Korpuslinguistik und Statistik

(050370)

Fabian Barteld, M.A.
Fabian.Barteld@ruhr-uni-bochum.de

10. Sitzung

Kennzahlen von Verteilungen (I)

Mittelwerte, Lagemaße

	Nominal	Ordinal	Metrisch
Modalwert	x	x	x
Median		x	x
(Quartile)		x	x
arithm. Mittel			x

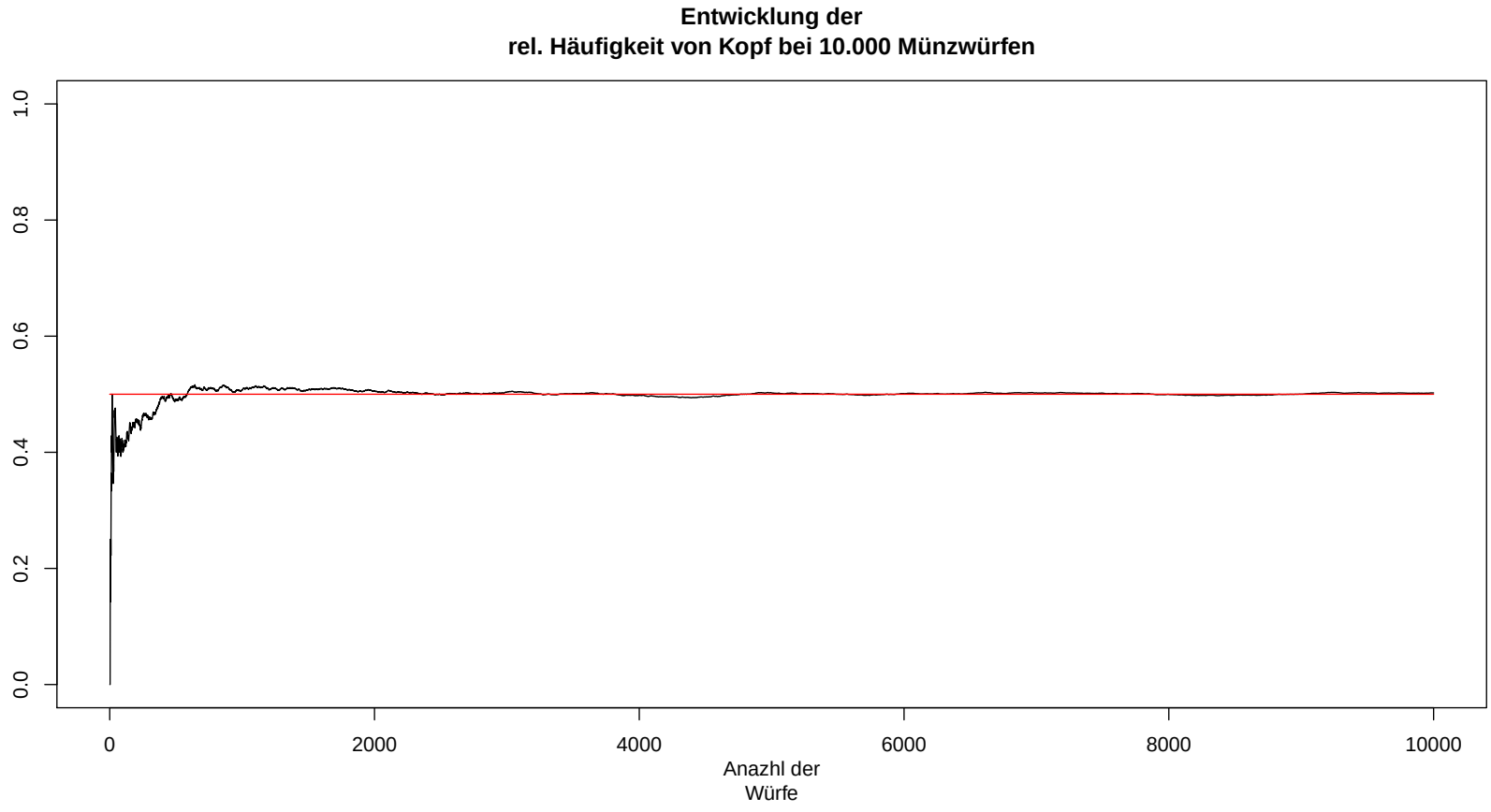
(vgl. Büchter & Henn, 2007, S. 79)

