Analysen politikwissenschaftlicher Datensätze mit Stata

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Sitzung 4: Deskriptive Statistik

Vorbereitung

- bitte starten Sie Stata (z:\profile.do)
- laden Sie anschließend den Datensatz z:\daten\allbus1980-2000.dta

Häufigkeitstabellen

- Einfachste Form der Datenauswertung
- Fragestellung: Wie häufig kommen die Ausprägungen einer einzigen kategorialen Variablen in der Stichprobe oder in der Population vor?

Häufigkeitstabellen

- Absolute Häufigkeiten, Prozente und kumulierte Prozente: `tab v378`
- Missing als Kategorie: `tab v378, mis`
- Einfache „graphische“ Darstellung: `tab v378, mis plot`

Kreuztabellen

- Kombinieren zwei kategoriale Merkmale
- Kirchgang × Region: `tab v378 v5` (Zeile × Spalte)
  - Spaltenprozente: `tab v378 v5,col`
  - Zeilenprozente: `tab v378 v5,row`
  - absolute Häufigkeiten unterdrücken: `tab v378 v5, col nofre`
  - Totalprozente: `tab v378 v5, cel nofre`

Kreuztabellen

- `tab2 varlist` erzeugt alle möglichen Kreuztabellen ohne Doubletten: `tab2 v378 v5 v6,col`
- `tab1 varlist` erzeugt einfache Tabellen (Häufigkeitstabellen) für alle angegebenen Variablen

Grafiken

- Kohler/Kreuter beziehen sich in ihrem Buch auf Stata 7
- Die Syntax der Grafik-Befehle hat sich Stata 8 grundlegend geändert
- Nach Eingabe von `version 7` könnten Sie die alten Befehle wieder benutzen (nicht empfehlenswert)

Grafiken

- Die Grafik-Befehle sind sehr mächtig, aber auch hochgradig komplex: `whelp graph`
- Diese Komplexität entsteht hauptsächlich durch Optionen und Unteroptionen
- Viele wichtige Grafiken erzeugen Sie mit Varianten von `graph twoway`
- Grafikbefehle können abgekürzt werden
  - `graph twoway scatter`
  - `twoway scatter` und
  - `scatter` sind äquivalent
- Grafiken können gruppenweise erzeugt und überlagert werden

Balkendiagramme

- Für nominale Variablen (z.B. Konfession)
- `graph bar v378` ergibt leider nicht das gewünschte Ergebnis
- Sie müssen zunächst für jede Kirchgangskategorie einen Dummy erzeugen: `tab v378,gen(kg)`
- `d kg1-kg6`
- `graph bar kg1-kg6` ist schon besser

Balkendiagramme

- Vermutlich ist das das nahe am Gewünschten: `graph bar kg1-kg6,percent bargap(25) legend(lab(1 ">1/Woche") lab(2 "1/Woche") lab(3 "1-3/Monat") lab(4 "mehrmals/Jahr") lab(5 "seltener") lab(6 "nie")) ytitle("Prozent")`

Balkendiagramme

- Gruppenweise Ausführung
- Holen Sie mit Bild↑ den vorherigen Befehl zurück
- Fügen Sie ganz am Ende `, by(v5)` hinzu

