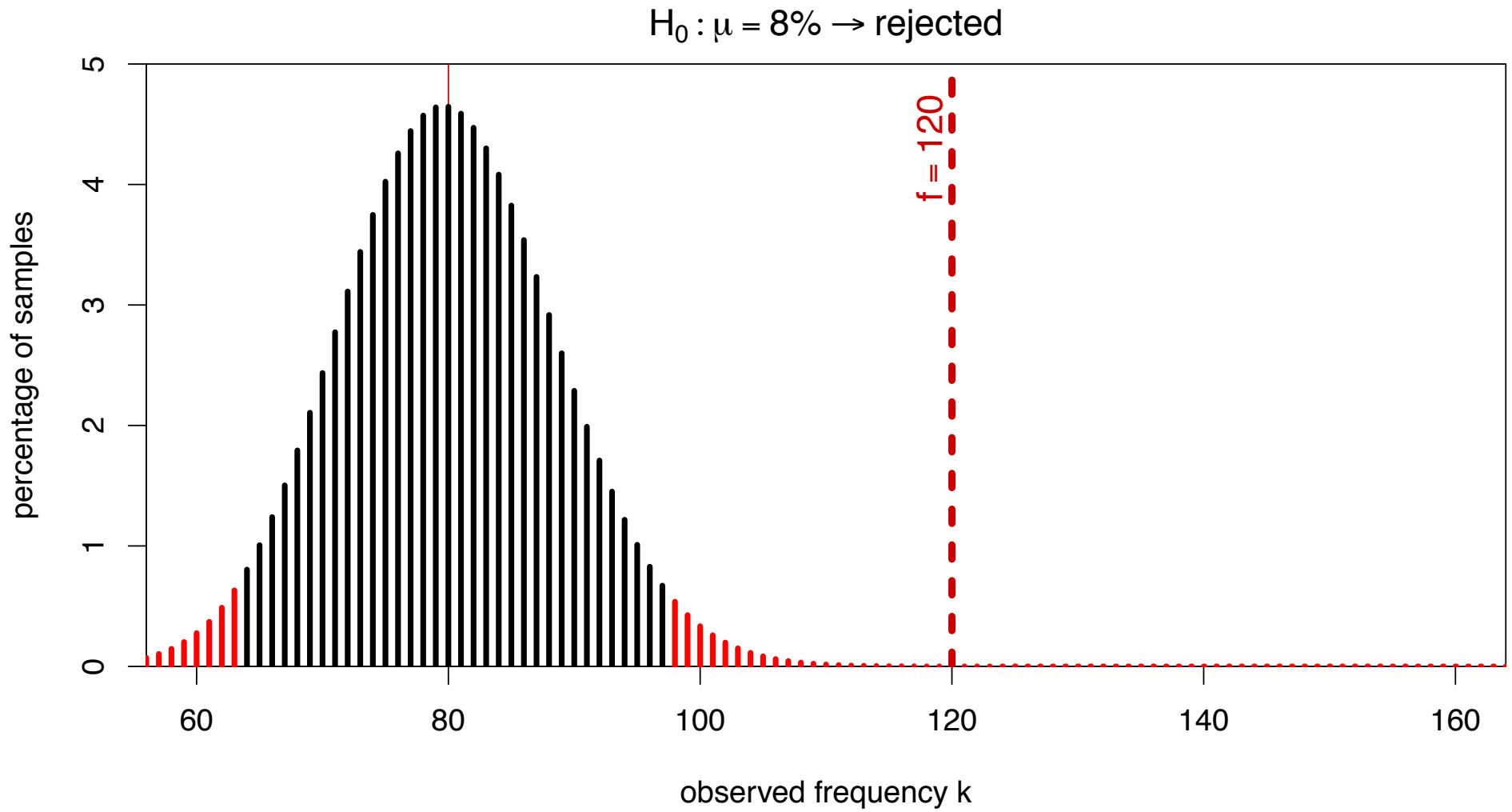
Konfidenzintervalle

- Mehrere Stichproben (jeweils $n = 1000$)
  1) $k = 120 \rightarrow \mu = 12.0\%$
  2) $k = 105 \rightarrow \mu = 10.5\%$
  3) $k = 129 \rightarrow \mu = 12.9\%$
  4) $k = 126 \rightarrow \mu = 12.6\%$
  5) $k = 111 \rightarrow \mu = 11.1\%$
  6) $k = 92 \rightarrow \mu = 9.2\%$
  7) $k = 117 \rightarrow \mu = 11.7\%$

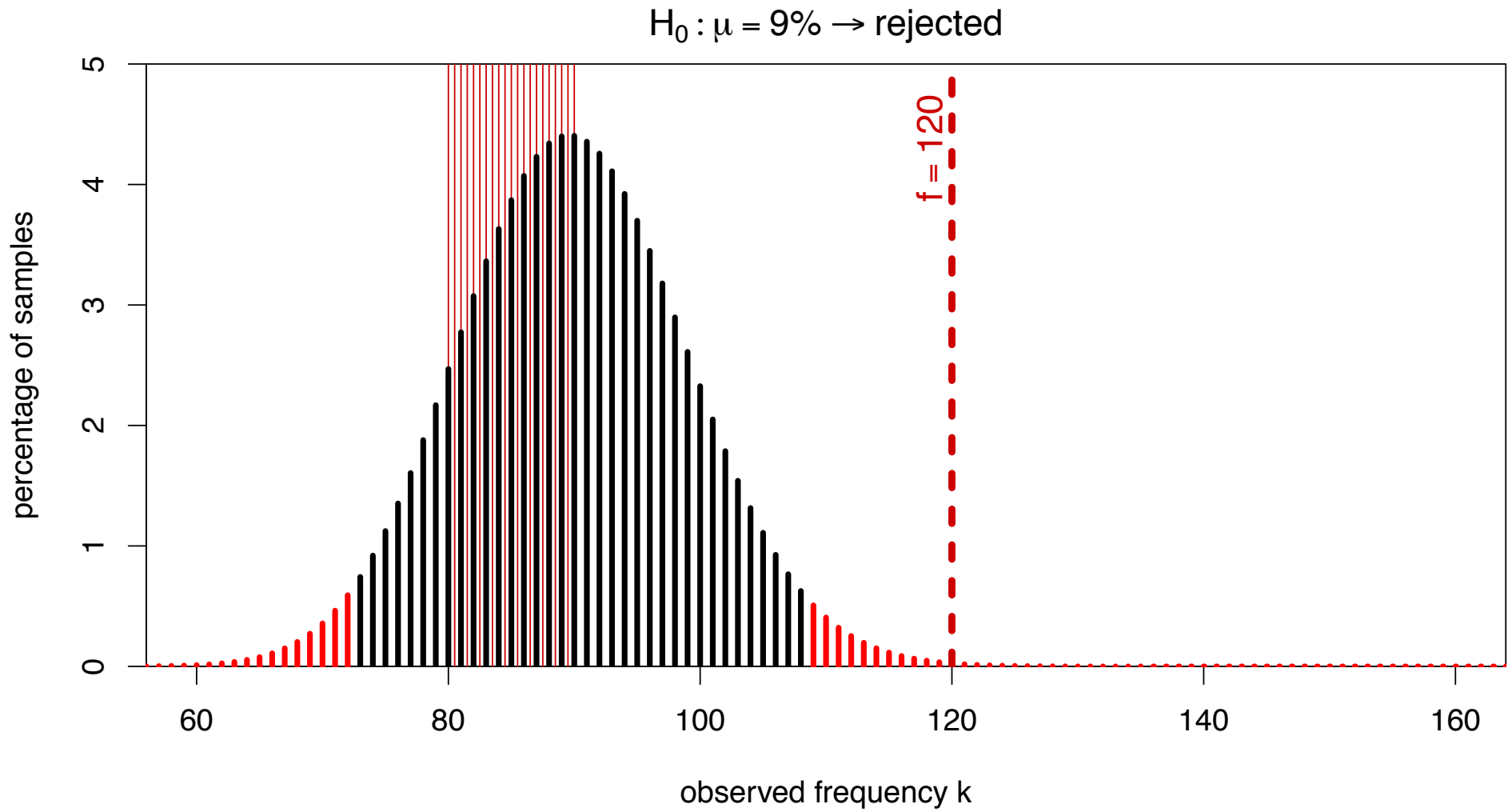
Konfidenzintervalle

- Welcher dieser Schätzwerte ist plausibel?
  - wir wollen auf Basis *einer* Stichprobe entscheiden
  - können tatsächlichen Wert von μ nicht genau bestimmen, sondern nur auf einen bestimmten Bereich eingrenzen
  - Bereich plausibler Schätzwerte = **Konfidenzintervall**
- Idee: Ausschlussverfahren
  - Stichprobe: 120 NK unter 1000 Substantiven
  - Ist der Schätzwert $\mu = 10\%$ plausibel? → Nullhypothese H_0
  - Binomialtest: $p < 0.05$* → $\mu = 10\%$ ist nicht plausibel
  - für alle möglichen Schätzwerte μ ausprobieren