Kennzahlen von Verteilungen (II)

Streuungsmaße

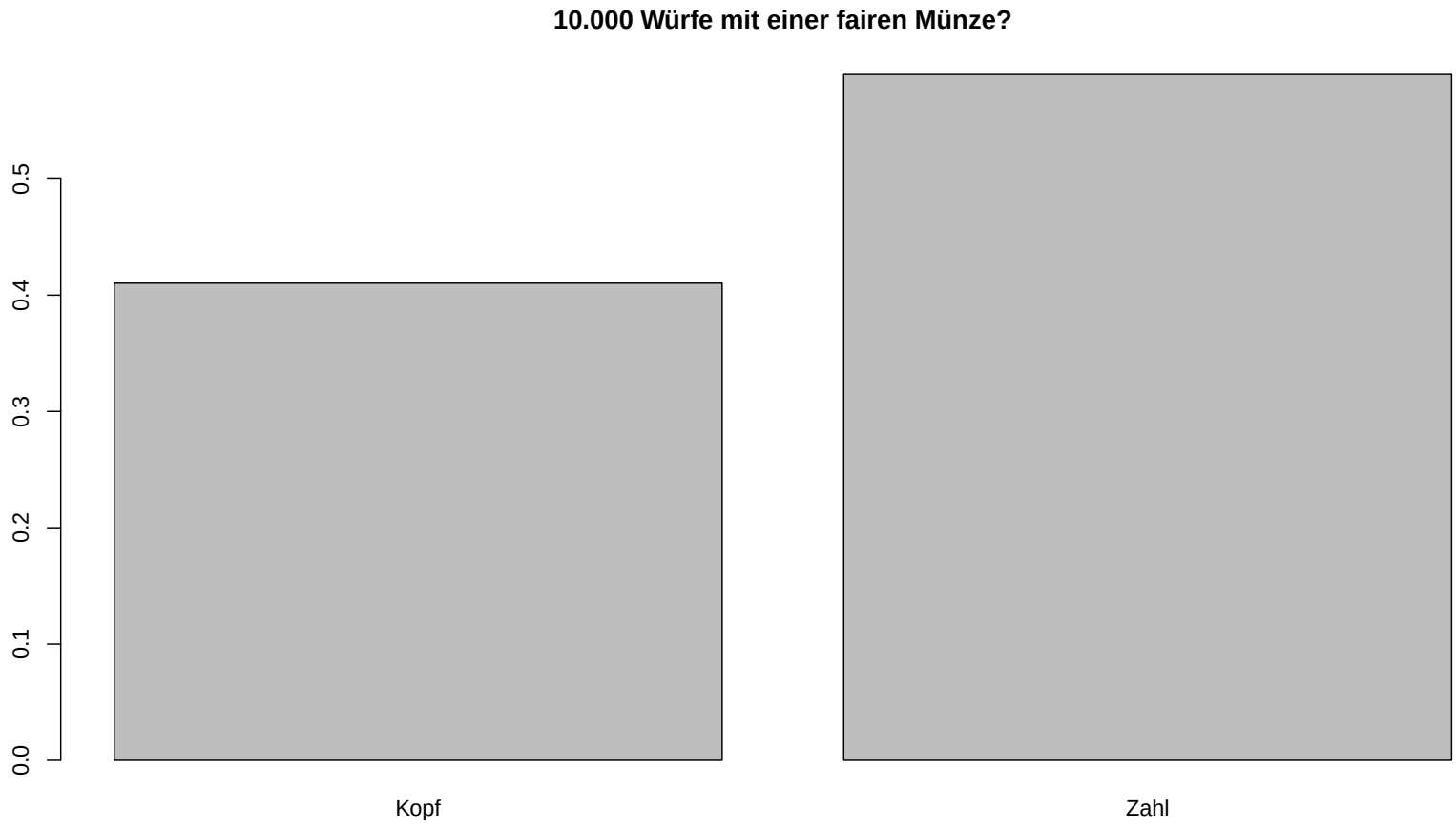
	Nominal	Ordinal	Metrisch
Quartile		x	x
Interquartilsabstand			x
Spannweite			x
Varianz			x

