Korpuslinguistik und Statistik

(050370)

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

11. Sitzung

Untersuchungsdesigns

univariat: eine Variable, deren Verhalten beschrieben werden soll

bivariat: unabhängige Variable erklärt abhängige Variable

multivariat: unabhängige Variablen erklären abhängige Variable

Gemeinsame Verteilung zweier Variablen

Häufigkeiten, mit der die Ausprägungen der beiden Variablen gemeinsam auftreten

Darstellung: Kontingenztafel

	V 1, Ausprägung 1	V 1, Ausprägung 2	...
V 2, Ausprägung 1	Hfgkt. Ausprägungen 1,1	Hfgkt. Ausprägungen 2,1	
V 2, Ausprägung 2	Hfgkt. Ausprägungen 1,2	Hfgkt. Ausprägungen 2,2	
...			...

Beispiel – Urliste

(Ausschnitt aus dem Datensatz `etymology` zu Baayen, 2008)

	Verb	WrittenFrequency	Auxiliary	Regularity
1	blijken	9.88	zijn	irregular
2	gloeien	6.91	hebben	regular
3	glimmen	7.03	zijnheb	irregular
4	rijzen	7.37	zijn	irregular
5	werpen	8.49	hebben	irregular
6	delven	6.17	hebben	irregular

