

Einführung

Statistik II

Einführung

Überblick

Wiederholung

Statistiksoftware

Deskriptive Statistik mit Stata

Tabellen, Häufigkeiten, Anteilswerte

Graphische Darstellungen

Kategoriale Daten

Kontinuierliche Daten

Zusammenfassung

The image shows a vintage punch card with columns for 'KONTR', 'NR.', 'TEXT', and 'BETRAG'. The card contains numerical data and a 'SUMMIERKARTE' section. The data is organized into rows, with some rows containing multiple columns of numbers. The 'SUMMIERKARTE' section is a specific row of numbers. The card is framed by a dark border.

Worum geht es hier?

- ▶ Eine von zwei zentralen Vorlesungen im Aufbaumodul I
- ▶ Für BA-Kernfach
- ▶ Vertiefung und Erweiterung der Grundkenntnisse
- ▶ **Plus praktische Anwendung**
- ▶ Berufsqualifizierende Übung (am Computer) zur Vorlesung (verpflichtend)

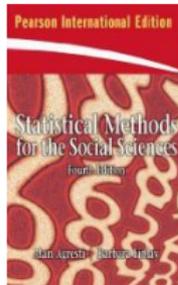
Worum geht es hier?

- ▶ Eine von zwei zentralen Vorlesungen im Aufbaumodul I
- ▶ Für BA-Kernfach
- ▶ Vertiefung und Erweiterung der Grundkenntnisse
- ▶ **Plus praktische Anwendung**
- ▶ Berufsqualifizierende Übung (am Computer) zur Vorlesung (verpflichtend)
- ▶ Weitere Bestandteile des Moduls: VL Methoden II + Seminar Berufsfeldqualifikation (z. B. Theorie der Umfrageforschung) + Übung „Anwendung von Forschungsmethoden“ (z. B. Auswertung von Mikro-Daten)

Worum geht es hier?

- ▶ Eine von zwei zentralen Vorlesungen im Aufbaumodul I
- ▶ Für BA-Kernfach
- ▶ Vertiefung und Erweiterung der Grundkenntnisse
- ▶ **Plus praktische Anwendung**
- ▶ Berufsqualifizierende Übung (am Computer) zur Vorlesung (verpflichtend)
- ▶ Weitere Bestandteile des Moduls: VL Methoden II + Seminar Berufsfeldqualifikation (z. B. Theorie der Umfrageforschung) + Übung „Anwendung von Forschungsmethoden“ (z. B. Auswertung von Mikro-Daten)
- ▶ **Gemeinsame Klausur für Statistik II/Methoden II**

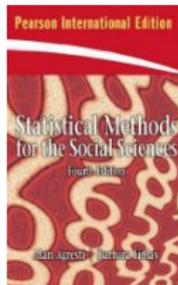
Literatur etc.



► Mehr Lehrbücher:

- Fox, John: Applied regression analysis, linear models, and related methods (2008)
- Gelman, Andrew & Hill, Jennifer: Data Analysis Using Regression and Multilevel/Hierarchical Models
- Kennedy, Peter: A Guide to Econometrics (5e)
- Wooldridge, Jefferey: Introductory Econometrics

Literatur etc.



- ▶ Mehr Lehrbücher:
 - ▶ Fox, John: Applied regression analysis, linear models, and related methods (2008)
 - ▶ Gelman, Andrew & Hill, Jennifer: Data Analysis Using Regression and Multilevel/Hierarchical Models
 - ▶ Kennedy, Peter: A Guide to Econometrics (5e)
 - ▶ Wooldridge, Jeffrey: Introductory Econometrics
- ▶ Homepage: <http://kai-arzheimer.com/Statistik-II/>

Nachschlagewerke Mathematik



- ▶ Fox, John: A mathematical primer for social statistics.
- ▶ Gill, J.: Essential Mathematics for Political and Social Research.