Das empirische Gesetz der großen Zahlen

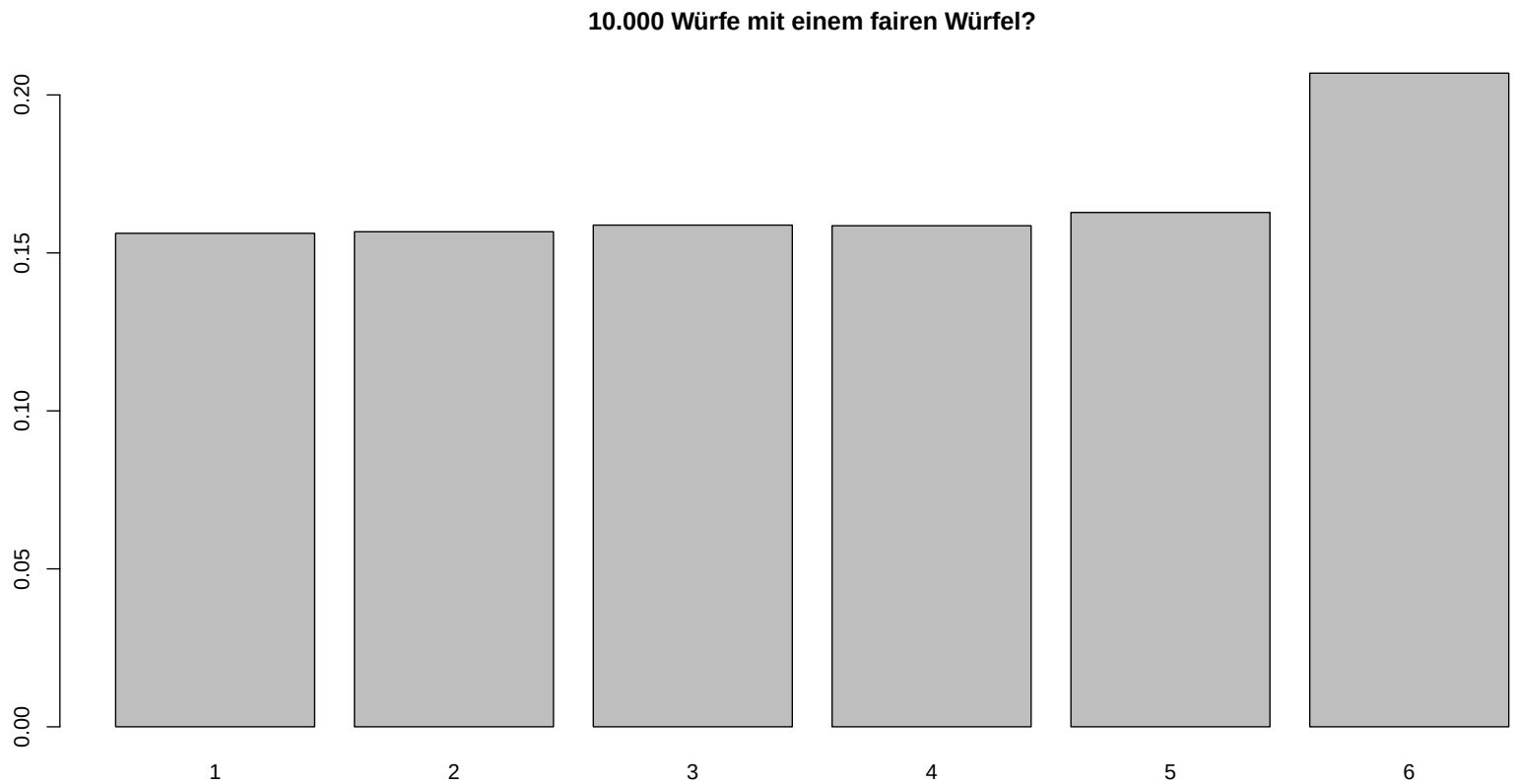