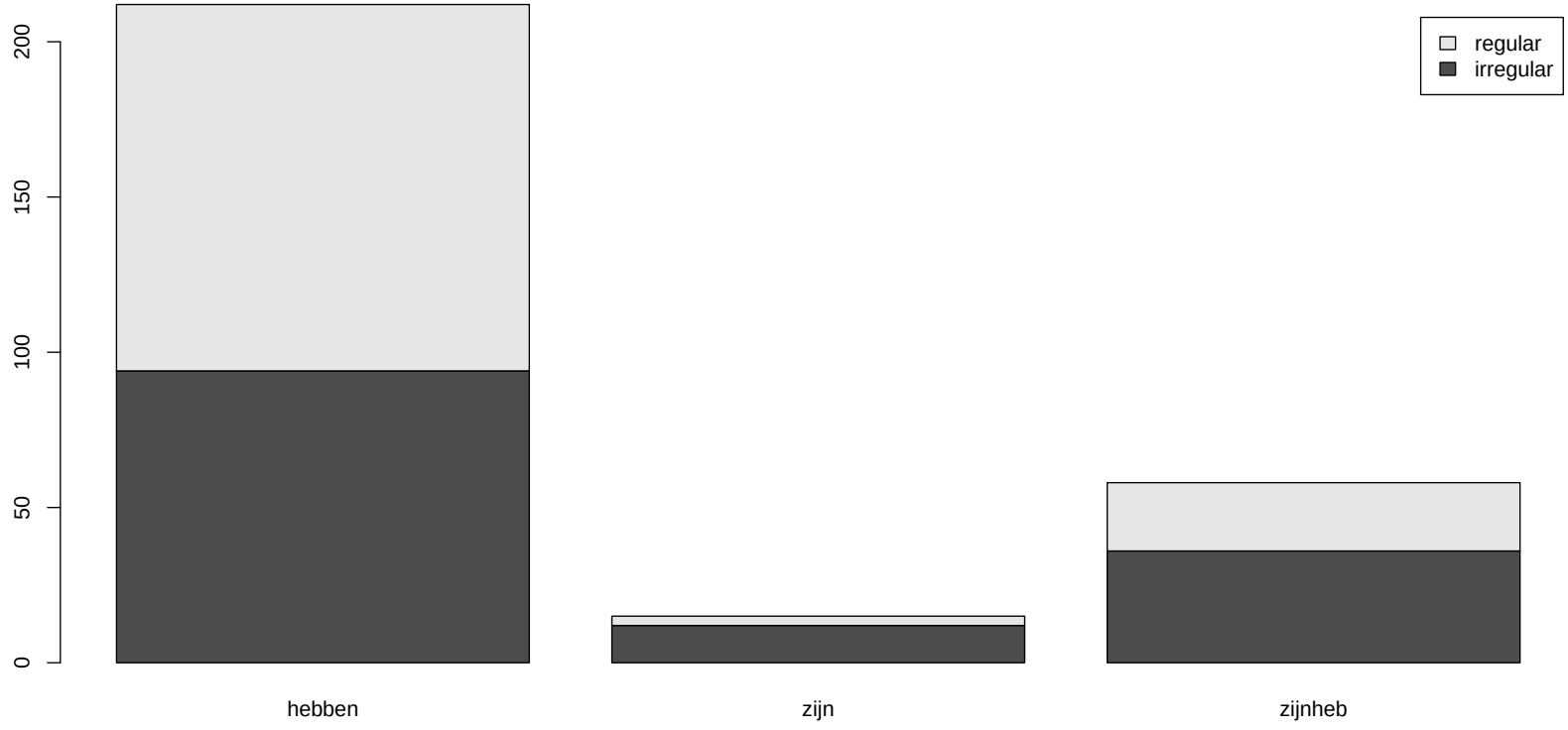
Beispiel – Kontingenztafel

Variablen: *Regularity* und *Auxiliary*

	irregular	regular
hebben	94	118
zijn	12	3
zijnheb	36	22

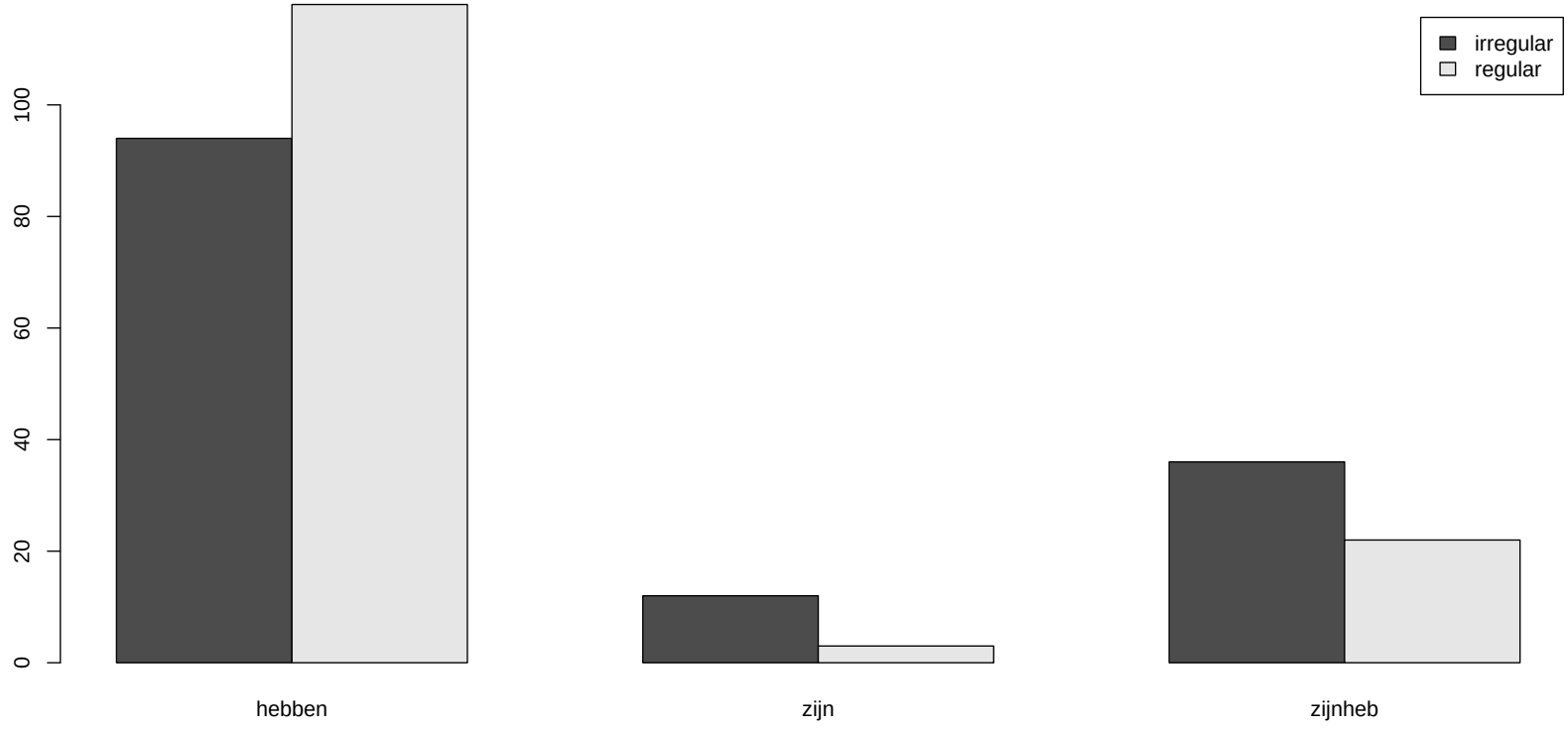
Graphische Darstellung – Balkendiagramm (I)