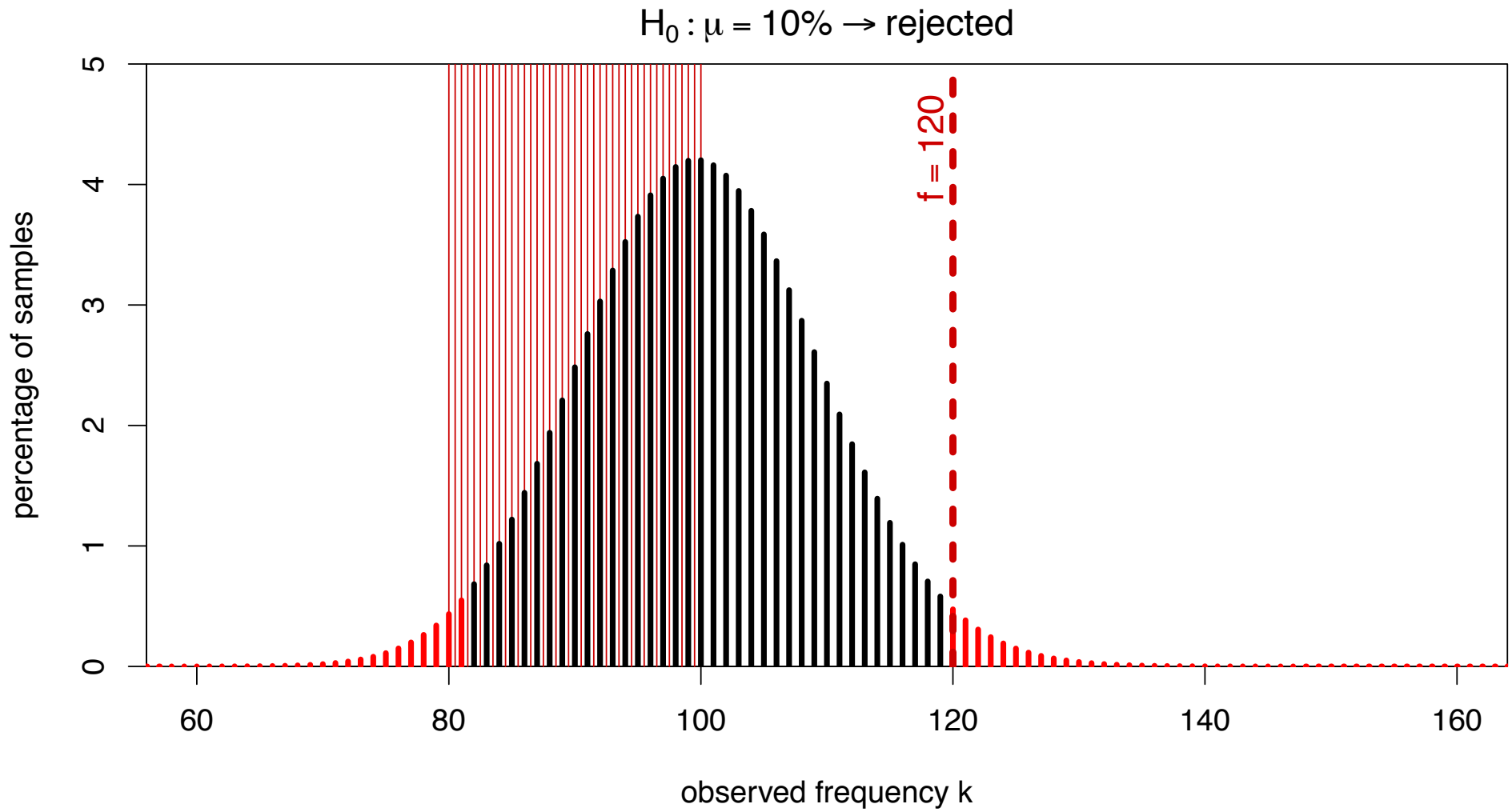
Konfidenz: invertierter Test



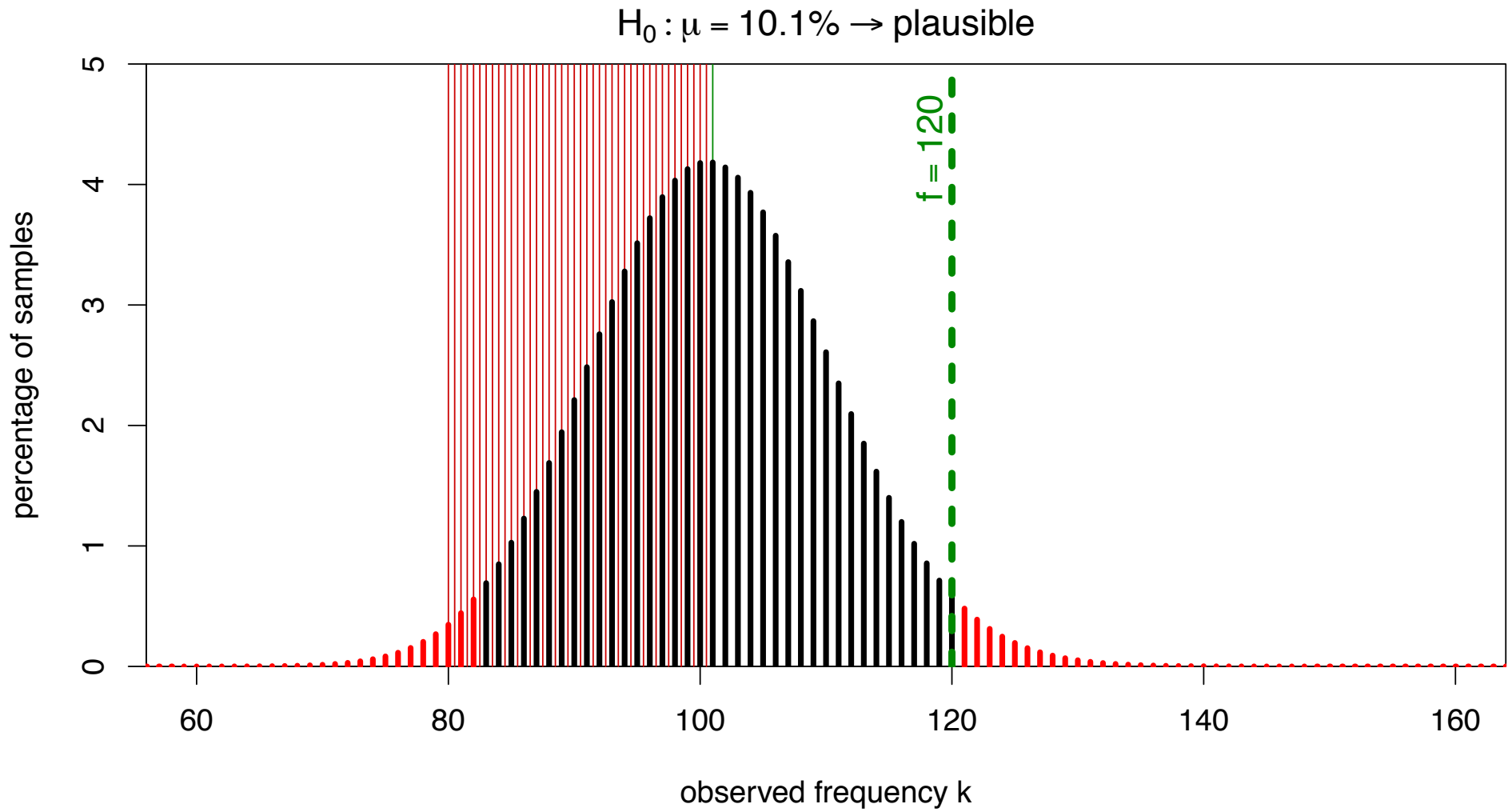
Konfidenz: invertierter Test



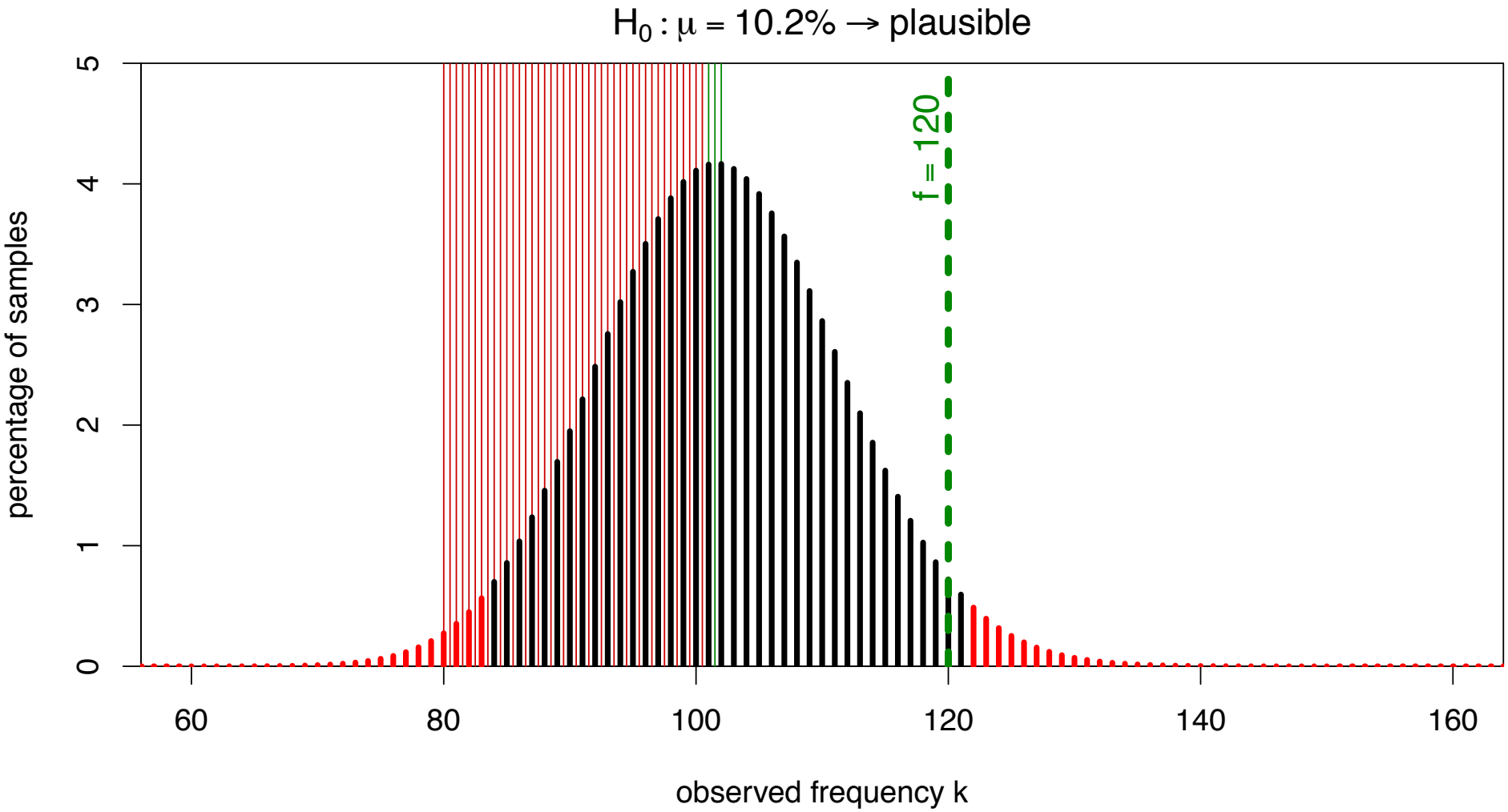
Konfidenz: invertierter Test



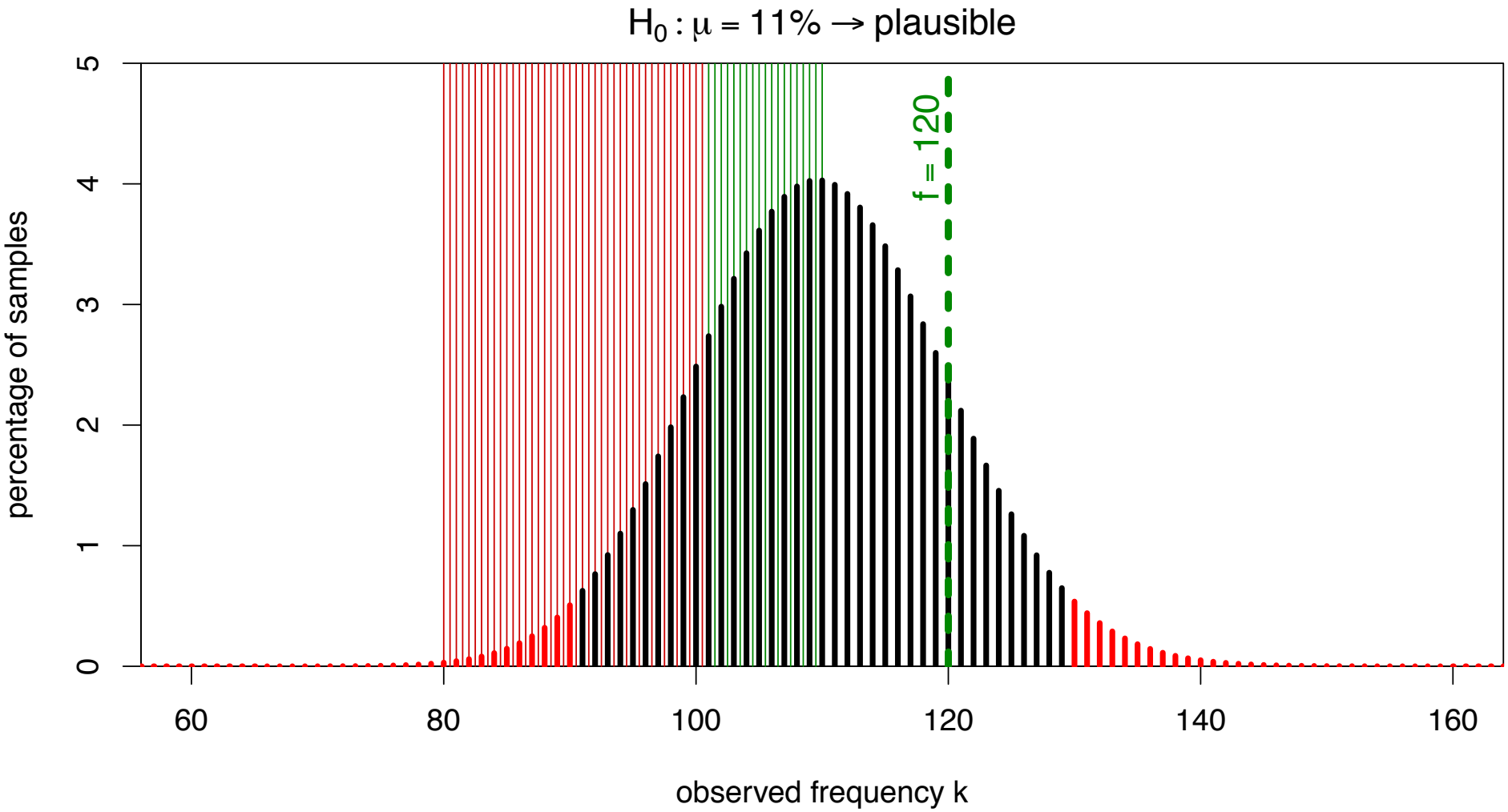
Konfidenz: invertierter Test



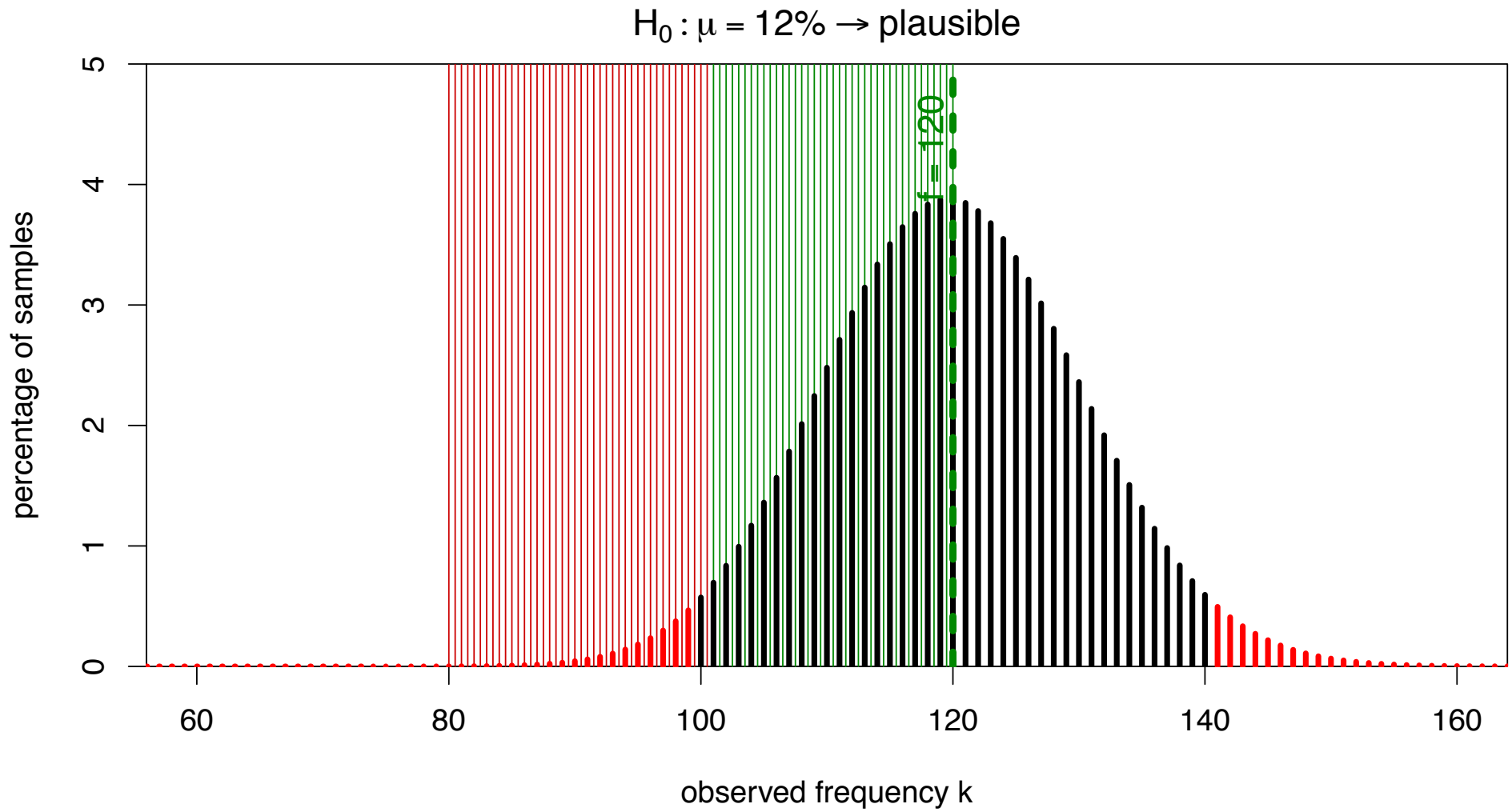
Konfidenz: invertierter Test



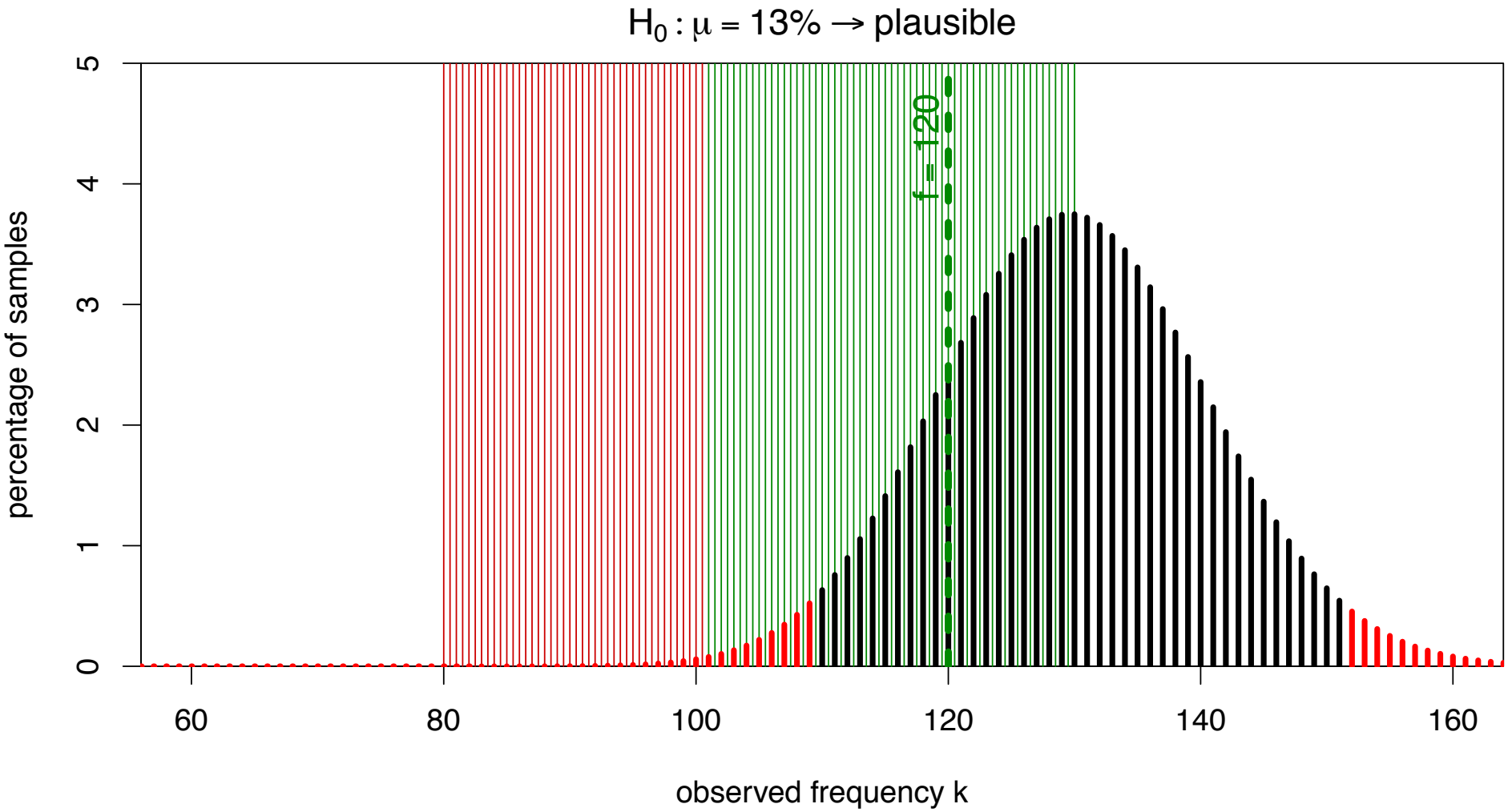
Konfidenz: invertierter Test



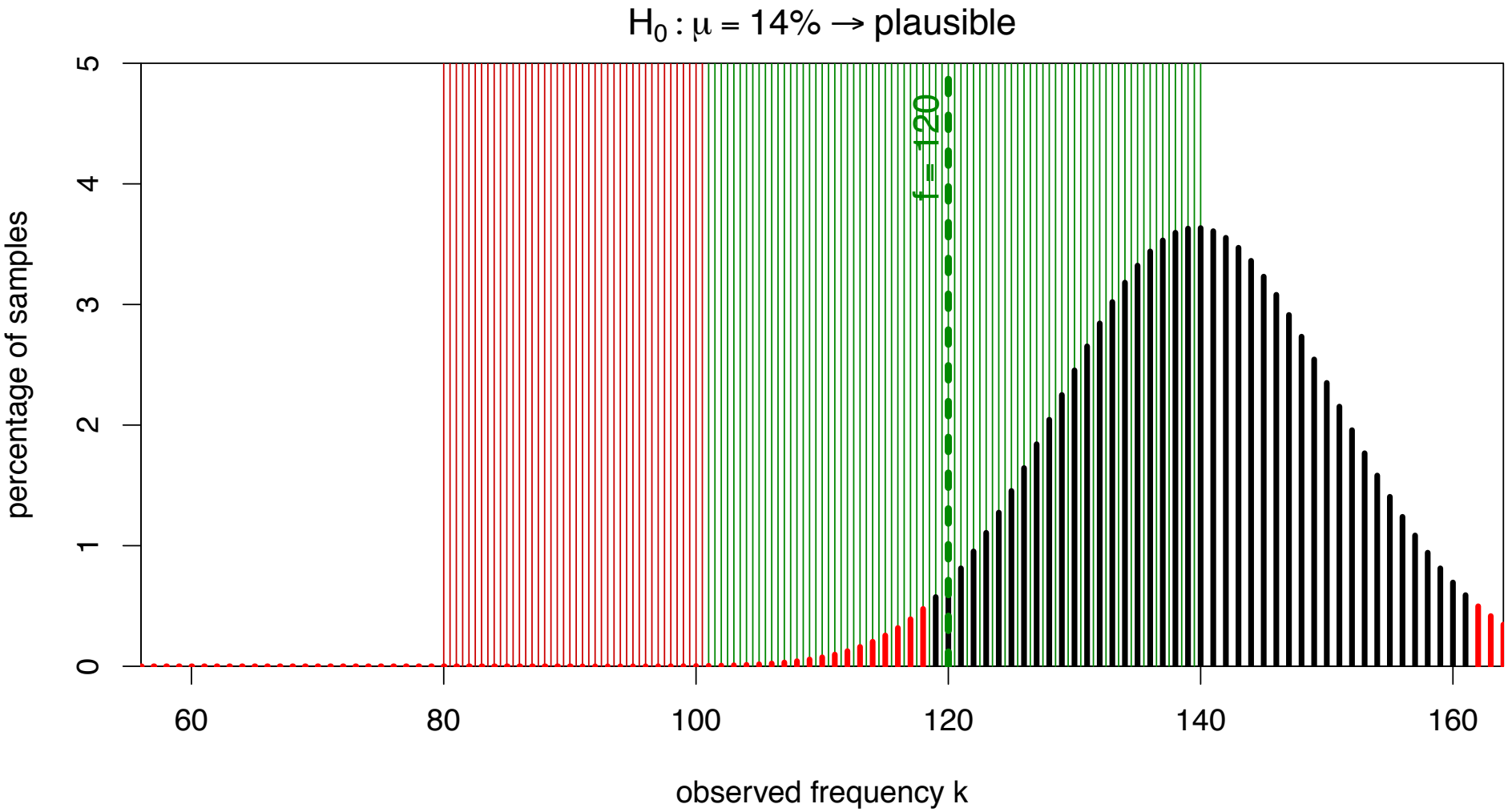
Konfidenz: invertierter Test



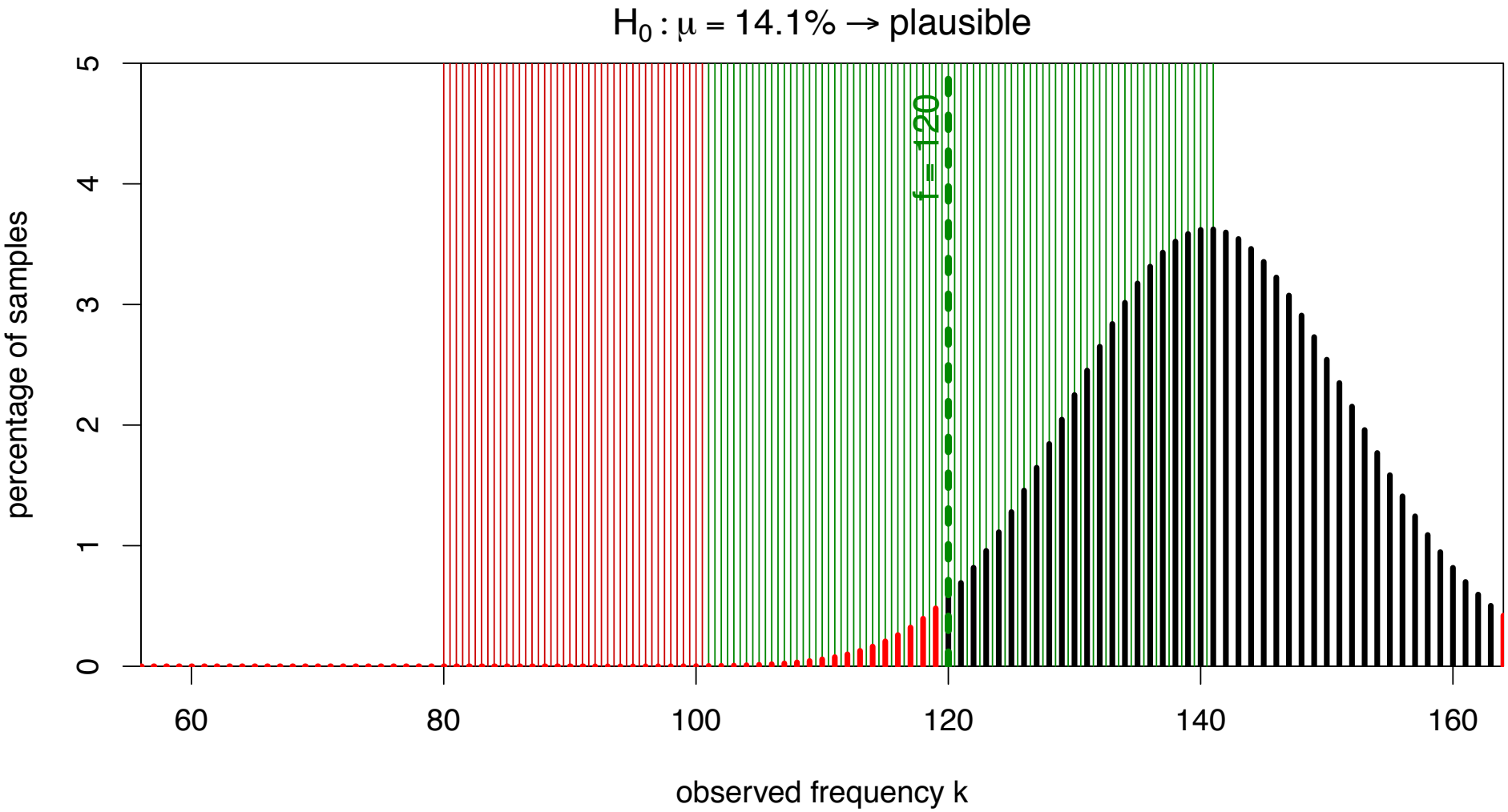