Binomialverteilung

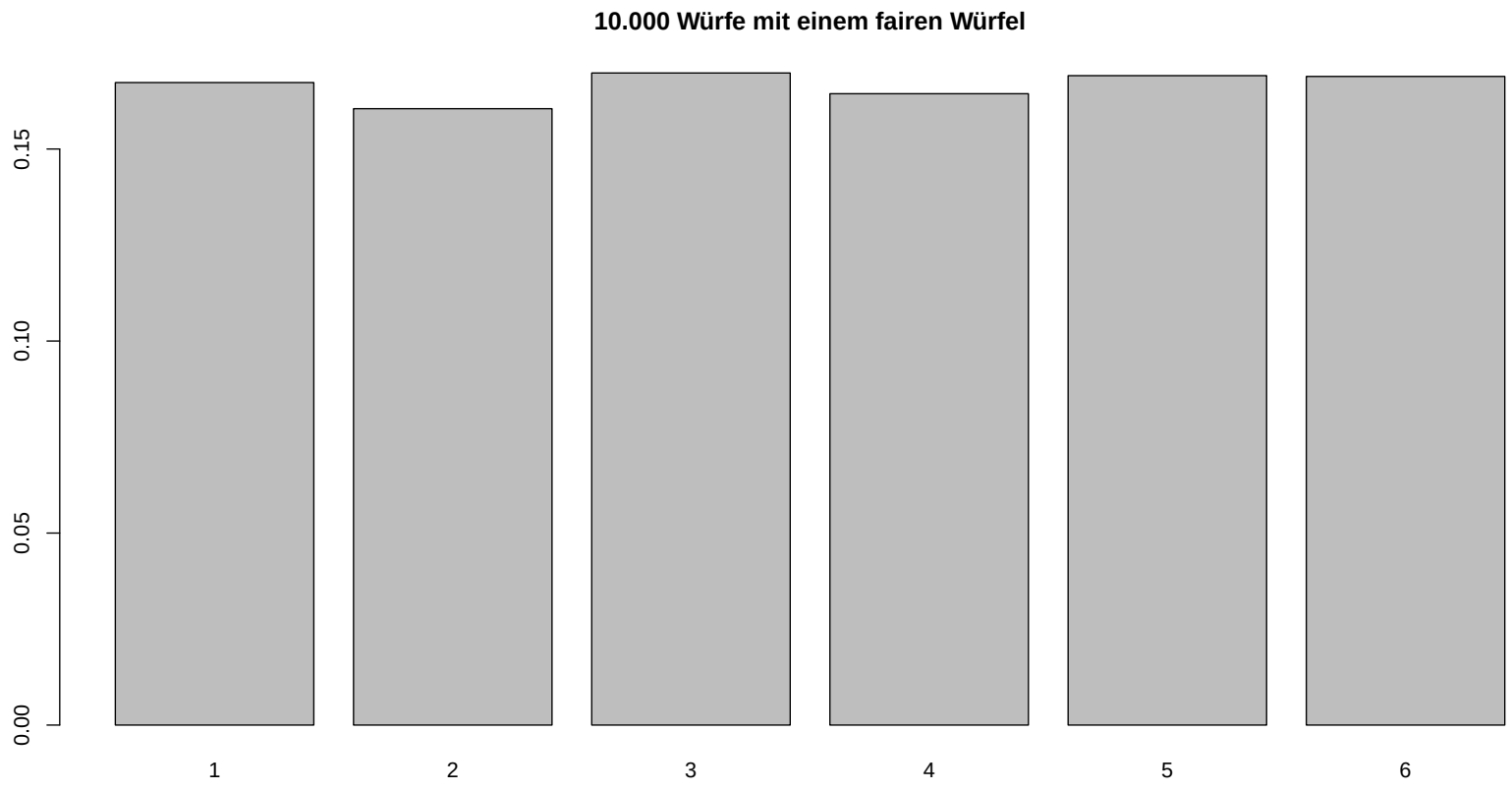
(Dichotome Variable)