Nach-/Vorbereitung

- ▶ Weitere Literatur im ReaderPlus Nr. 7850
- ▶ Registrierunschlüssel STATA
- ▶ Für heute: Kohler/Kreuter Seite 7-28, <https://www.zdv.uni-mainz.de/readerplus/mav/1/miv/0/nph-ld/14545/31197.07803!1!1256220117!/document.pdf?Log=1>
- ▶ Außerdem: Kohler/Kreuter Seite 145-157, <https://www.zdv.uni-mainz.de/readerplus/mav/1/miv/0/nph-ld/14545/31197.07803!1!1256220610!/document.pdf?Log=1>
- ▶ Für nächste Woche: in Kohler/Kreuter stöbern

Was steht auf dem Programm?

Vorlesung	Inhalt
1	Wiederholung, Stata, Deskriptive Statistik
2	Mittelwerte, Zusammenhangsmaße, Hypothesentests in Stata
3	Was ist Regression?
4	Partielle Korrelation, Drittvariablenkontrolle
5	Signifikanztests, Gewichtung, Multikollinearität und Kohortenanalyse
6	ANOVA, Transformationen
7	Schätzverfahren, Annahmen, Verletzungen

Was steht auf dem Programm? II

Vorlesung	Inhalt
8+9	Einführung Logit und Probit
10	Multinomiale Logitmodelle
11	Einführung Count/Event Data
12	Einführung Zeitreihen
13	Einführung Panel/TSCS
14	Einführung Faktorenanalyse
15	Reserve/Wiederholung

Was sind Daten?

- ▶ Empirische Beobachtungen/Messungen an einem politischen Objekt (z.B. Person, Staat, Parteiprogramm, Parlament)
- ▶ Quantitativ
- ▶ Verschiedene „Skalenniveaus“
 1. Ratio-Skalen
 2. Intervall-Skalen
 3. Ordinal-Skalen
 4. Nominal-Skalen
- ▶ Speicherung auf Tontafeln/Papyrus/Papier (früher)
- ▶ Heute mit Computer (Lochkarten, Magnetbänder, Festplatten, CD/DVD ...)

Die Welt früher: Lochkarten

KUNDEN	A-I		TEXT	A-I		BETRAG	
	KA	J-R		J-R	DM	PF	
NR.							
I. C. T. GMBH DOSSELDORF, JAN-WELLEM-PLATZ 3	0	0	0	0	0	0	0
	1	1	1	1	1	1	1
	2	2	2	2	2	2	2
	3	3	3	3	3	3	3
	4	4	4	4	4	4	4
	5	5	5	5	5	5	5
	6	6	6	6	6	6	6
	7	7	7	7	7	7	7
	8	8	8	8	8	8	8
D-445	9	9	9	9	9	9	9
	1	2	3	4	5	6	7
	8	9	10	11	12	13	14
	15	16	17	18	19	20	21
	22	23	24	25	26	27	28
	29	30	31	32	33	34	35
	36	37	38	39	40		

SUMMENKARTE

BERÜHER PRESSE VERTIKAL

Was ist eine „(Roh)datenmatrix“?

- ▶ „Rohe“ (nicht-bearbeitete) Meßwerte → Tabelle
 - ▶ Speicherung intern und auf Festplatte/Speichermedium (Datei/Format)
1. Zeilen = „Fälle“ (Untersuchungsobjekt = Personen, Länder, Départements etc.)
 2. Spalten = „Variablen“ (Eigenschaft = Nationalität, Links-Rechts-Wert etc.)

	cntry	idno	lrscale	trstplt	trstplc	vote
1.	IT	3183200	6	2	7	yes
2.	SE	202429	6	3	6	yes
3.	AT	602	5	no trust	8	yes
4.	AT	1934	5	6	9	not elig
5.	IT	3583300	left	6	8	yes
6.	IT	3457000	4	7	9	yes
7.	AT	1216	4	4	7	yes
8.	IT	3120400	6	5	7	yes
9.	IT	3029500	left	5	complete	yes

Wie kommen die Daten in den Computer?