Streifendiagramme

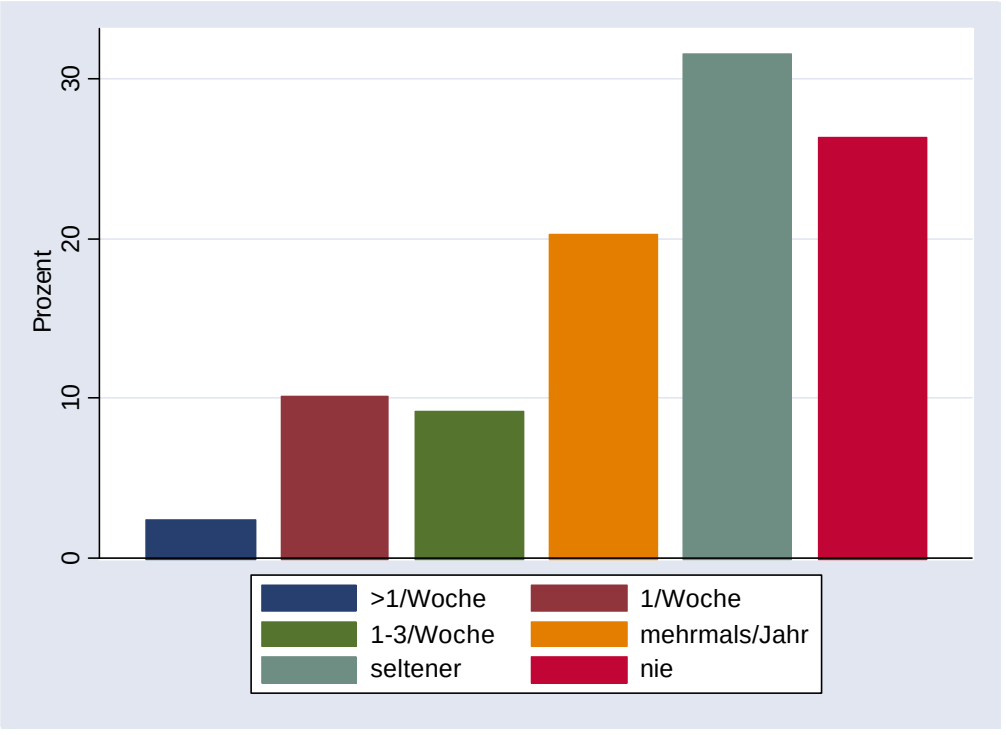
- Holen Sie mit Bild↑ den letzten oder vorletzten Befehl zurück
- Ersetzen Sie `graph bar` durch `graph hbar`

Schemata

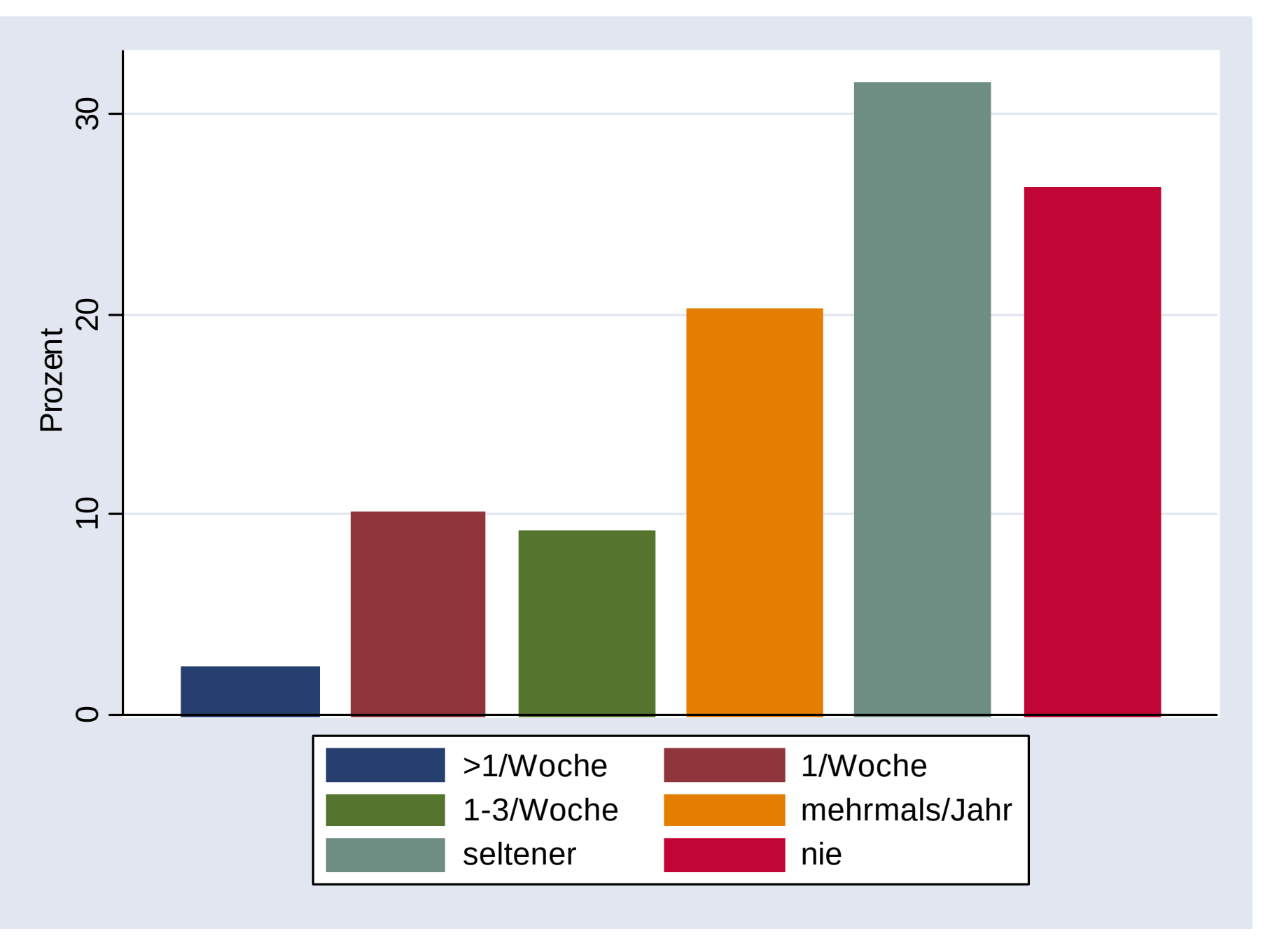
- Für Haus-/Magisterarbeiten: besser Graustufen-Darstellung mit weißem Hintergrund
- `set scheme s1mono`
- Letzten Grafik-Befehl mit Bild↑ zurückholen und ausführen
- `set scheme s2color` kehrt zur Bildschirmdarstellung zurück
- Mit entsprechenden Suboptionen können Sie außerdem Farben, Linien- und Füllmuster verändern