unabhängige Variable: kategorial, ordinal
abhängige Variable: kategorial, ordinal



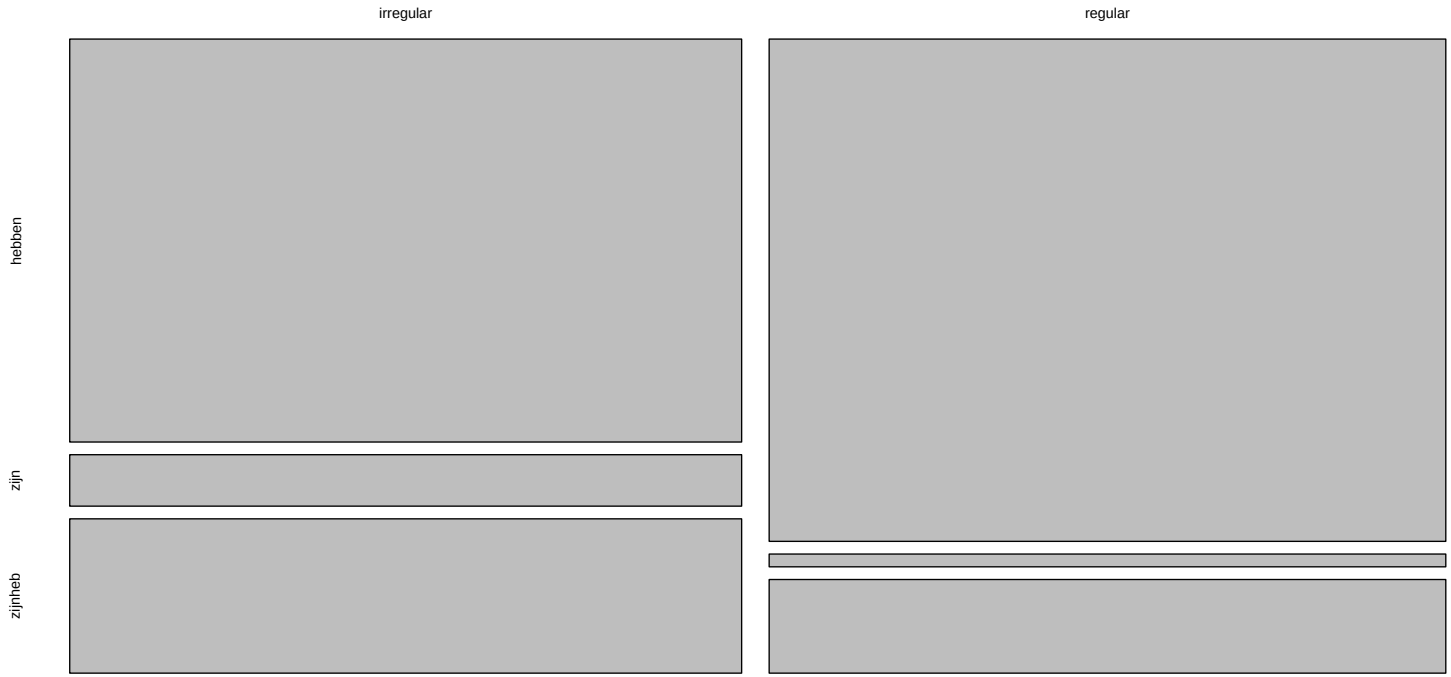
Graphische Darstellung – Balkendiagramm (II)

unabhängige Variable: kategorial, ordinal
abhängige Variable: kategorial, ordinal