- ▶ Früher: pencil & paper →Eingabe von Hand (Fehler)
- ▶ Heute: (oft) CAPI/CATI, Web oder Datenbanken
→„maschinenlesbare“ Daten

Wie kommen die Daten in den Computer?

- ▶ Früher: pencil & paper →Eingabe von Hand (Fehler)
- ▶ Heute: (oft) CAPI/CATI, Web oder Datenbanken
→„maschinenlesbare“ Daten
- ▶ **Aber:** Vielzahl von Formaten (Arbeit + Fehler)
- ▶ Unzahl von verschiedenen *logischen* Formaten z. B. für Zeit und Datum (1. November 2009, 01/11/09, 11/01/2009, Nov 1 2009 ...)
- ▶ Unterschiedliche *physische Speicherkonventionen* (Unix vs. DOS/Windows, 32/64 bit, big-endian/little-endian)

Wie kommen die Daten in den Computer?

- ▶ Früher: pencil & paper →Eingabe von Hand (Fehler)
- ▶ Heute: (oft) CAPI/CATI, Web oder Datenbanken
→„maschinenlesbare“ Daten
- ▶ **Aber:** Vielzahl von Formaten (Arbeit + Fehler)
- ▶ Unzahl von verschiedenen *logischen* Formaten z. B. für Zeit und Datum (1. November 2009, 01/11/09, 11/01/2009, Nov 1 2009 ...)
- ▶ Unterschiedliche *physische Speicherkonventionen* (Unix vs. DOS/Windows, 32/64 bit, big-endian/little-endian)
- ▶ Vielzahl von Programmen mit (proprietären) Formaten (logischer Aufbau der Dateien im Computer)

Was sind die wichtigen maschinenlesbaren Formate?

1. Generische (Text) Formate
 - ▶ ASCII (frei, fest, komma-separiert)
 - ▶ Extension z. B. *.txt, *.dat, *.csv
 2. Proprietäre (binäre) Formate
 - ▶ Excel (*.xls)
 - ▶ SPSS (*.sav, *.por)
 - ▶ Stata (*.dta)
 - ▶ ...
- ▶ (Fast) alle Programme lesen Textformate
 - ▶ Probleme: Ineffizient (Dateigröße), Details
 - ▶ Proprietäre Formate erfordern Konversionsprogramme/-routinen
 - ▶ Probleme: Details, konzeptuelle Unterschiede (z. B. missing values)

Wozu braucht man ein Statistikprogramm?

1. Zugriff auf Daten
2. Verwaltung von Daten
 - ▶ Auffinden von Informationen
 - ▶ Kombinieren von Daten aus verschiedenen Quellen
 - ▶ *Kontrolliertes* Verändern/Kopieren von Rohdaten + Dokumentation → Skriptfähigkeit
3. *Reproduzierbare* Analyse von Daten → Skriptfähigkeit
4. Erstellen von Grafiken und Tabellen → Skriptfähigkeit wünschenswert

Welche wichtigen Programme gibt es?

1. Spezielle Programme für besondere Verfahren:
 - ▶ AMOS, Lisrel, MPlus, Mlwin
 - ▶ Interoperabilität, beschränkte Möglichkeiten zur Datenverwaltung
2. Generelle Programme für (fast) alle Aufgaben
 - ▶ Excel (nicht wirklich)
 - ▶ SAS (Dinosaurier, in manchen kommerziellen Bereichen)
 - ▶ SPSS (point & click, Marktforschung, noch an Universitäten)
 - ▶ R (open source, Programmiersprache, an Hochschulen in *fortgeschrittenen* Bereichen)
 - ▶ Stata (zunehmend in Wirtschafts- und Sozialwissenschaften)

Welche wichtigen Programme gibt es?