Import/Export

- Bei mehreren Grafiken: `graph export graf1.emf,replace` speichert die aktuelle Grafik im EMF-Format und ersetzt eine evtl. vorhande Version
- quick & dirty: mit der rechten Maustaste auf die Grafik klicken – kopieren – in Office einfügen



EMF (Enhanced Meta File) – Dateien sind frei skalierbar


Prozent
0
10
20
30
>1/Woche
1/Woche
1-3/Woche
mehrmals/Jahr
seltener
nie

Intervallskalierte Daten

- Darstellung als Histogramm
- `graph twoway histogram v372`
- Fläche proportional zur Häufigkeit
- Die Klassenbreite ist bei Stata konstant
- Die Zahl der Klassen können Sie mit ,bin(#) selbst festlegen
- `graph twoway histogram v372,bin(10)`

Kern-Dichte-Schätzer

- Histogramm faßt notwendigerweise kontinuierliche Daten zu Gruppen zusammen
- Kern-Dichte-Schätzer versuchen, Verteilung kontinuierlich zu schätzen – interessant insbesondere für Stichproben
- Stellen eine Art gleitendes Mittel dar
- Gewichtung der Fälle hängt von gewähltem „Kern“ ab
- Arbeiten die Form einer Verteilung heraus
- `graph twoway kdensity v372`
- Beide Graphen können kombiniert werden: `graph twoway (histogram v372,bin(10)) (kdensity v372)`

Liniendiagramme

- Zeitreihen sind Verteilungen, die besonders gut durch Linienzüge repräsentiert werden können
- Tippen Sie bitte `preserve`
- Laden Sie anschließend `z:\daten\pi-gesamt-77-01.dta`
- `graph twoway line piwest zeitpunkt`
- Daten entweder vorher sortieren oder `,sort` angeben

Liniendiagramme

- Sie können mehrere Zeitreihen in einer Grafik darstellen: `graph twoway line piwest piost zeitp`
- Der Variablen zeitpunkt ist ein besonderes Format zugewiesen, daß Sie als Datum (in Monaten seit Januar 1960) kennzeichnet
  - `d zeitp`
  - `list in 1/10`
  - `format zeitp %9.0g`
  - `list in 1/10`

Mittelwerte und Streuungsmaße

- `restore`
- Versuchen, die wesentlichen Eigenschaften einer Verteilung numerisch zu erfassen
- Mittelwerte
  - Modus
  - Median
  - Arithmetisches Mittel
- Streuungsmaße
  - Spannweite
  - Varianz
  - Standardabweichung

Mittelwerte und Streuungsmaße

- Alter errechnen `gen alter=v2-v372`
- `graph twoway kdensity alter`
- Arithmetisches Mittel, Varianz, Standardabweichung, Median und Perzentile: `summ alter,det`
- Alternativ z.B. `tabstat alter,stat(range median mean sd var)`
- Modus ist etwas komplizierter:
  - `egen dummy=mode(alter)`
  - `summ dummy`
  - `whelp egen`