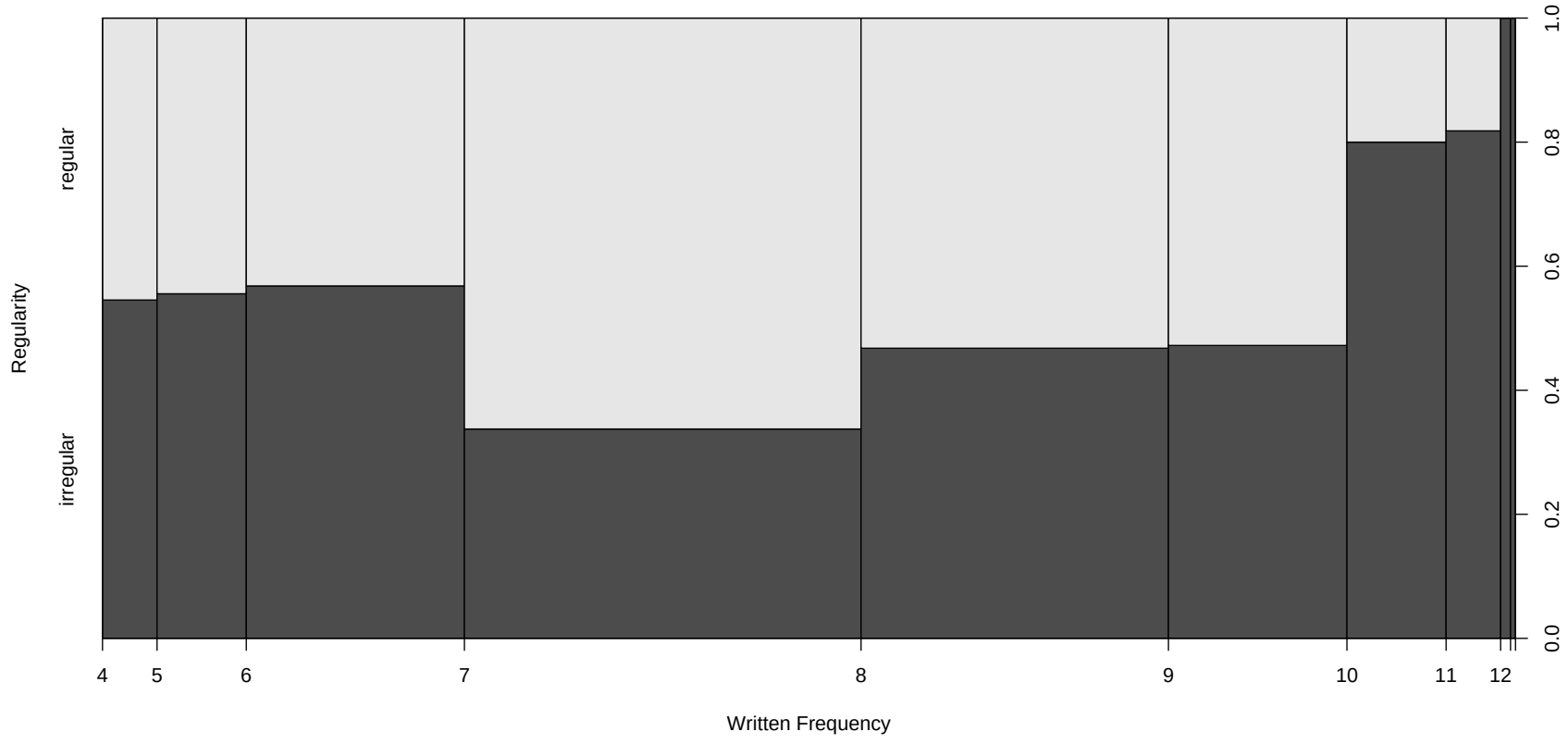
Graphische Darstellung – Mosaikplot (III)

unabhängige Variable: kategorial, ordinal
abhängige Variable: kategorial, ordinal