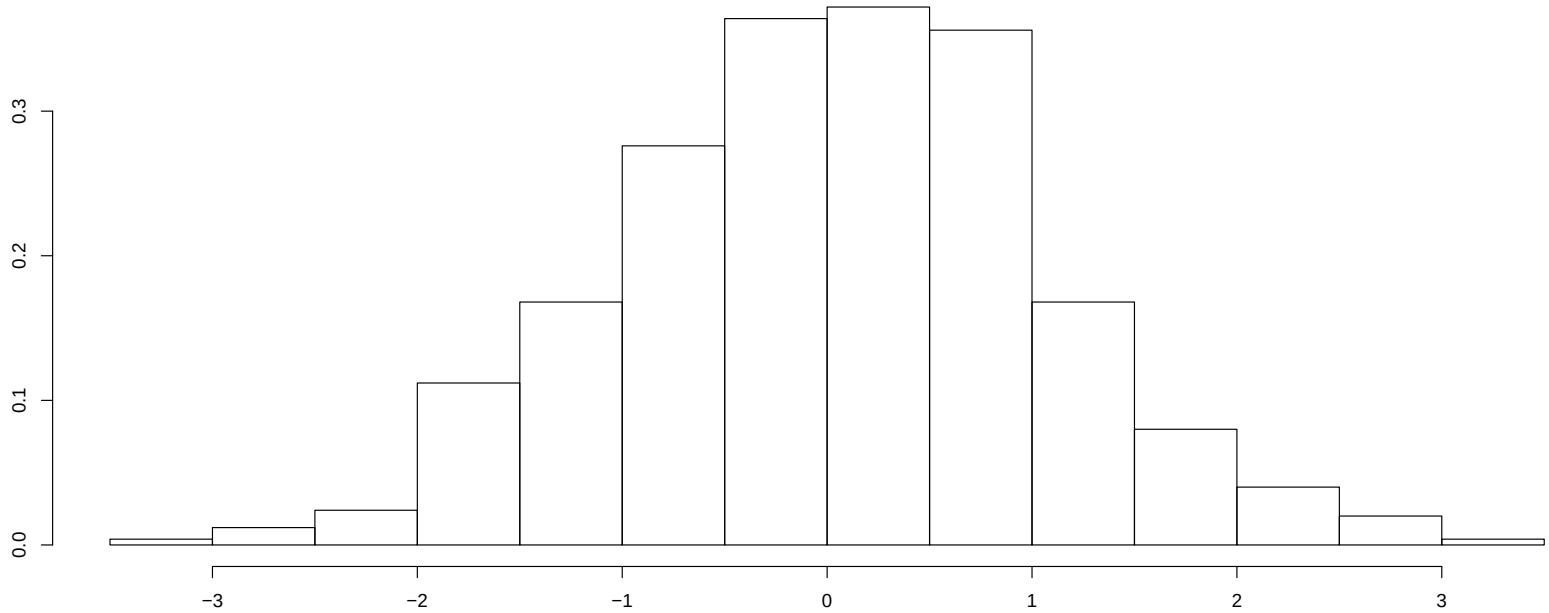
Multinomialverteilung



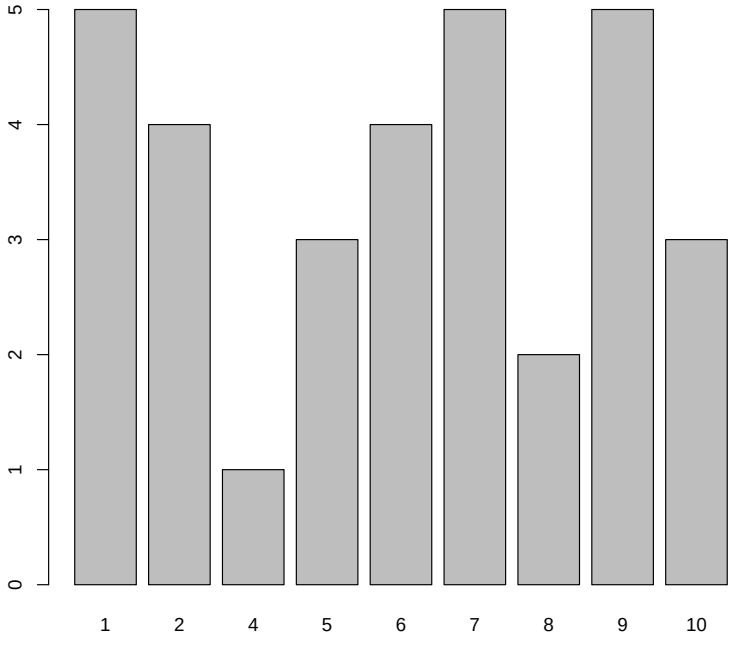
Gleichverteilung

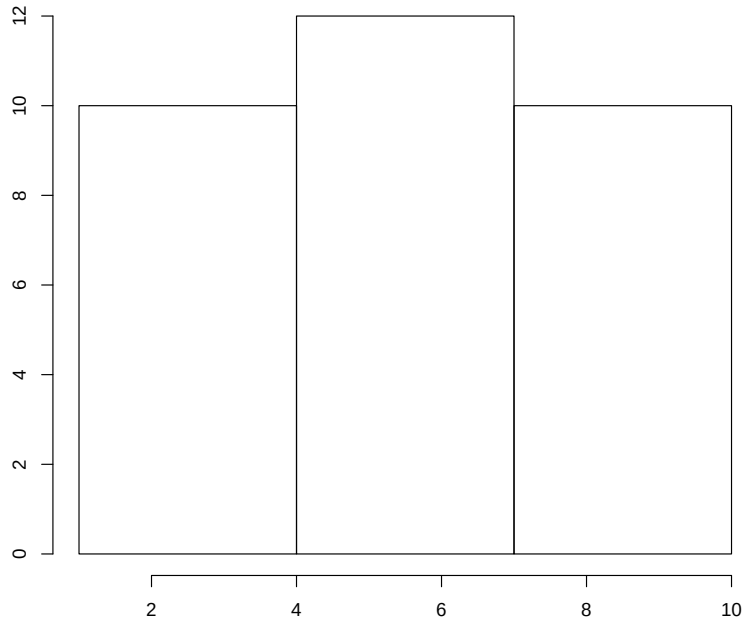


Normalverteilung

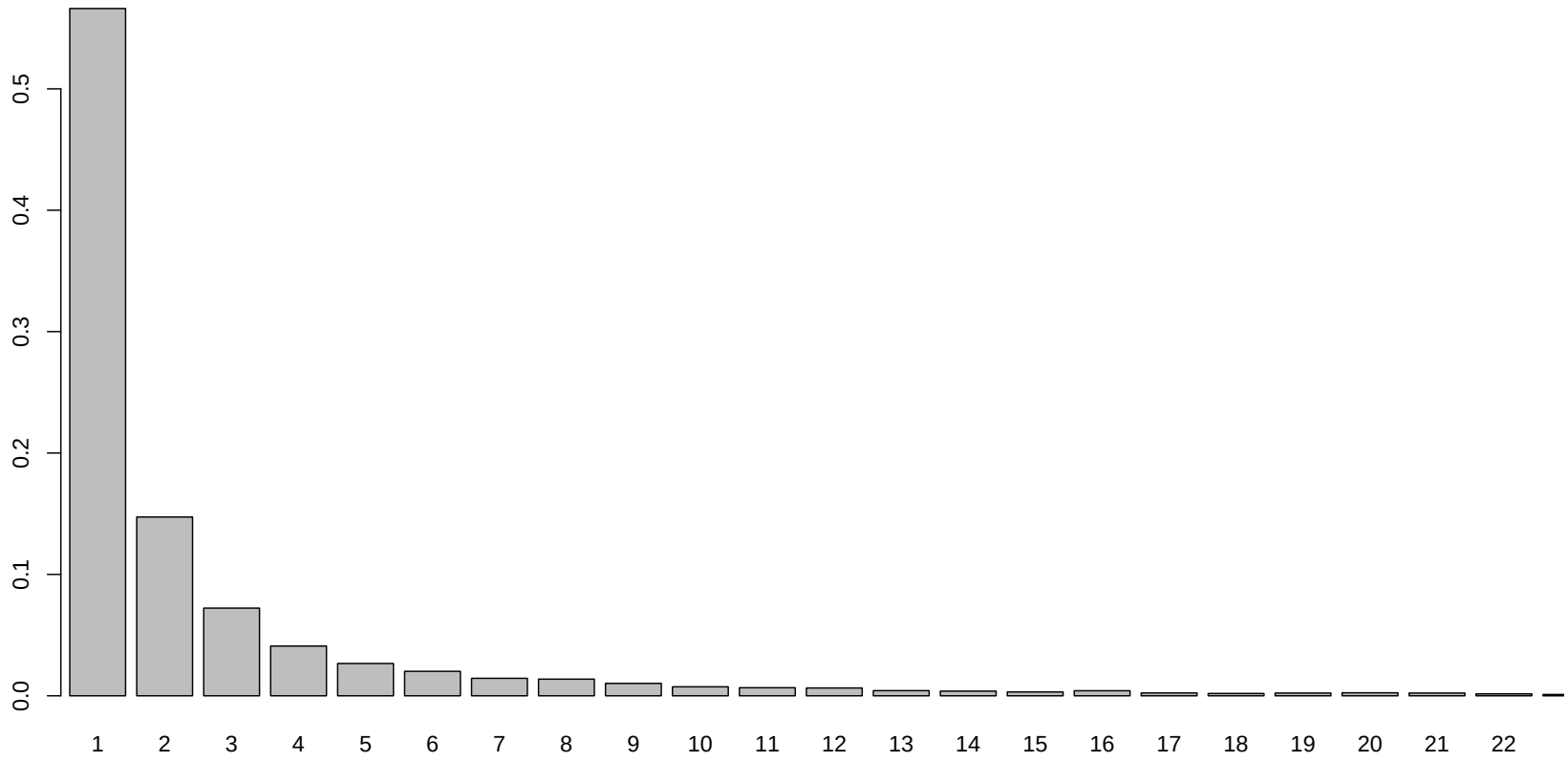


Balkendiagramm und Histogramm





LNRE-Verteilung



Belege von *Gott* bei Goethe (I)

Ausgabe von Cosmas und Koka:

Nr.	Sigle	Quelle	Beleg
1	GOE/-AGA.-00000	Goethe: Campagne in Frankreich, [Autobiographie], (Geschr. 1820-1822), In: Goethes Werke, Bd. 10. - München, 1982 [S. 251]	... immer [[[Gott]]] danken ...

Belege von *Gott* bei Goethe (II)

Wünschenswerte Informationen:

- Wortform: in Beleg, markiert durch *[[[...]]]*
- Textname: in Quelle, zwischen *Goethe:␣* und *„␣[*
- Textsorte: in Quelle, nach dem Textnamen, markiert durch *[...]*
- Jahresangabe: in Quelle, zwischen *„␣(* und *),*

Belege von *Gott* bei Goethe (III)

Aufbereitung mit dem Programm **grep**:

```
grep -Pon '(?<=Goethe: ).+?(?=,)' gott.csv
```

Erklärung:

(?<=Goethe:) *Goethe:*␣ muss vor dem gesuchten Text stehen

(?=,) , muss hinter dem gesuchten Text stehen

. beliebiges Zeichen

+ mindestens einmal, sonst beliebig oft

? so selten wie möglich

Der Vokal bei *Glaube* im Frnhd. (I)

Daten liegen in XML-Dateien vor

```
<text>
...
<wortform lemma=”glaube” vokal=”au” ...>
... gelaub ...
</wortform>
...
</text>
```

Pfadangabe	//wortform/@lemma
Pfadangabe mit Auswahl	//wortform[@lemma=’glaube’]

Der Vokal bei *Glaube* im Frnhd. (II)

Umwandeln mit **XSLT**:

- jede Box durchgehen
- eine Anweisung ausführen
  (z.B. den Text der Box ausgeben)
- Anweisungen für die Boxen können über Pfadangaben ausgewählt werden

Literatur

- zu **grep**
  Anhang: Unix-Befehle In: Bubenhofer, 2006–2011
- zu **XML** und **XSLT**
  3.5 Von Rohdaten zum Korpus In: Perkuhn, Keibel und Kupietz, 2012

Daten