Konfidenz: invertierter Test



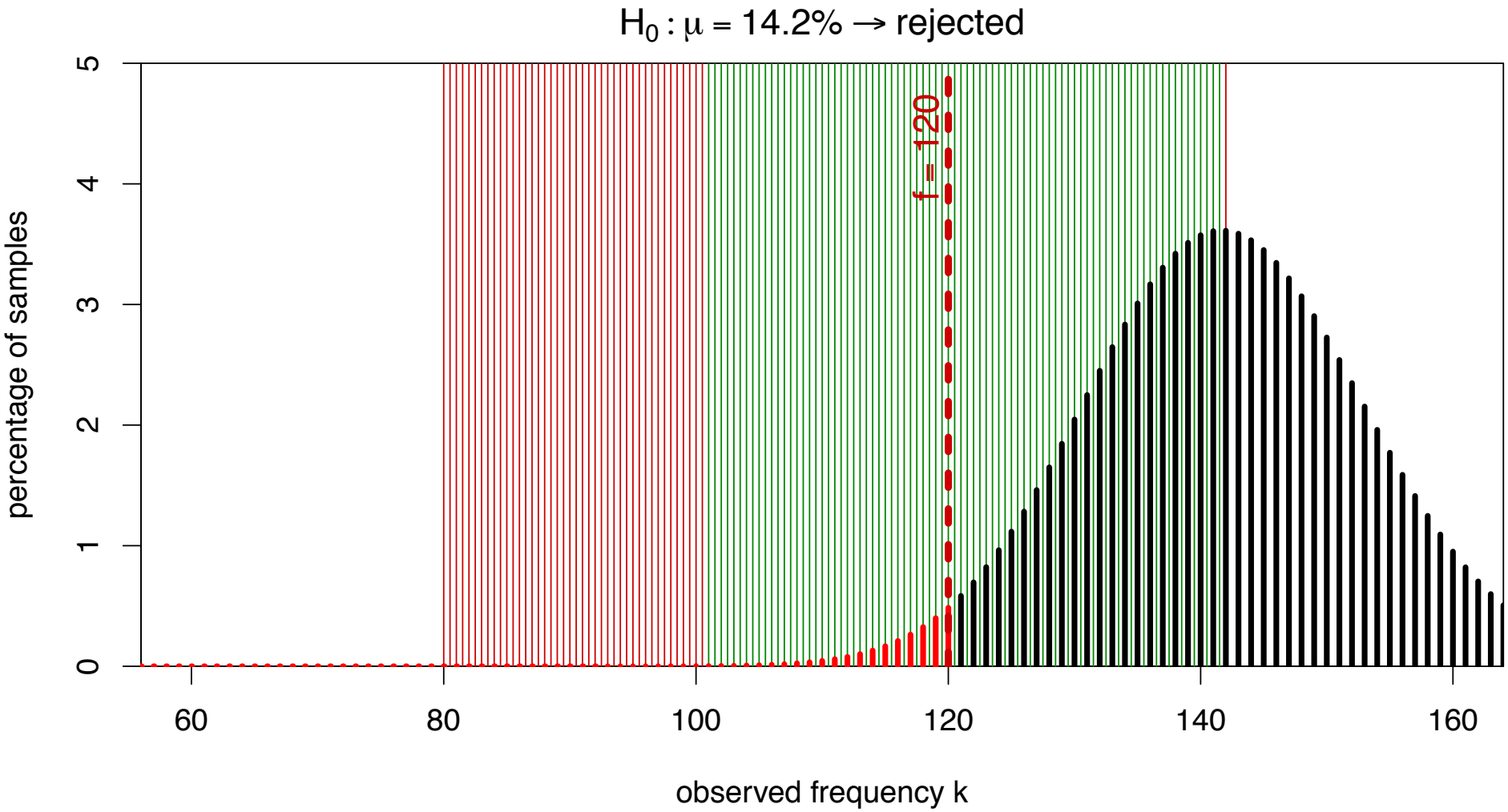
Konfidenz: invertierter Test



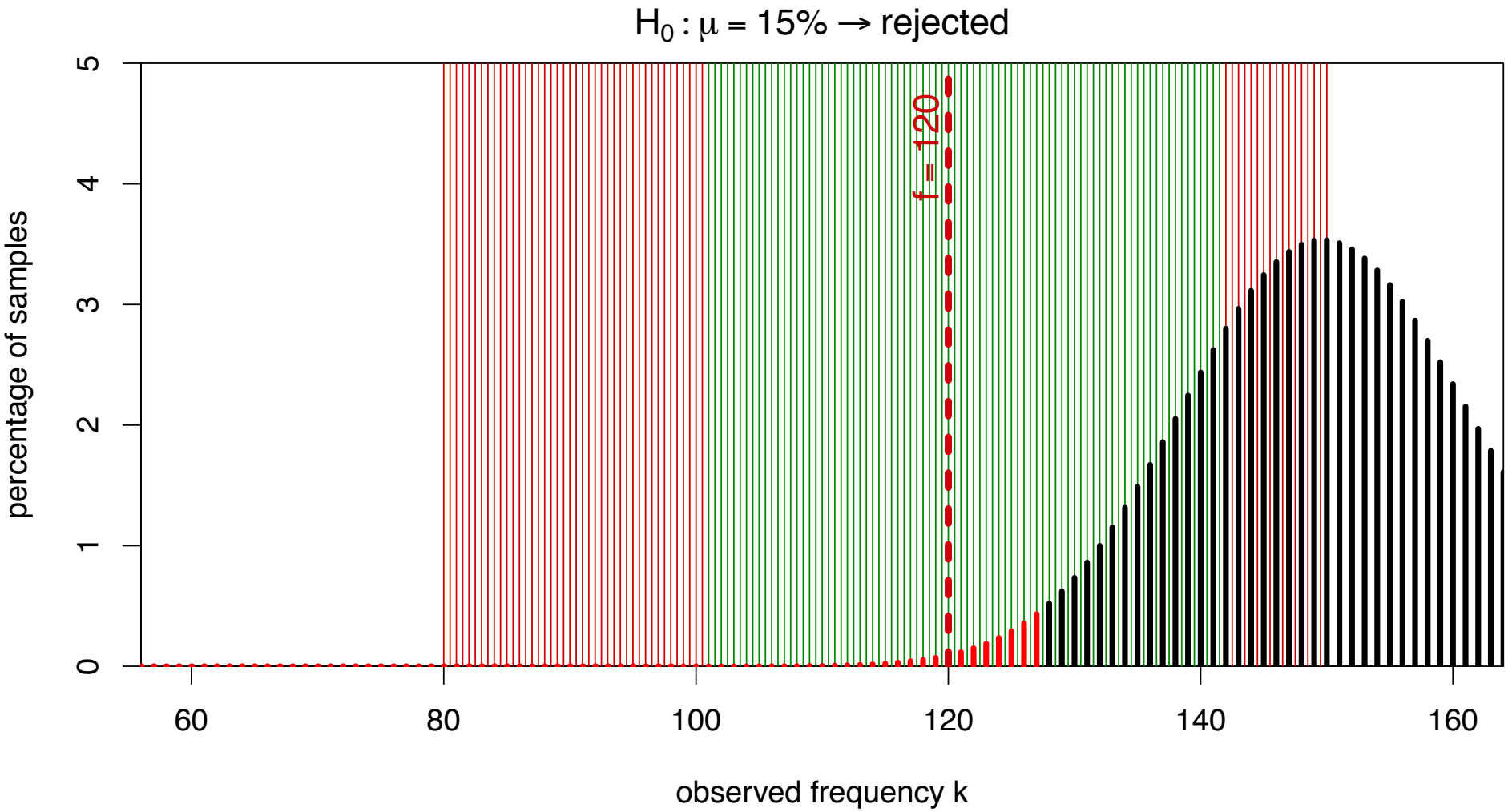
Konfidenz: invertierter Test



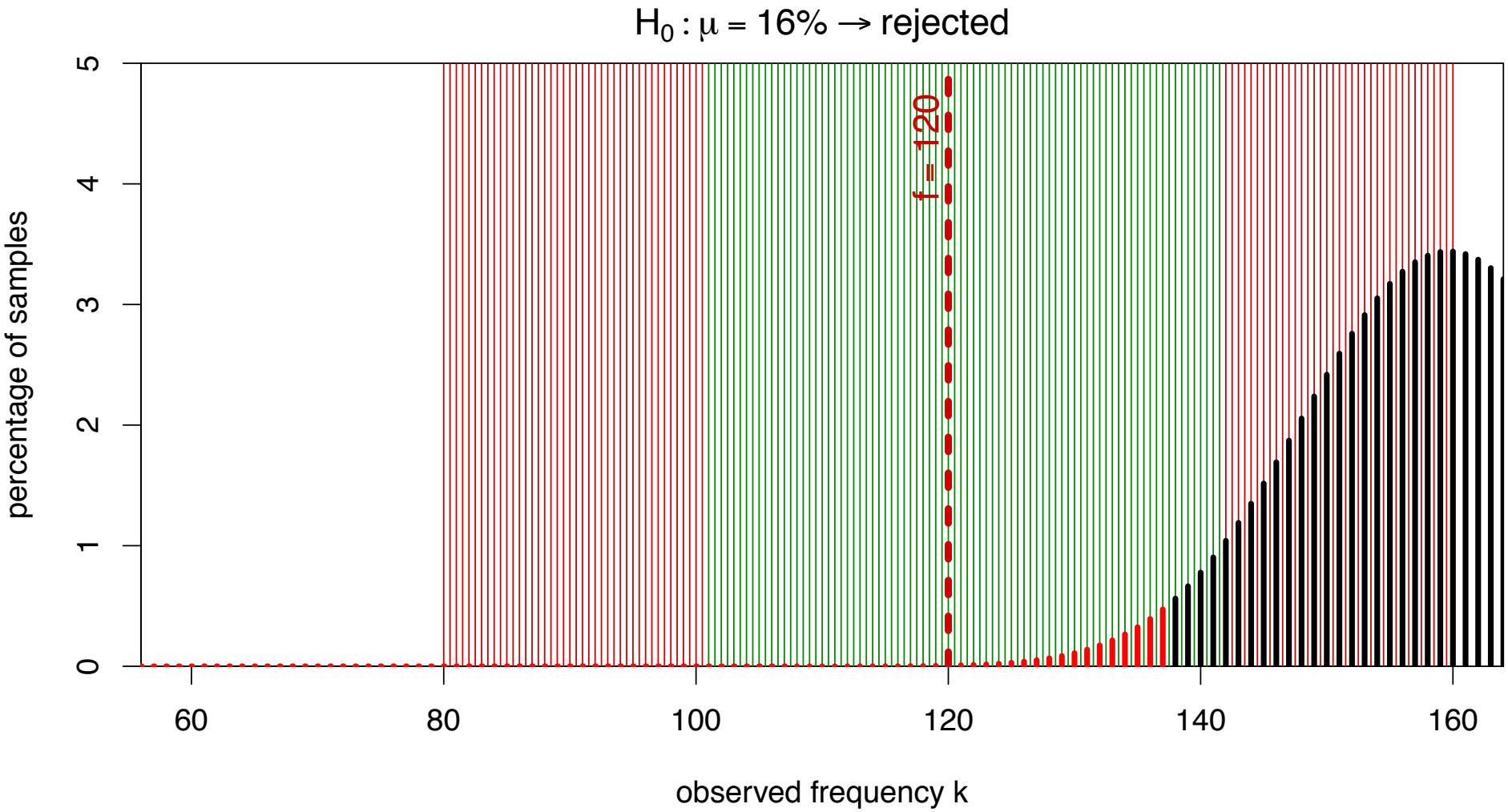
Konfidenz: invertierter Test



Konfidenz: invertierter Test



Konfidenz: invertierter Test



Konfidenzintervall in R

```
> binom.test(120, n = 1000)

    Exact binomial test

data:  120 and 1000
number of successes = 120, number of trials = 1000,
    p-value < 2.2e-16
alternative hypothesis: true probability of success
    is not equal to 0.5
95 percent confidence interval:
 0.1005009 0.1417669
...
```

→ Tatsächliche Häufigkeit im Bereich 10.1% … 14.2%
mit 95% Konfidenz (bei Signifikanzniveau $p < 0.05$)

Konfidenzintervall für Übung 1

```
> prop.test(c(226, 185), c(5000, 5000))

   2-sample test for equality of proportions with …

data:  c(226, 185) out of c(5000, 5000)
X-squared = 4.0598, df = 1, p-value = 0.04392
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0002197596 0.0161802404

…
```

→ Tatsächlicher Unterschied der Fehlerhäufigkeiten liegt zwischen 0.02 und 1.62 Fehlern / 100 Wörter!