Zusammenhang

- Zusammenhänge:
  - Arbeiter wählen häufiger die SPD als andere Gruppen
  - Hochgebildete vertreten häufiger postmaterialistische Werte als Niedriggebildete
  - Männer haben ein höheres Durchschnittsgehalt als Frauen
  - Je älter ein Befragter ist, desto höher ist auch sein Wert auf einer Konservatismusskala
- Ein Zusammenhang zwischen zwei Variablen besteht dann, wenn bestimmte Ausprägungen *häufiger gemeinsam* auftreten, als bei einer *zufälligen* Verteilung zu erwarten wäre

Zusammenhangsmaße

- beschreiben einen Zusammenhang zwischen zwei Variablen
- Mit Hilfe von Zusammenhangsmaßen kann die Stärke verschiedener Zusammenhänge leichter miteinander verglichen werden
- Zusammenhangsmaße sollten einen Wertebereich von 0 bis 1 bzw. von -1 bis +1 aufweisen
- Wahl des Zusammenhangsmaßes hängt vom Skalenniveau der Variablen ab

Maße auf der Basis von χ^2

- Zwei nominale Variablen
- vergleichen eine empirische Kreuztabelle mit einer Tabelle, in der die Häufigkeiten eingetragen sind, die zu erwarten wäre, wenn kein Zusammenhang zwischen den Merkmalen bestünde (Indifferenztabelle)
- Hier ist besonders leicht zu erkennen, daß Zusammenhänge sich auf das überzufällig häufige gemeinsame Auftreten von Ausprägungen beziehen

Konfession × Region

- Beobachtete Werte, Zeilenprozente, erwartete Werte: `tab v377 v5, row exp`
- Cramers V: `tab v377 v5, V`

Bildung × pol. Interesse

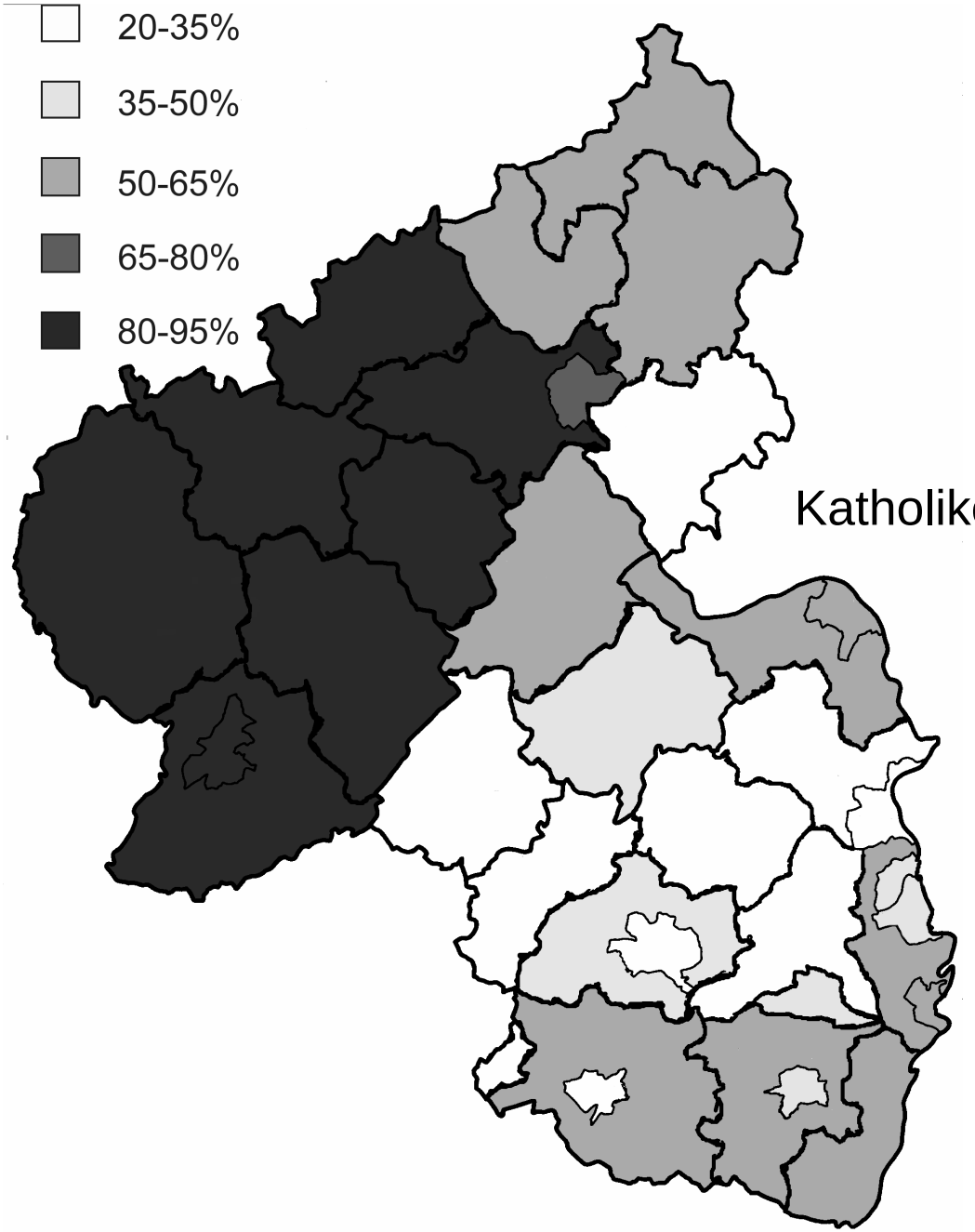
- zwei ordinale Variablen
- Bildung rekodieren: `recode v382 (1/2=1 niedrig) (3=2 mittel) (4/5=3 hoch) (else=.),gen(bildung)`
- `tab v20 bildung,col`
- Gamma basiert auf der Logik des Paarvergleichs
  - `tab v20 bildung,gamma`
  - Interesse „falsch“ kodiert:
  - `numlabel v20,add`
  - `tab v20`