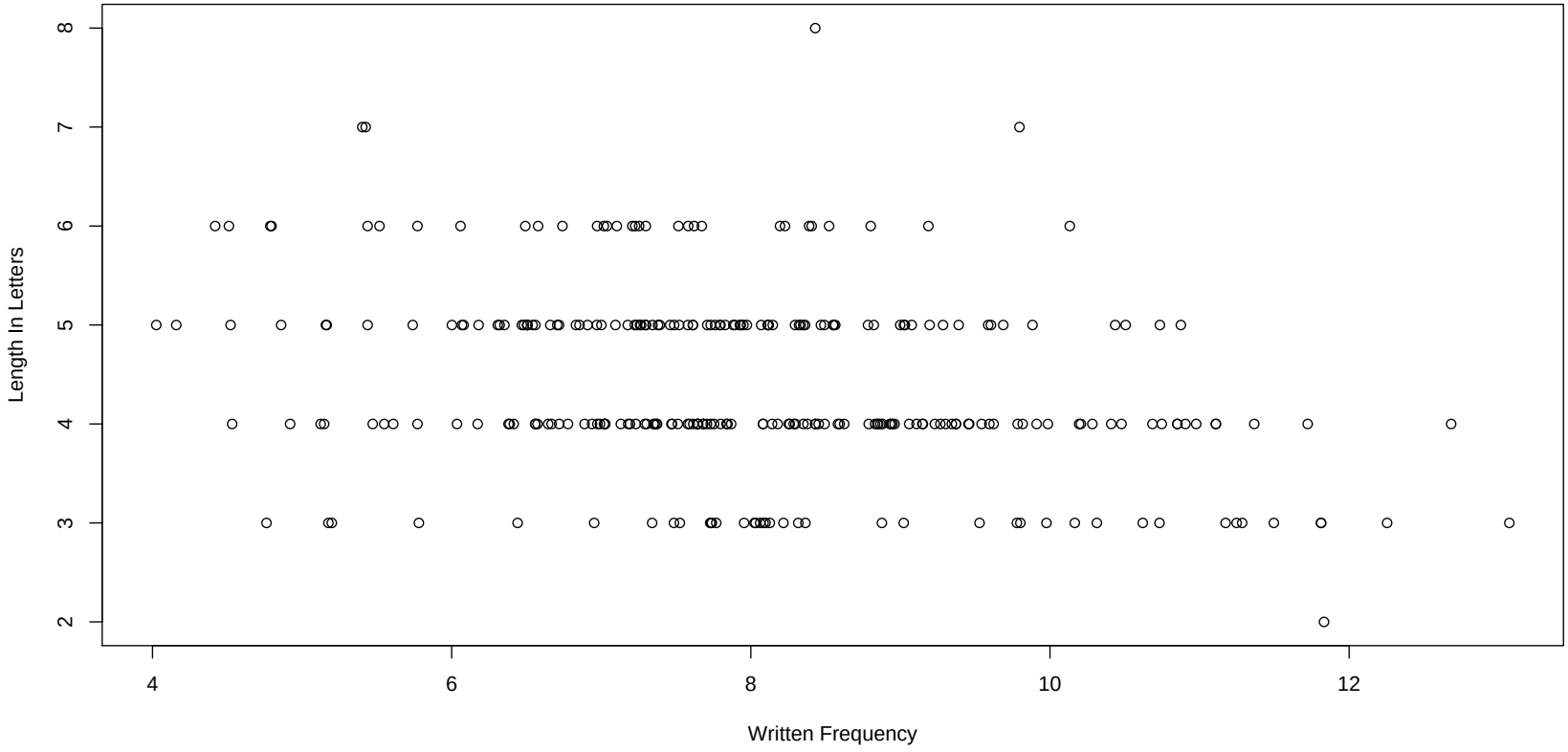
Graphische Darstellung – Spineplot (IV)

unabhängige Variable: (ordinal), metrisch
abhängige Variable: kategorial, ordinal