Konfidenzintervall für Übung 1

- Unterschied in Übung 1 war **signifikant** ($p < 0.05$*)
  - bilinguale Schüler machen mehr Fehler als monolinguale
- Wir können aber nur mit (95%iger) Sicherheit sagen, dass die bilingualen Schüler mindestens 0.02 Fehler mehr je 100 Wörter machen → nicht **relevant**
  - Signifikanz (Fehlerrisiko) vs. Relevanz (Effektgröße)
- Tatsächlicher Unterschied könnte aber auch bei 1.62 Fehlern / 100 Wörter liegen → wäre relevant
- Wie können wir die Effektgröße genauer bestimmen?
  - größere Stichprobe → bessere Trennschärfe des Tests → Konfidenzintervall wird kleiner

Signfikanz vs. Relevanz

- **Studie 1:** $k_1 = 11$, $n_1 = 90$ | $k_2 = 35$, $n_2 = 110$
- **Studie 2:** $k_1 = 12500$, $n_1 = 51000$ | $k_2 = 11200$, $n_2 = 48000$
- Welche Studie ist „interessanter“?

Signfikanz vs. Relevanz

- **Studie 1:** $k_1 = 11$, $n_1 = 90$ | $k_2 = 35$, $n_2 = 110$

  → p = 0.001888
- **Studie 2:** $k_1 = 12500$, $n_1 = 51000$ | $k_2 = 11200$, $n_2 = 48000$

  → p = 0.000015
- Welche Studie ist „interessanter“?

Signfikanz vs. Relevanz

- **Studie 1:** $k_1 = 11$, $n_1 = 90$ | $k_2 = 35$, $n_2 = 110$

  → p = 0.001888 $\quad\quad$ **$\mu_2 - \mu_1 \geq 7.56$ / 100 Wörter**

- **Studie 2:** $k_1 = 12500$, $n_1 = 51000$ | $k_2 = 11200$, $n_2 = 48000$

  → p = 0.000015 $\quad\quad$ **$\mu_1 - \mu_2 \geq 0.64$ / 100 Wörter**

- Welche Studie ist „interessanter“?
  - Unterschied bei Studie 2 ist höher signifikant, aber linguistisch (vermutlich) nicht relevant

ÜBUNG 2

Übung – Komposita in L2-Deutsch

- Wir nehmen nun eine größere (echte ☺) Stichprobe von Substantiven, um herauszufinden, ob der Unterschied zwischen L1 und L2 tatsächlich signifikant ist (Daten aus dem Falko-Korpus, Reznicek et al. 2010)
- Falls kein Zufall dahinter steckt, werden auch weitere Daten einen signifikanten Unterschied aufweisen

Daten herunterladen

- Wer die Daten noch nicht hat:
  http://u.hu-berlin.de/falko_comp
- Datei speichern:

  `compound_noun_falko_all_v2.2.tab`

Daten einlesen

```
> comp_data <- read.table(file.choose(), header=TRUE, as.is=TRUE)
> head(comp.data,12)
              tok           lemma          head modifier trancription_name  L1     type
1      Videospiele      Videospiel         Spiel    Video    dcs001_2007_10 deu compound
2      Haftstrafen      Haftstrafe        Strafe     Haft    dcs001_2007_10 deu compound
3        Volksmund       Volksmund          Mund    Volks    dcs001_2007_10 deu compound
4         TV-Shows         TV-Show          Show      TV-    dcs001_2007_10 deu compound
5  Kleinkriminelle Kleinkriminelle   Kriminelle    Klein    dcs001_2007_10 deu compound
6  Extrembeispiele  Extrembeispiel     Beispiel   Extrem    dcs001_2007_10 deu compound
7  Verkehrsunfälle  Verkehrsunfall       Unfall Verkehrs    dcs001_2007_10 deu compound
8     Feststellung    Feststellung     Stellung     Fest    dcs001_2007_10 deu compound
9     Gegenbeweise     Gegenbeweis       Beweis    Gegen    dcs001_2007_10 deu compound
10        Personen          Person       Person     NULL    dcs001_2007_10 deu  simplex
11         Ansicht         Ansicht      Ansicht     NULL    dcs001_2007_10 deu  simplex
12    Kriminalität    Kriminalität Kriminalität     NULL    dcs001_2007_10 deu  simplex
```

* Wer keine Umlaute sieht, kann folgende Option verwenden:

```
> read.table(file.choose(), header=TRUE, fileEncoding="UTF-8")
```

attach()

- Mit attach() kann man auf einzelne Spalten leichter zugreifen:

```
> attach(comp_data) #Direktzugriff auf Spalten

> head(modifier) #erste Werte von modifier
[1] Video  Haft   Volks  TV-    Klein  Extrem
1124 Levels: 1-Euro- Abbestellungs Abbrecher ... Zwang

> levels(L1) #alle Ausprägungen von L1
 [1] "afr" "cat" "ces" "cma" "dan" "deu" "ell" "eng"
       "fin" "fra" "hbs" "hin" "hun" "iii" "ita" "jpn"
    "kik" "kor" "kua" "lub" "luy" "nde" "nld" "nor"
    "pol" "ron" "rus" "slk" "sme"
[30] "spa" "sqi" "swe" "tat" "tur" "ukr" "uzb" "vie"
    "zho"

> length(levels(L1)) #Länge der Liste der Ausprägungen
[1] 38
```

Daten darstellen

- Uns interessiert die Verteilung von Komposita/Simplizia:

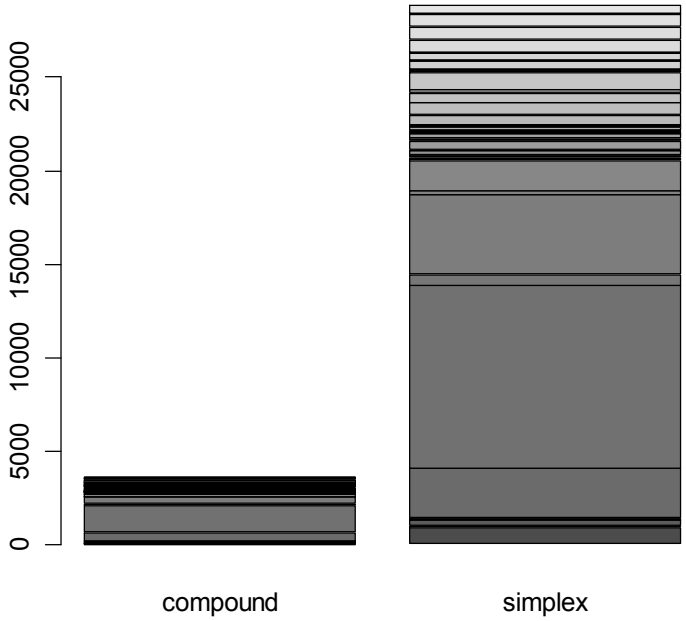
```
> table(L1,type)
     type
L1     compound simplex
  afr        81     864
  cat         9      88
  ces        26     358
  cma        17      87
   ...
   vie        22     106
  zho        42     400
```

Unser 1. Balkendiagramm ☺

- Das war schwer zu lesen...
- Wir hätten gern ein Diagramm dieser Daten
- Das geht einfach mit R:

```
> barplot(table(L1,type))
```

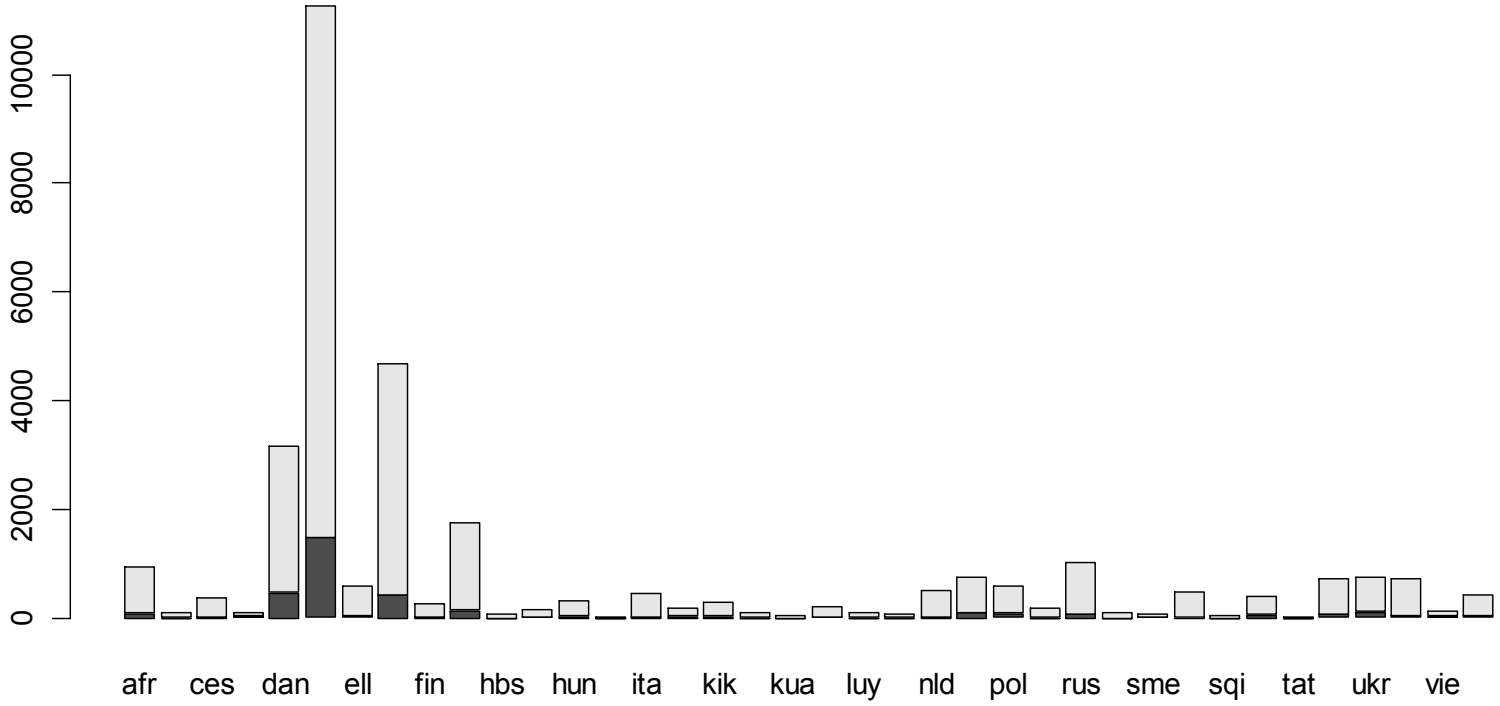
???