1. Spezielle Programme für besondere Verfahren:
 - ▶ AMOS, Lisrel, MPlus, Mlwin
 - ▶ Interoperabilität, beschränkte Möglichkeiten zur Datenverwaltung
 2. Generelle Programme für (fast) alle Aufgaben
 - ▶ Excel (nicht wirklich)
 - ▶ SAS (Dinosaurier, in manchen kommerziellen Bereichen)
 - ▶ SPSS (point & click, Marktforschung, noch an Universitäten)
 - ▶ R (open source, Programmiersprache, an Hochschulen in *fortgeschrittenen* Bereichen)
 - ▶ Stata (zunehmend in Wirtschafts- und Sozialwissenschaften)
- ▶ Vorteile von Stata
- ▶ Kann (fast) alles, was man bis zur Promotion braucht
 - ▶ Hervorragende Dokumentation, aktive Nutzergemeinschaft
 - ▶ Relativ leicht programmier- und erweiterbar
 - ▶ Wer Stata kann, lernt sehr schnell SPSS

Wie kann man Stata benutzen?

- ▶ Anmelden auf Windows-Terminalserver
ts.zdv.uni-mainz.de (aus PC-Pool oder von zu Hause,
Erläuterung: <http://www.zdv.uni-mainz.de/3222.php>)
- ▶ Stata aus Start-Menü auswählen
- ▶ Befehle per Menü/Eingabebox oder über Eingabezeile
(bevorzugt) eingeben
- ▶ Befehle und Variablennamen können abgekürzt werden
- ▶ Ausführen → Return-Taste; Befehle müssen nicht mit Punkt,
Komma oder Semikolon abgeschlossen werden
- ▶ Komplexere Befehle per Editor
bearbeiten/speichern/ausführen → *.do-Files
- ▶ Ein- und Ausgaben können protokolliert werden →
Reproduzierbarkeit und Automatisierung

Stata gestartet

Stata/IC 10.1 - Z:\home\kai\Work\15-Jahre-Deutsche-Einheit-Beck\Jugend\allbus2004.dta

File Edit Data Graphics Statistics User Window Help

Menüs und Buttons

bisherige Kommandos (History)

```

1 use "Z:\home\kai\Work\15-Jahre-... 901
2 set mem 20m
3 use "Z:\home\kai\Work\15-Jahre-...
4 clear
5 use "Z:\home\kai\Work\15-Jahre-...
6 tab v3

```

Ergebnisse

```

2. Drop some variables or observations; see help drop.
3. Increase the amount of memory allocated to the data area using the set
memory command; see help memory.
t(901);
, set mem 20m
(20480k)
, use "Z:\home\kai\Work\15-Jahre-Deutsche-Einheit-Beck\Jugend\allbus2004.dta",
> clear
, clear
, use "Z:\home\kai\Work\15-Jahre-Deutsche-Einheit-Beck\Jugend\allbus2004.dta",
> clear
, tab v3

```

erhebungsbiet:	Freq.	Percent	Cum.
west - ost			
alte bundeslaender	1,982	67.28	67.28
neue bundeslaender	964	32.72	100.00
Total	2,946	100.00	

Variables

Name	Label	Type	Forma
v1	studienummer ...	int	%8.0
v2	erhebungsbiet: w...	int	%8.0
v3	erhebungsbiet: w...	byte	%8.0
v4	deutsche staatsang...	byte	%8.0
v5	wirtschaftslage in d...	byte	%8.0
v6	wirtschaftslage, bef...	byte	%8.0
v7	wirtschaftslage der ...	byte	%8.0
v8	wirtschaftslage, bef...	byte	%8.0
v9	freizeit: buecher les...	byte	%8.0
v10	freizeit: zeitschrifte...	byte	%8.0
v11	freizeit: lp, cd, cass...	byte	%8.0
v12	freizeit: videokasse...	byte	%8.0
v13	freizeit: mit dem co...	byte	%8.0
v14	freizeit: internet, onl...	byte	%8.0
v15	freizeit: sich privat ...	byte	%8.0
v16	freizeit: einfach ric...	byte	%8.0
v17	freizeit: spaziereng...	byte	%8.0
v18	freizeit: yoga, medit...	byte	%8.0
v19	freizeit: essen oder ...	byte	%8.0
v20	freizeit: besuch nac...	byte	%8.0
v21	freizeit: besuch fam...	byte	%8.0
v22	freizeit: menschenhal...	byte	%8.0