Graphische Darstellung – Scatterplot (V)

unabhängige Variable: metrisch
abhängige Variable: metrisch



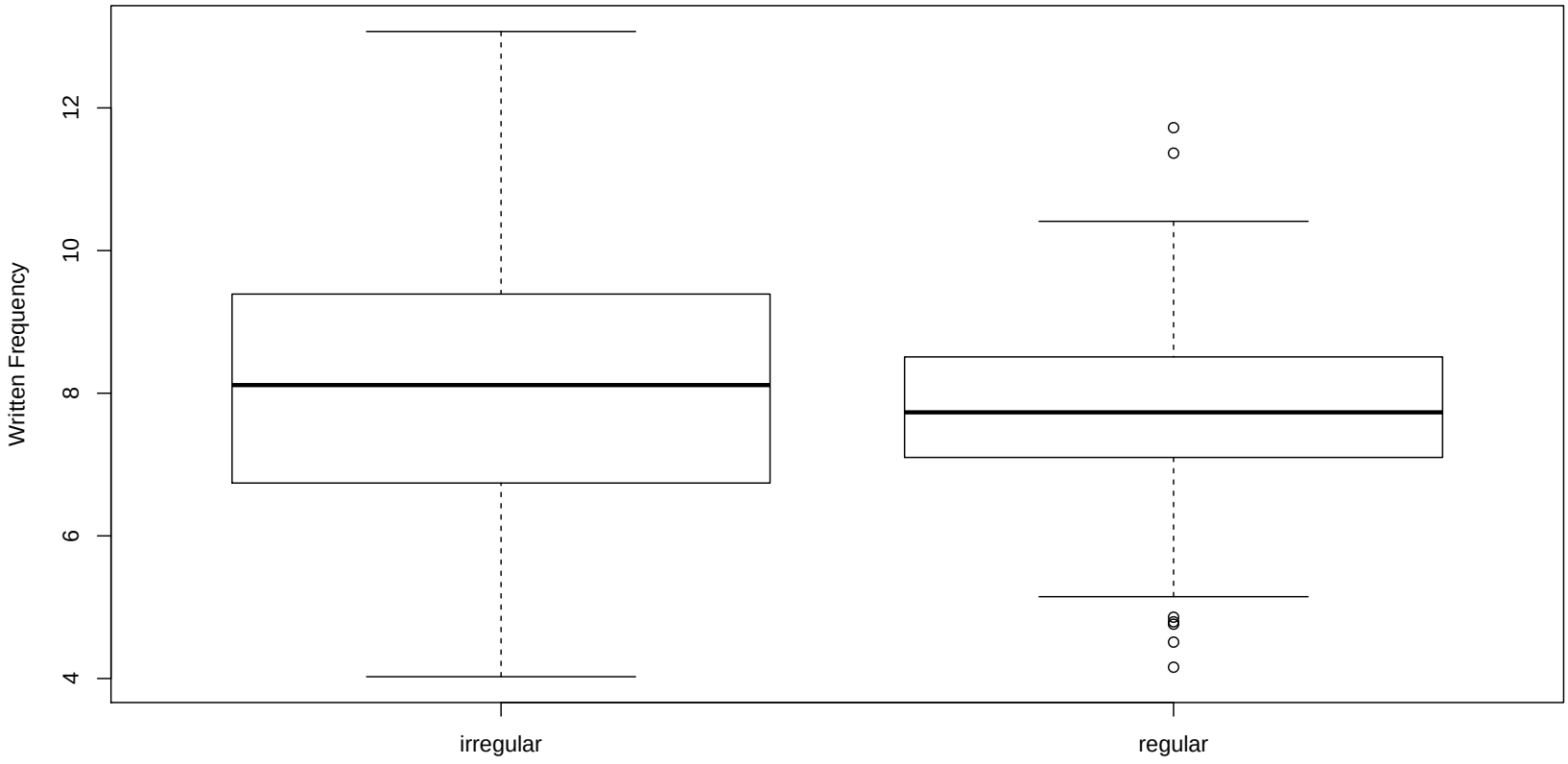
Kennzahlen einzelner Gruppen

unabhängige Variable: nominal, ordinal

Written Frequency:

	irregular	regular
Min.	4.03	4.16
1st Qu.	6.75	7.10
Median	8.11	7.73
Mean	8.19	7.79
3rd Qu.	9.39	8.51
Max.	13.10	11.70