Unser 2. Balkendiagramm

- die Funktion barplot() erwartet eine Spalte für jeden Balken, keine Zeile
- Tabelle transponieren mit t():

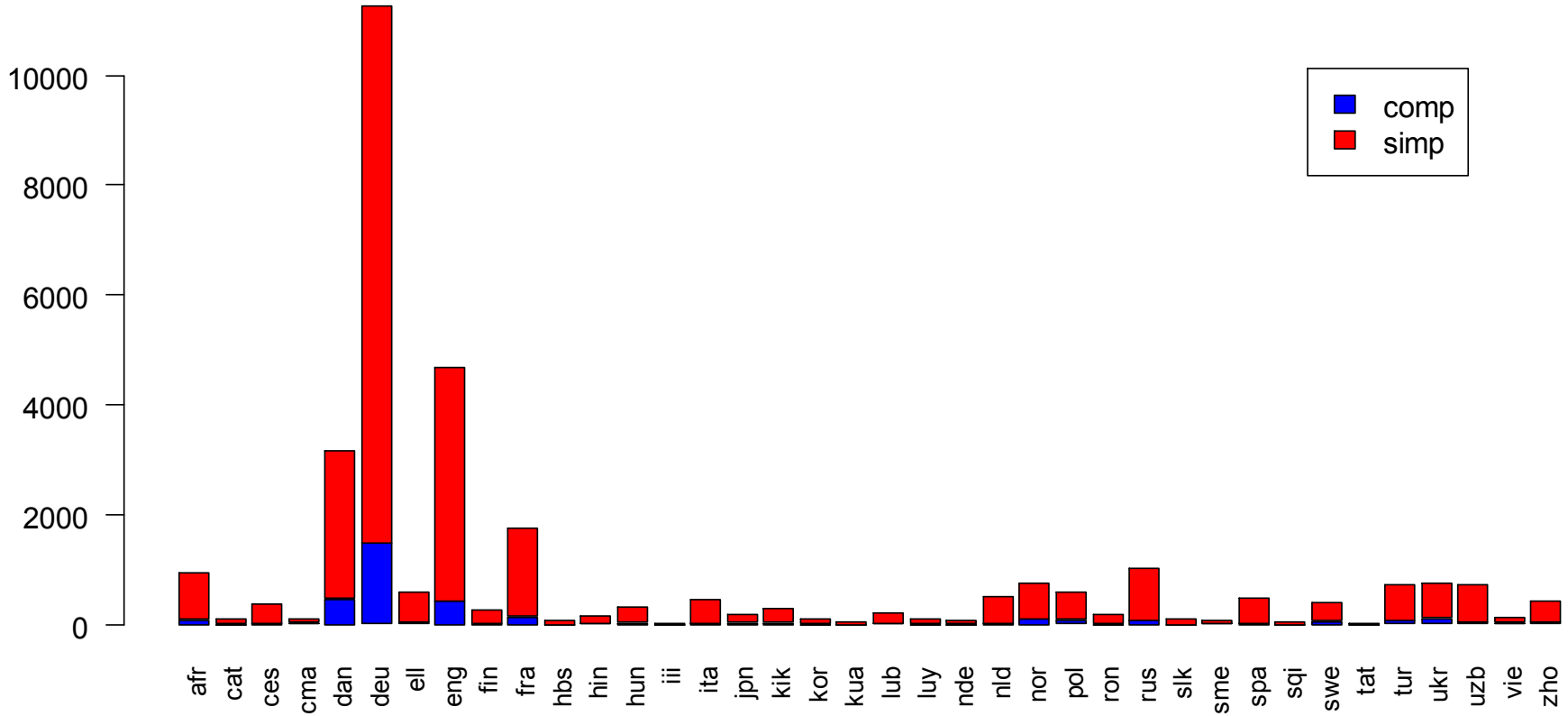
```
> barplot(t(table(L1,type)))
```



Unser 3. Balkendiagramm

- Und ein bisschen hübscher:

```
> barplot(t(table(L1,type)), cex.names=0.6, las=2,
col=c("blue","red"))
> legend("topright", c("comp","simp"), fill=c("blue","red"))
```



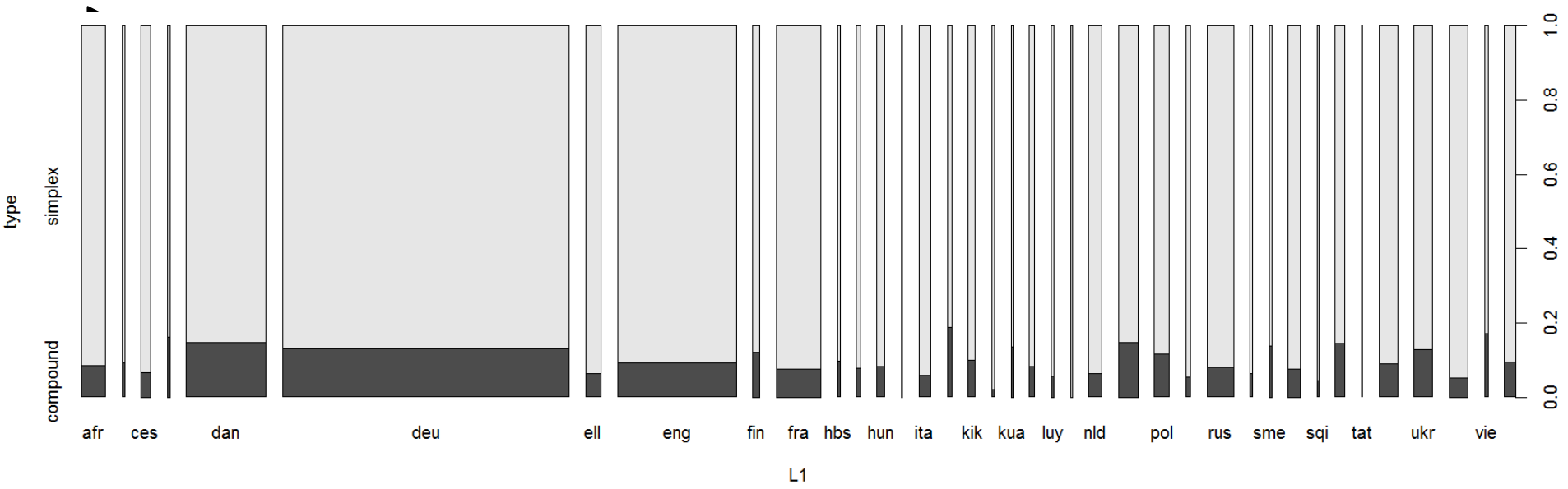
Ein besseres Diagramm

- Das ist nicht so hilfreich...
- Unterschiedlich viele Daten aus jeder L1
  - Anteile nicht direkt vergleichbar
  - Zuverlässigkeit der Zahlen nicht vergleichbar

Ein besseres Diagramm

- Eine bessere Darstellung mit spineplot()
- spineplot() möchte Zeilen, nicht Spalten, daher **nicht** transponieren:

```
> spineplot(table(L1,type))
```



L1 vs. L2

- Als nächstes interessiert uns die allgemeine Frage: verwenden Lerner weniger Komposita?
- Wir brauchen eine neue Tabelle:

```
> L1L2_tab <- table(L1=="deu",type)
> L1L2_tab
       type
        compound simplex
  FALSE     2113   19067
  TRUE      1476    9789
```

prop.test() mit Tabellen

- Wir könnten die Zahlen in `prop.test()` eingeben
- Bequemerweise darf man Tabellen direkt benutzen:

```
> prop.test(L1L2_tab)

        2-sample test for equality of proportions with
    continuity correction

data:  L1L2_tab
X-squared = 72.7311, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03875335 -0.02376939
sample estimates:
    prop 1     prop 2
0.09976393 0.13102530
```

Verhalten sich die Lerner gleich?

- Wir hatten bisher immer einen Vergleich zweier Stichproben
- Jetzt möchten wir wissen, ob alle Lerner Gruppen sich vergleichbar verhalten
- prop.test() kann auch mehrere Gruppen vergleichen:

```
> prop.test(c(a,b,c,...,z),c(A,B,C,...,Z))
```

Verhalten sich die Lerner gleich?

- Noch einfacher geht es mit einer großen Tabelle (immer nur mit 2 Spalten):

```
> L2.data=table(L1[L1!="deu"],type[L1!="deu"])
> prop.test(L2.data)

        37-sample test for equality of proportions without continuity correction

data:  L2.data
X-squared = 248.0218, df = 36, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
    prop 1     prop 2     prop 3     prop 4     prop 5     prop 6     prop 7
0.08571429 0.09278351 0.06770833 0.16346154 0.14711477 0.06435644 0.09308284
   ...

Warning message:
In prop.test(L2.data) : Chi-squared approximation may be incorrect
```

Ein Unterschied

- Mindestens eine Gruppe ist mit den anderen nicht vergleichbar
- Wir schauen uns den Anteil der Komposita in jeder Gruppe an:

```
> L2.props = L2.data[,"compound"]/(L2.data[,"simplex"]+L2.data[,"compound"])
> L2.props
       afr        cat        ces        cma        dan        ell        eng
0.08571429 0.09278351 0.06770833 0.16346154 0.14711477 0.06435644 0.09308284

...
```

- Einige Gruppen sind ähnlich, bspw. "afr", "eng"

Vergleich ausgewählter Gruppen

```
> eng_afr.data=table(L1[L1=="eng"|L1=="afr"], type[L1=="eng"|
L1=="afr"]) #englisch oder afrikaans
> eng_afr.data
        compound simplex
  afr         81     864
  eng        436    4248
> prop.test(eng_afr.data)
        2-sample test for equality of proportions with
continuity correction
data:  eng_afr.data
X-squared = 0.4273, df = 1, p-value = 0.5133
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02769702  0.01295993
sample estimates:
    prop 1     prop 2
0.08571429 0.09308284
```

Weitere Übungen

- Sortieren Sie die Proportionen mit sort()
- Stellen Sie die sortierten Proportionen als Balkendiagramm mit barplot() dar
- Vergleichen Sie die zwei Gruppen mit den meisten Komposita mit prop.test() (nutzen Sie wieder die Tabelle L2.data)
- Verhalten sich die romanischen Sprache ähnlich? (cat,fra,ita,spa,ron) Testen Sie, ob die Unterschiede signifikant sind und stellen Sie die Daten mit spineplot() dar

Endlich!

KAFFEEPAUSE

ASSOZIATION UND UNABHÄNGIGKEITSTEST

Vergleich von Merkmalen

- Bisher: Vergleich von zwei Stichproben
  - aus verschiedenen Grundgesamtheiten
- Jetzt: Vergleich von zwei Merkmalen
  - unterschiedliche Eigenschaften derselben Token
  - eine Stichprobe aus einer Grundgesamtheit
- Fallbeispiel: englische Dativalternation
  - Peter gave [$_{NP}$ his friend] the book
    vs. Peter gave the book [$_{PP}$ to his friend]
  - Besteht ein Zusammenhang mit Informationsstatus?

Dativalternation

- Was für eine Stichprobe wird benötigt?
  - Token = Instanzen von VPen mit Dativobjekt
  - (manuell) annotiert: Dativ-Realisierung (NP/PP), Informationsstatus (new, given, accessible), …
  - aus welchen Textquellen?
- Hier: Teilmenge von Bresnan et al. (2007)
  - *give*-VPen aus *Wall Street Journal* (Zeitungsartikel) und Switchboard-Dialogen (gesprochene Sprache)
  - komplett in R-Paket `languageR` (Baayen 2008)

Dativalternation

```
> Give <- read.delim("dative_give.txt")
> Give <- read.delim(file.choose()) # alternativ
> dim(Give)
[1] 250   6
> head(Give, 5)
```

Voreinstellung für TAB-getrennte Tabellen mit Kopfzeile

	Recip	AccessRec	AccessTheme	AnimRec	AnimTheme	VerbClass
1	PP	new	new	animate	inanimate	transfer
2	PP	given	accessible	animate	inanimate	transfer
3	NP	new	accessible	inanimate	inanimate	abstract
4	PP	new	new	inanimate	inanimate	abstract
5	NP	accessible	accessible	animate	inanimate	abstract

Dativalternation

```
> summary(Give) # Überblick über Kategorien

 Recip          AccessRec       AccessTheme
 NP:199   accessible: 70   accessible:134
 PP: 51   given     :146   given     : 15
          new       : 34   new       :101

     AnimRec         AnimTheme           VerbClass
 animate  :209   animate  :  2   abstract     :224
 inanimate: 41   inanimate:248   communication: 10
                                 transfer     : 16
```

Assoziation

- Wie können wir feststellen, ob es einen Zusammenhang zwischen den Merkmalen Recip und AccessRec gibt? (Assoziation)
- Wie oft kommen bestimmte Merkmalsausprägungen miteinander vor?
  - sog. Kreuztabelle („contingency table“)
  - Wie erstellt man eine Kreuztabelle in R?

```
> kt <- table(Give$Recip, Give$AccessRec)
```

Kreuztabellen

```
> kt <- table(Give$Recip, Give$AccessRec)
> kt
```

	access	given	new
NP	48	137	14
PP	22	9	20

- Zusammenhang erkennbar?
- Woran?
- Insgesamt oder für einzelne Felder?
- Signifikanz?