Geschlecht × Einkommen

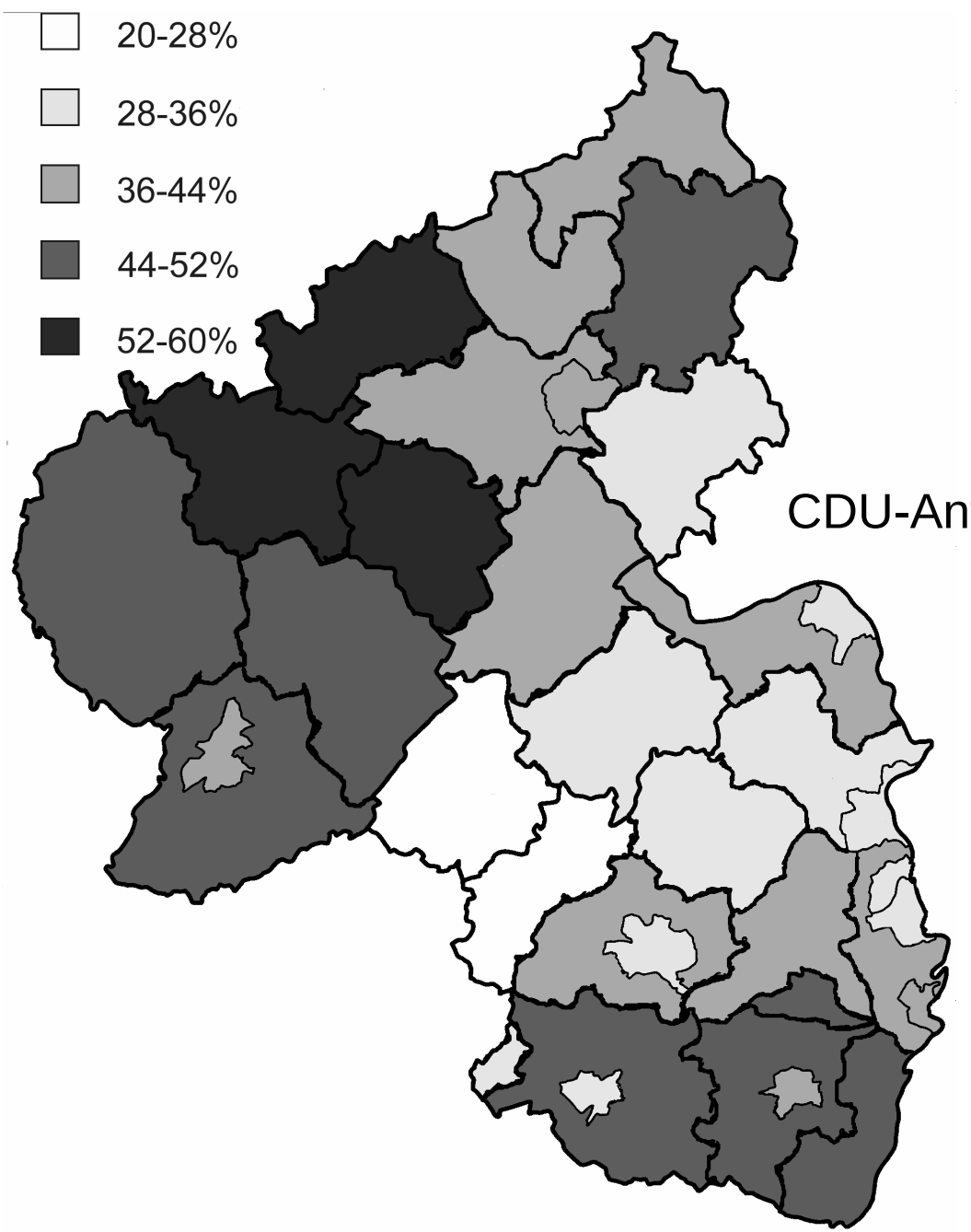
- Ein nominales, ein intervallskaliertes Merkmal
- η bzw. η^2
- Vergleicht Streuung innerhalb der Gruppen mit Gesamtstreuung
- Entspricht einfacher Varianzanalyse
  - `tabstat v495,by(v376)`
  - `loneway v495 v376`
  - R-squared entspricht η^2