Boxplot



Zusammenhangsmaß für Kontingenztafeln

Was heißt Zusammenhang?
die Ausprägung der unabhängigen Variable bestimmt die Ausprägung der abhängigen Variablen

oder umgekehrt:
wenn die Ausprägung der unabhängigen Variable keinen Einfluß auf die abhängige Variable hat, existiert kein Zusammenhang

Beispiel

beobachtet

	irregular	regular	Sum
hebben	94.00	118.00	212.00
zijn	12.00	3.00	15.00
zijnheb	36.00	22.00	58.00
Sum	142.00	143.00	285.00

erwartet

(wenn unabhängig)

	irregular	regular	Sum
hebben	105.63	106.37	212.00
zijn	7.47	7.53	15.00
zijnheb	28.90	29.10	58.00
Sum	142.00	143.00	285.00

χ^2 (Chi-Quadrat)-Koeffizient

erwartete Häufigkeit in einem Feld:

$$\frac{\text{Summe Zeile} * \text{Summe Spalte}}{\text{Anzahl Beobachtungen}}$$

Abweichung in einem Feld:

$$(\text{beobachtet} - \text{erwartet})^2$$

Gewichtete Abweichung:

$$\frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}}$$

Gesamtabweichung (Summe der Felder):

$$\boxed{\chi^2 = \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}}}$$

χ^2-Koeffizient als Zusammenhangsmaß?

Problem
maximale Größe von χ^2 (χ^2_{max}) ist:
Anzahl Beobachtungen $*$ (min(Anzahl Zeilen, Anzahl Spalten) $-$ 1)

also abhängig von:

- der Stichprobengröße
- Anzahl der Ausprägungen

Cramers V

daher: Normierung auf das Intervall $[0, 1]$

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}}$$

0 heißt: unabhängig

1 heißt: vollständig abhängig
(kenne ich die Ausprägung der unabhängigen Variable,
kenne ich die Ausprägung der abhängigen Variable)

Beispiel – vollständig abhängig

	irregular	regular	Sum
hebben	212.00	0.00	212.00
zijn	0.00	15.00	15.00
zijnheb	0.00	58.00	58.00
Sum	212.00	73.00	285.00

(geht nicht immer ohne die Randhäufigkeiten zu verändern)

$\chi^2 = 285$

Beispiel – Cramers V

	irregular	regular	Sum
hebben	94.00	118.00	212.00
zijn	12.00	3.00	15.00
zijnheb	36.00	22.00	58.00
Sum	142.00	143.00	285.00

$V = 0.2008$

geringe Abhängigkeit

Zusammenhang von metrischen Daten

	WrittenFrequency	LengthInLetters
1	9.88	5
2	6.91	5
3	7.03	4
4	7.37	4
5	8.49	4
6	6.17	4

Korrelationskoeffizient

gemeinsame Abweichung der einzelnen Werte:
$(x - \text{Mittelwert}(x)) * (y - \text{Mittelwert}(y))$

(Stichproben)Kovarianz
Mittelwert der einzelnen gemeinsamen Abweichungen:
$\text{Cov} = \frac{1}{n} \sum (x - \text{Mittelwert}(x)) * (y - \text{Mittelwert}(y))$

Problem: abhängig von der Größe der Ausprägungen
daher: Normierung (nach Pearson)

$$\boxed{\text{Cor} = \frac{\textit{Kovarianz}}{\textit{Standardabweichung}(x) * \textit{Standardabweichung}(y)}}$$