Statistische Unabhängigkeit

- Ein Zusammenhang besteht, wenn bestimmte Merkmalskombinationen auffallend häufig oder selten auftauchen
  - sog. Kookkurrenz von Merkmalsausprägungen
- Hängt von Häufigkeit einzelner Kategorien ab
  - Erwartung: viele Kookkurrenzen bei zwei häufigen Kategorien, wenige bei zwei seltenen Kategorien
  - unter der Hypothese, dass die Merkmale statistisch unabhängig sind, lässt sich die erwartete Häufigkeit mit einer mathematischen Formel berechnen

Notation für Kreuztabellen

```
> kt <- table(Give$Recip, Give$AccessRec)
> kt
```

	access	given	new
NP	48	137	14
PP	22	9	20

	access	given	new	
NP	n_{11}	n_{12}	n_{13}	$= n_{1\bullet}$
PP	n_{21}	n_{22}	n_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Notation für Kreuztabellen

```
> kt <- table(Give$Recip, Give$AccessRec)
> kt
> addmargins(kt)
```

	access	given	new	
NP	48	137	14	= 199
PP	22	9	20	= 51
	= 70	= 146	= 34	250

	access	given	new	
NP	n_{11}	n_{12}	n_{13}	$= n_{1\bullet}$
PP	n_{21}	n_{22}	n_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Erwartete Häufigkeit

- H_0: Unabhängigkeitshypothese
  - Wk für NP = $n_{1\bullet} / n$
  - Wk für given = $n_{\bullet 2} / n$
  - Kookkurrenz-Wk = $n_{1\bullet}\, n_{\bullet 2} / n^2$
  - Erwartete Häufigkeit e_{12} für n Token

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

E	access	given	new	
NP				= 199
PP				= 51
	= 70	= 146	= 34	250

E	access	given	new	
NP		e_{12}		$= n_{1\bullet}$
PP				$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Erwartete Häufigkeit

- H_0: Unabhängigkeitshypothese
  - Wk für NP = $n_{1\bullet} / n$
  - Wk für given = $n_{\bullet 2} / n$
  - Kookkurrenz-Wk = $n_{1\bullet}\, n_{\bullet 2} / n^2$
  - Erwartete Häufigkeit e_{12} für n Token

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

E	access	given	new	
NP	55.7	116.2	27.1	= 199
PP	14.3	29.8	6.9	= 51
	= 70	= 146	= 34	250

E	access	given	new	
NP	e_{11}	e_{12}	e_{13}	$= n_{1\bullet}$
PP	e_{21}	e_{22}	e_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Erwartete Häufigkeit

- Erwartete Häufigkeit in R (für Profis)

```
> n_rows <- rowSums(kt)
> n_cols <- colSums(kt)
> n <- sum(kt)
> kt.e <- outer(n_rows, n_cols) / n
> round(kt.e, 1)
> addmargins(kt.e)
```

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

E	access	given	new	
NP	55.7	116.2	27.1	= 199
PP	14.3	29.8	6.9	= 51
	= 70	= 146	= 34	250

E	access	given	new	
NP	e_{11}	e_{12}	e_{13}	$= n_{1\bullet}$
PP	e_{21}	e_{22}	e_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Assoziation & Chi-Quadrat-Test

- Vergleich von erwarteten und tatsächlichen Häufigkeiten
- Hypothesentest für H_0: Merkmale sind unabhängig
- Chi-Quadrat-Statistik → Signifikanzwert

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

E	access	given	new	
NP	55.7	116.2	27.1	= 199
PP	14.3	29.8	6.9	= 51
	= 70	= 146	= 34	250

	access	given	new	
NP	48	137	14	= 199
PP	22	9	20	= 51
	= 70	= 146	= 34	250

Chi-Quadrat-Test in R

```
> ergebnis <- chisq.test(kt)
> ergebnis

        Pearson's Chi-squared test

data:  kt
X-squared = 54.376, df = 2, p-value = 1.557e-12
```

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Resultat: $p = 1.557 \times 10^{-12} = 0.000000000001557$

  - R-Profis können X^2 auch direkt berechnen:

    ```
    > X2 <- sum( (kt - kt.e)^2 / kt.e )
    ```

Chi-Quadrat-Test in R

- Welche Merkmalskombinationen sind auffällig?
- Standardisierte Abweichung (z-score) für jede Kombination
- Unter H_0: jedes z_{ij} folgt einer Standardnormalverteilung

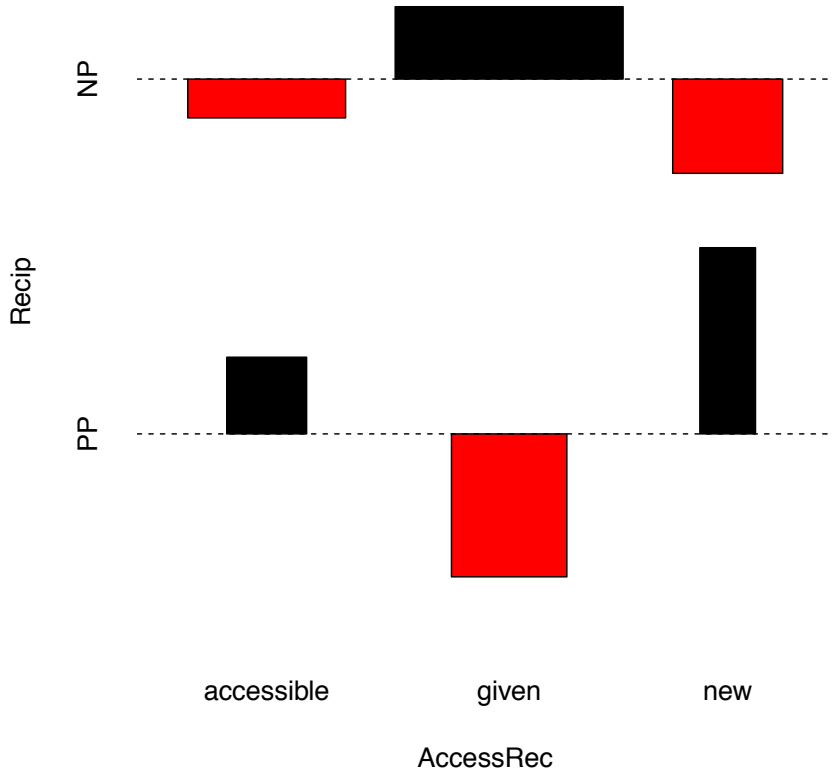
$$z_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

```
> round(ergebnis$expected, 1)
> round(ergebnis$residuals, 2) # = z_ij
```

Chi-Quadrat-Test in R

```
> round(ergebnis$residuals, 2)
> 2 * pnorm(2.51, lower=FALSE) # NP / new
> assocplot(t(kt)) # t() = transponieren
```

Z	access	given	new
NP	–1.03	1.93	–2.51
PP	2.04	–3.81	4.96



Signifikanz vs. Relevanz

- Auch standardisierte Abweichung z_{ij} ist ein Signifikanzwert → Trennschärfe
  - Wie kann man die Effektgröße messen?
- Globales Maß: *V*-Koeffizient von Cramér

```
> X2 = ergebnis$statistic
> n = sum(kt)
> faktor = min(dim(kt)) - 1
> V = sqrt(X2 / (n * faktor))
> V  # Wertebereich 0 … 1
```

$$V = \sqrt{\frac{X^2}{n \cdot \min\{r-1, c-1\}}}$$

Mini-Übung

- Gibt es Evidenz für andere Einflüsse auf die Dativalternation (z.B. Belebtheit)?
- Zusammenhang zw. Informationsstatus von Dativobjekt und Akkusativobjekt?
- Ändert sich der Signifikanzwert bei einer größeren Stichprobe?

Mini-Übung: Lösungen

```
> kt2 <- table(Give$Recip, Give$VerbClass)
> chisq.test(kt2)

    Pearson's Chi-squared test

data:  kt2
X-squared = 3.8243, df = 2, p-value = 0.1478 (?)
Warning message:
In chisq.test(kt2) : Chi-squared approximation may be
incorrect
> fisher.test(kt2) # exakter Test (aufwendig)

    Fisher's Exact Test for Count Data

data:  kt2
p-value = 0.1301
alternative hypothesis: two.sided
```

Mini-Übung: Lösungen

```
> kt3 <- table(Give$AccessRec, Give$AccessTheme)
> kt3
             accessible given new
  accessible         49     2  19
  given              76    10  60
  new                 9     3  22
> chisq.test(kt3)
    Pearson's Chi-squared test
data:  kt2
X-squared = 18.0605, df = 4, p-value = 0.001201
Warning message: …
> fisher.test(kt3)
```

Mini-Übung: Lösungen

```
> fisher.test(kt2)$p.value
[1] 0.1301211

> fisher.test(2 * kt2)$p.value
[1] 0.02515787

> fisher.test(5 * kt2)$p.value
[1] 0.000151227

> fisher.test(10 * kt2)$p.value
[1] 3.426715e-08
```

Mini-Übung: Lösungen

```
> cramer = function (kt) {
    X2 = chisq.test(kt)$statistic
    n = sum(kt)
    faktor = min(dim(kt)) - 1
    sqrt(X2 / (n * faktor))
  }

> chisq.test(kt2)$p.value
[1] 0.001200939
> cramer(kt2)
0.1900553
> chisq.test(10 * kt2)$p.value
[1] 5.528046e-38
> cramer(10 * kt2)
0.1900553
```

ÜBUNG 3

Mehrere kategoriale Variablen

- Mit prop.test() haben wir nur Gruppen im Hinblick auf einen Wert verglichen
- Wir hatten keinen Weg, mehr als zwei Variablenausprägungen zu behandeln
  (Kompositum oder nicht; was ist mit Simplex, Derivation und Komposition? Zwei- bzw. mehrgliedrige Komposita?)
- Was machen wir, wenn wir mehr haben?

Übung

- Wir arbeiten als nächstes mit informationsstrukturell annotierten Daten
- Zwei Variablen: (Guidelines des SFB632, Dipper et al. 2007)
  - infstat: **giv**(en), **new**, **acc**(essible), **idiom**
  - topic: **ab**(outness), **fs** = framesetter, **nt** = not-topic
- Genauere Unterteilung in "active" bzw. "inactive", "inferrable", "generic" etc.

Übung

- Forschungsfragen:
  - Wie hängt Topikalität mit Bekanntheit zusammen?
  - Gibt es Unterschiede dabei zwischen Vorerwähntheit und Erschließbarkeit?
  - Framesetter und Aboutness-Topik?
- Formulierung von Hypothesen
- Erstellung von passenden Kreuztabellen

Daten herunterladen

- Wer die Daten noch nicht hat:

  http://u.hu-berlin.de/infstat_data
- Datei speichern:

  `infstat_data.tab`

Daten einlesen

```
> infstruct.data <- read.table(file.choose(), header=TRUE, as.is=TRUE, fileEncoding="UTF-8")
> head(infstruct.data)
  ID                  referent infstat topic infstat_fine
1  1           Die_Jugendlichen     new    ab          new
2  2                     Zossen     new    ab          new
3  3              ein_Musikcafé     new    nt          new
4  4                        Das     giv    nt   giv-active
5  5                        sie     giv    ab   giv-active
6  6 der_ersten_Zossener_Runde     new    fs          new
```

Erste Hypothese

- Frage: hängt Informationsstatus mit Topikalitätzusammen?
- Nullhypothese **H0**: Kein Zusammenhang zwischen den Variablen
- Wir fangen an mit einer groben Untersuchung: referentieller Inf-Status und binäre Topikalität

  → Ausprägung "Idiom" entfernen

  → Ausprägungen "fs" und "ab" zusammentun

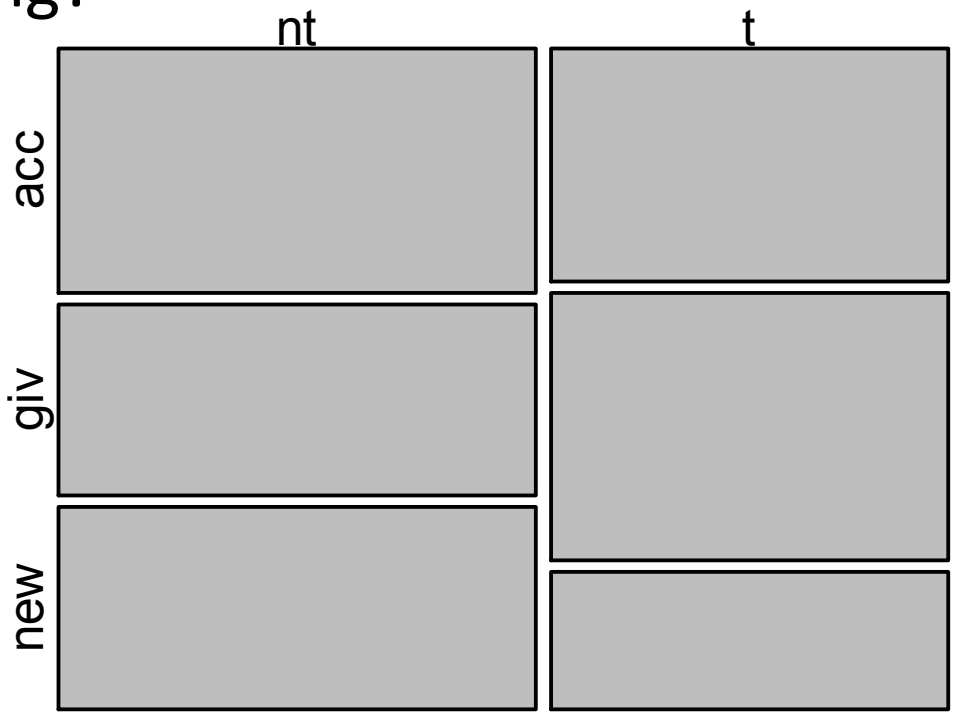
Tabelle erstellen

```
> no_idiom = subset(infstruct.data,infstat!="idiom")
> topic_infstat = table(ifelse(no_idiom$topic=="nt","nt","t"), no_idiom$infstat)
> topic_infstat
```

```
   nt  72  56  60
   t   57  65  34
```

plot(topic_infstat)

Zusammenhang?