Katholikenanteil × CDU-Anteil

- Zwei intervallskalierte Merkmale: Pearsons r
- `use z:\daten\rpstrukt,replace`
- `list kreis pwbcdu71 pkathv70`
- Wie hängen beide Merkmale zusammen?
  - `summ pwbcdu71 pkathv70`
  - `summ pwbcdu71 if pkathv70>54`
  - `summ pwbcdu71 if pkathv70<54`



Katholikenanteile, Volkszählung 1970



CDU-Anteile, LTW 1971

Katholikenanteil × CDU-Anteil

- Berechnung Pearsons R
  - Abweichungsprodukte
  - Kovarianz
  - Normieren
- `graph twoway scatter pwbcdu71 pkathv70,ylabel(,nogrid)`
- Mittelwerte eintragen: `graph twoway scatter pwbcdu71 pkathv70,xline(54.1) yline(39.2) ylabel(,nogrid)`

Katholikenanteil × CDU-Anteil

- Komplexere Darstellungen sind möglich und bei geringer Fallzahl sinnvoll
- Kreise mit besonders hohem/niedrigem Katholikenanteil + Ausreißer links oben markieren:
  - `graph twoway (scatter pwbcdu71 pkathv70,xline(54.1) yline(39.2) ylabel(,nogrid)) (scatter pwbcdu71 pkathv70 if pkathv70>80,mlabel(kreis) ) (scatter pwbcdu71 pkathv70 if pkathv70 <28,mlabel(kreis) legend(off)) (scatter pwbcdu71 pkathv70 if pkathv70 >40 & pkathv70 <50 &pwbcdu71>40,mlabel(kreis))`
  - do z:\rlpplot
- Pearsons r: `corr pwbcdu71 pkathv70`

Hausaufgabe

- Schreiben Sie unter Verwendung von muster.do eine Datei zusammenhang.do, die
  - den kumulierten ALLBUS-Datensatz z:\daten\allbus1980-2000.dta öffnet
  - ein Histogramm mit überlagerter Kernel-Density-Schätzung für das Alter der Interviewer erzeugt
  - Arithmetisches Mittel, Standardabweichung, Modus und Median des Alters der Interviewer bestimmt
  - getrennt für Ost- und Westdeutschland den Zusammenhang zwischen Konfession (katholisch / protestantisch / andere) und Wahlverhalten (CDU / SPD / andere) bestimmt
  - den Zusammenhang zwischen der formalen Bildung der Befragten und der Interviewer ermittelt („anderer Abschluß“ / „noch Schüler“ auf missing setzen)
  - feststellt, ob das Durchschnittsalter der Befragten mit dem Geschlecht des Interviewers variiert und ob eine Korrelation zwischen dem Alter von Interviewern und Befragten besteht (vgl. die Aufgaben bei Gehring/Weins)
- Schicken Sie die Lösung bis zum 16. Juni an do-files@politik.uni-mainz.de; verwenden Sie das bekannte Schema