Zur Interpretation der Korrelationskoeffizienten

Korrelations-koeffizient	Bezeichnung der Korrelation	Art der Korrelation
$0,7 < r \leq 1$	sehr hoch	positive Korrelation:
$0,5 < r \leq 0,7$	hoch	je mehr/höher ..., desto mehr/höher ...
$0,2 < r \leq 0,5$	mittel	je weniger/niedriger ..., desto weniger /niedriger ...
$0 < r \leq 0,2$	gering	
$r = 0$	Nullkorrelation: kein statistischer Zusammenhang	
$0 > r \geq -0,2$	gering	negative Korrelation:
$-0,2 > r \geq -0,5$	mittel	je mehr/höher ..., desto weniger/niedrieger ...
$-0,5 > r \geq -0,7$	hoch	je weniger/niedrieger ..., desto mehr/höher ...
$-0,7 > r \geq -1$	sehr hoch	

(Gries, 2008, S. 144)

Beispiel

Korrelation zwischen *Written Frequency* und *Length In Letters*:
-0.3257

negative, mittlere Korrelation
Je häufiger ein Wort, desto kürzer ist es.

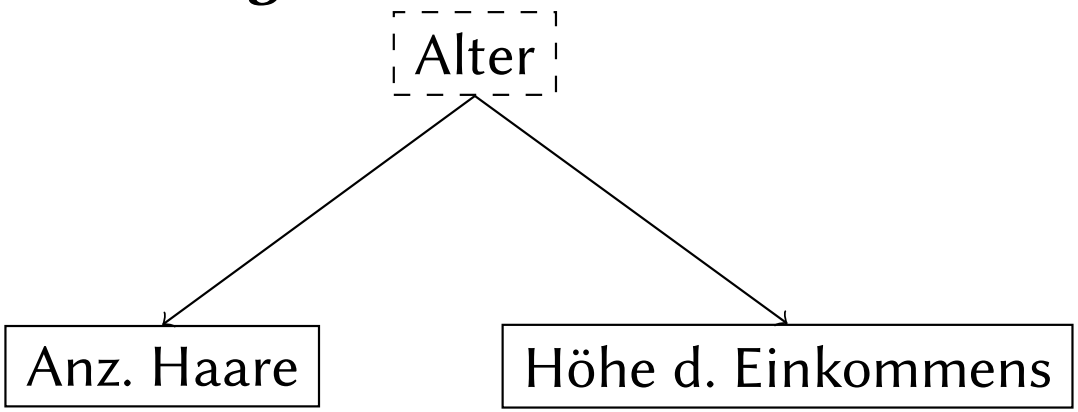
Zusammenfassung: Zusammenhangsmaße

	Nominal	Ordinal	Metrisch
Cramers V	x	x	(x)
Korrelationskoeffizient			x

Ursache und Wirkung?

Das Einkommen von Männern korreliert in Deutschland mit der Anzahl der Haare, die sie auf dem Kopf haben.

Erklärung:



Eine Kausalitätsbeziehung wird theoretisch begründet!

Korrelation kann dann als Beleg für diese Kausalität dienen.

Encodings in R

Option: `fileEncoding` in der Funktion `read.table` setzen

```
read.table("...", fileEncoding="ENCODING")
```

wichtige Encodings: utf8, latin1

Optionen für Graphiken

`main`	Überschrift
`xlab`	Beschriftung der x-Achse
`ylab`	Beschriftung der y-Achse
`legend`	Legende Angabe: `c("Element1", "Element2", ...)`

Cramers V

existiert nicht als Befehl

Berechnung: `sqrt(chisq.test(tabelle)$statistic / (sum(tabelle) * (min(dim(tabelle))-1)))`

als Befehl hinzufügen:

`cramers.V <- function(tabelle) {` Berechnung `}`

Wichtige R-Befehle (II)

Daten einlesen	`read.table("...")`
Daten entfernen	`rm(x)`
Kontingenztafel	`table(y,x)`
Balkendiagramm	`barplot(table(y,x))`
Mosaikplot	`mosaicplot(table(y,x))`
Spineplot	`spineplot(y,x)`
Scatterplot	`plot(y,x)`
Boxplot	`boxplot(y ~ x)`
gruppierte Kennzahlen	`tapply(y,x, fkt)`
Cramers V	`(cramers.V(table(y,x)))`
Korrelationskoeffizient	`cor(y,x)`