Zur Erinnerung: was ist H_0?

- Heißt kein Zusammenhang: alles in jeder Zelle
-

```
> addmargins(topic_infstat)

     acc giv new Sum
nt    72  56  60 188
t     57  65  34 156
Sum  129 121  94 344
```

- Kombination?

Zur Erinnerung: was ist H_0?

- Nicht wirklich: es kann sein, dass es mehr Nicht-Topiks gibt als Topiks
- Es kann sein, dass die meisten Referenten neu sind
- Aber es darf keine Interaktion geben (mehr neu wenn nicht Topik)

```
> chisq.test(topic_infstat)$expected
```

```
       acc      giv      new
nt 70.5 66.12791 51.37209
t  58.5 54.87209 42.62791
```

gleich verteilt

gleich verteilt

Zusammenhang testen

```
> chisq.test(topic_infstat)

        Pearson's Chi-squared test

data:  topic_infstat
X-squared = 6.6862, df = 2, p-value = 0.03533
```

- Irgendwo sind Zeilen und Spalten **nicht** unabhängig, H0 gilt nicht

Residuen

- Die residuen drücken den Unterschied zwischen Erwartung und Beobachtung aus:
  - Residuals = (observed - expected) / sqrt(expected)

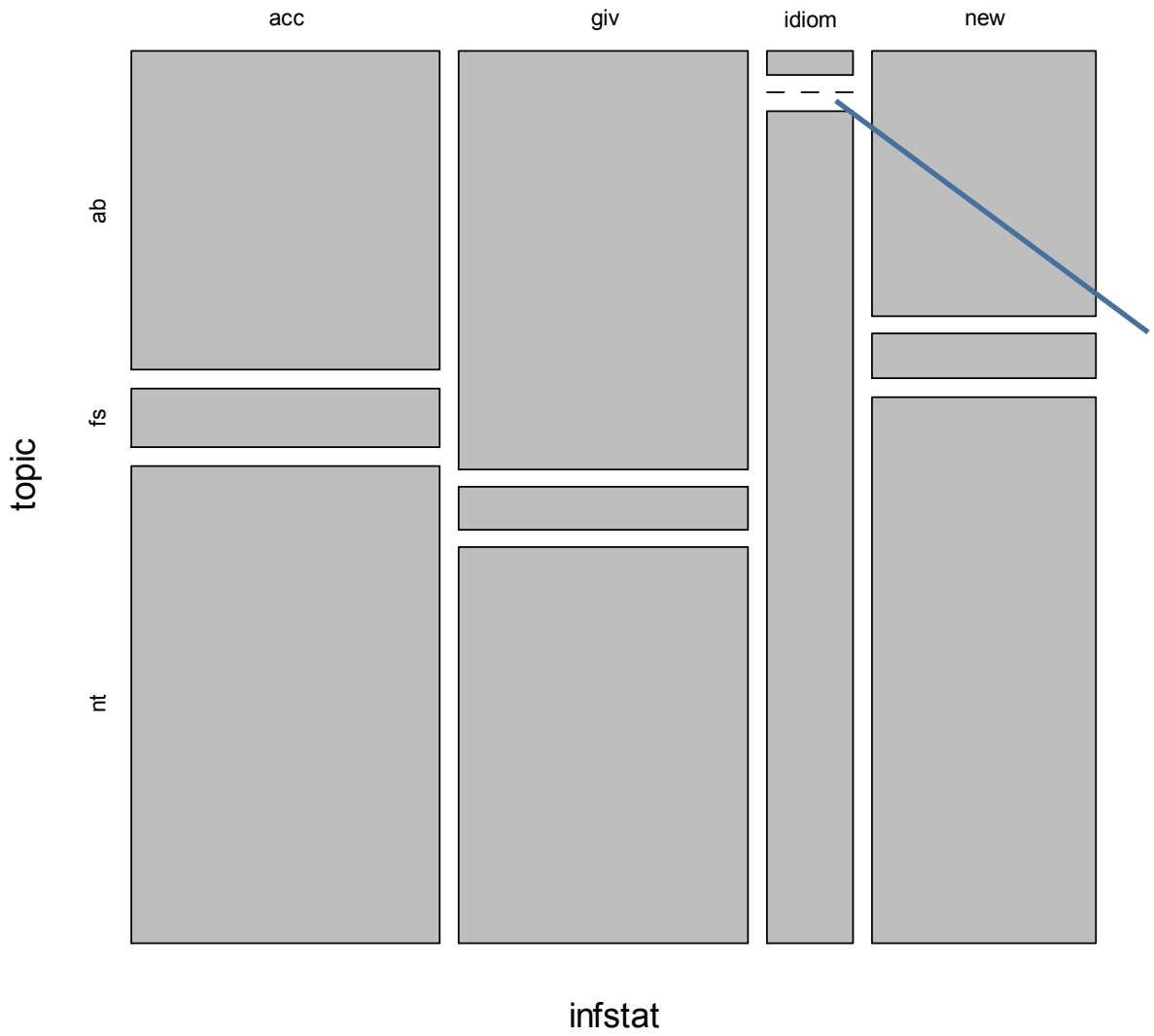
```
> chisq.test(topic_infstat)$residuals

            acc        giv        new
nt   0.1786474 -1.2454529  1.2037653
t   -0.1961161  1.3672374 -1.3214735
```

Zweite Kreuztabelle

- Bekommen wir ein besseres Bild mit allen Kategorien von topic ~ infstat?

```
> is_tab = table(infstat, topic)
> is_tab
        topic
infstat ab fs nt
  acc   48  9 72
  giv   59  6 56
  idiom  1  0 35
  new   29  5 60

> plot(is_tab)
```

plot(is_tab)


acc
giv
idiom
new
ab
fs
nt
topic
infstat
Keine Fälle von idiom & fs

Wann gibt es "ab" und "idiom"?

```
> infstat_data[infstat_data$infstat=="idiom" &
infstat_data$topic=="ab",]
     ID     referent infstat topic infstat_fine
155 155 Das_Füllhorn   idiom    ab        idiom
```

- Satz:

"***Das Füllhorn*** *schließlich schüttet man über Kitas und Schulen aus.*"

chisq.test

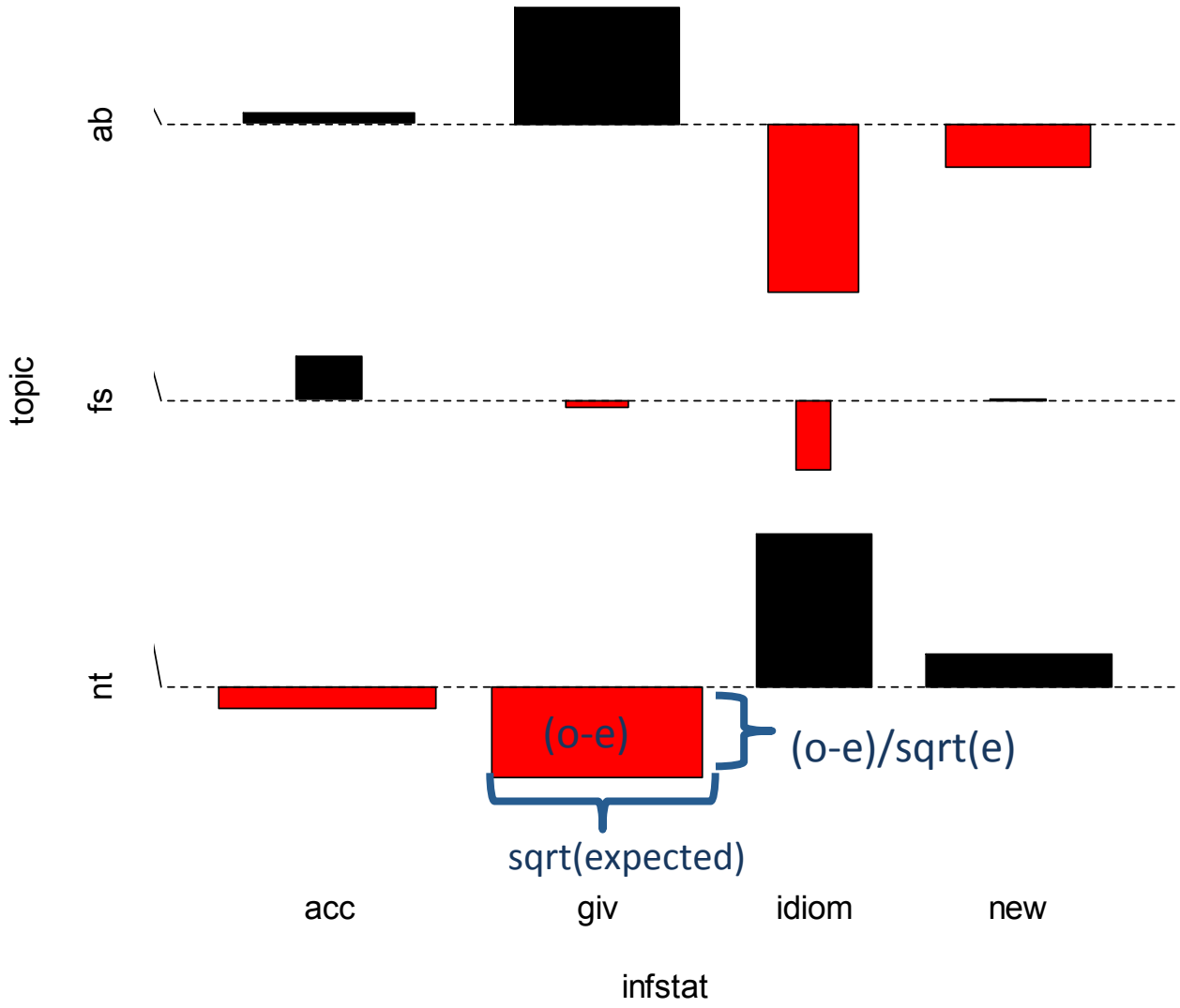
- Gibt es Zusammenhänge mit allen Variablenausprägungen?

```
> is_test = chisq.test(is_tab)
Warning message:
In chisq.test(is_tab) : Chi-squared approximation
   may be incorrect

> is_test

        Pearson's Chi-squared test

data:  is_tab
X-squared = 32.7535, df = 6, p-value = 1.17e-05
```

Welche Zusammenhänge fallen auf?

```
> is_test$residuals
        topic
infstat           ab          fs           nt
  acc     0.21879436  0.84835502 -0.42555433
  giv     2.32804365 -0.14599183 -1.78101040
  idiom  -3.32505567 -1.37649440  3.01842175
  new    -0.83990410  0.02366243  0.65123443
```

assocplot(is_tab)



Warnung?

- Zu wenig Daten zu den Idiomen:

```
> is_tab
        topic
infstat ab fs nt
   acc   48  9 72
   giv   59  6 56
   idiom  1  0 35
   new   29  5 60
```

Genauerer Test mit fisher.test()

```
> fisher.test(xtabs(~ topic+infstat, data=infstat_data), workspace=2e6)

        Fisher's Exact Test for Count Data

data:  xtabs(~topic + infstat, data = infstat_data)
p-value = 9.734e-07
alternative hypothesis: two.sided
```

Weitere Übungen

- Was passiert, wenn man die noch feineren Kategorien in *infstruct.data$infstat_fine* nimmt?
- Was passiert, wenn man "acc" und "giv" zusammen mit "new" kontrastiert?

SCHLUSSWORTE + EVALUATION

Literaturempfehlungen

Für den Einstieg:

- Gries, Stefan Th. (2008). *Statistik für Sprachwissenschaftler.* Göttingen: Vandenhoeck & Ruprecht.
- Gries, Stefan Th. (2009). *Statistics for Linguistics with R: A Practical Introduction.* Berlin: Mouton de Gruyter, Berlin.
- Oakes, Michael P. (1998). *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.
- Bortz, Jürgen (2005). *Statistik für Sozialwissenschaftler.* Heidelberg: Springer.

Fortgeschrittene Methoden:

- Baayen, R. Harald (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics.* Cambridge: Cambridge University Press.
- Rietveld, Toni & van Hout, Roeland (2005). *Statistics in Language Research: Analysis of Variance.* Berlin/New York: Mouton de Gruyter.

Online-Materialien

- Handouts & Daten zum Tutorium:
  http://wordspace.collocations.de/doku.php/corpus_tutorial:dgfs2013
- SIGIL-Kurs (Baroni & Evert):
  http://sigil.r-forge.r-project.org/
- Butler, C. (1985). *Statistics in Linguistics.* Oxford: Blackwell.
  http://www.uwe.ac.uk/hlss/llas/statistics-in-linguistics/bkindex.shtml
- Galtonbrett-Simulation:
  http://www.math.psu.edu/dlittle/java/probability/plinko/
- R-Homepage (u.a. Lehrbücher & Online-Tutorien):
  http://www.r-project.org/