- Gruppe 1
  - Jahreszahlen aus der FAZ, Jahrgang 1994
    (aus Baayen, 2008, verändert)
    `faz.csv`
  - Glaube im Frnhd.
    `glaube.csv`
- Gruppe 2
  - Füllwörter (Variable: Filler)
    (aus Gries, 2008, S. 103)
    `filler.csv`
  - Gott bei Goethe
    `gott.csv`

Gruppenarbeit

- Fassen Sie die Daten Ihrer Gruppe möglichst knapp, aber aussagekräftig zusammen.
- Nutzen Sie dazu **Verteilungen**, **graphische Darstellungen** und **Kennzahlen**.
- Präsentieren Sie am Ende der anderen Gruppe Ihre Daten.

Wichtige R-Befehle

Häufigkeitsverteilung	`table(x)`
rel. Häufigkeitsverteilung	`table(x)/length(x)`
Modalwert	`which.max(table(x))`
Median	`median(x)`
arithm. Mittel	`mean(x)`
Quartile	`quantile(x,type=1)[c(2,4)]`
Interquartilsabstand	`diff(Quartile von x)`
Spannweite	`diff(range(x))`
Standardabweichung	`sd(x)`
Balkendiagramm	`barplot(table(x))`
Histogramm	`hist(x)`
Boxplot	`boxplot(x,outline=FALSE)`