Eingaben

```

Command

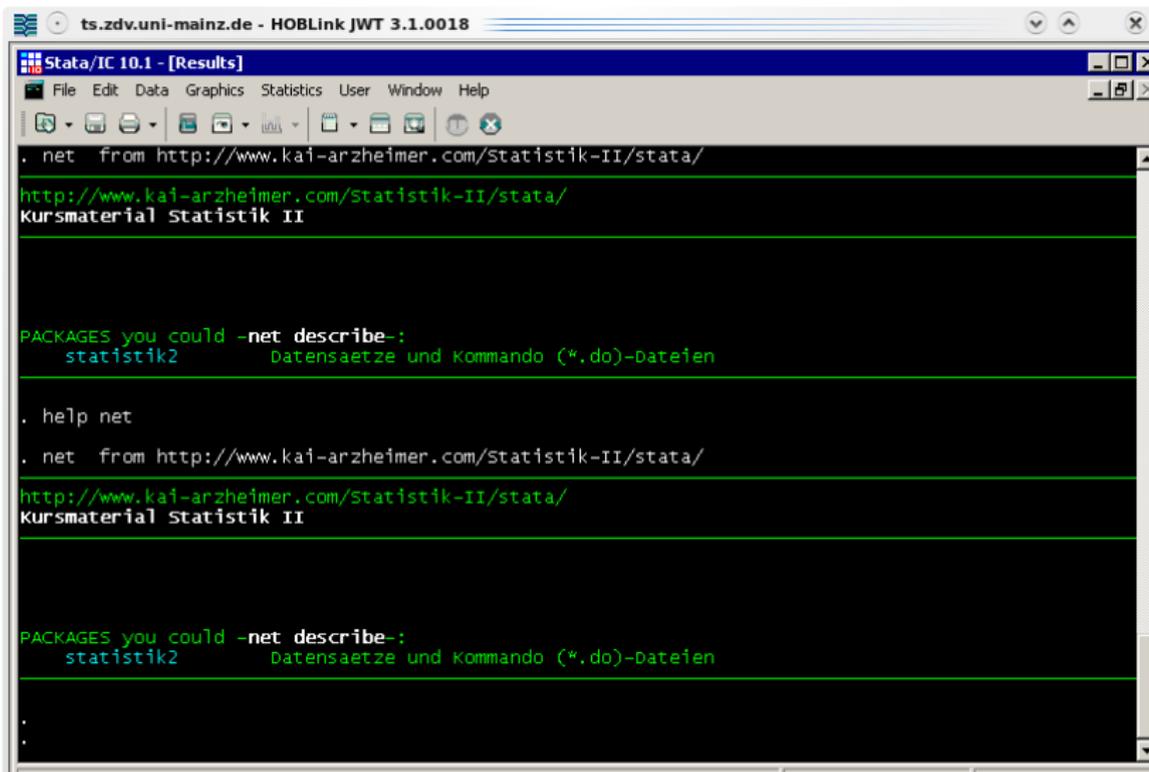
```

Z:\home\kai

Interaktivität

- ▶ Datensätze und Kommandodateien stehen über die Homepage zur Verfügung
- ▶ Zugriff ist aus Stata heraus möglich
- ▶ `net from`
`http://www.kai-arzheimer.com/Statistik-II/stata/`

Paket zur Vorlesung



```
ts.zdv.uni-mainz.de - HOBLink JWT 3.1.0018
Stata/IC 10.1 - [Results]
File Edit Data Graphics Statistics User Window Help
. net from http://www.kai-arzheimer.com/Statistik-II/stata/
http://www.kai-arzheimer.com/Statistik-II/stata/
Kursmaterial Statistik II

PACKAGES you could -net describe-:
  statistik2      Datensätze und Kommando (*.do)-Dateien

. help net
. net from http://www.kai-arzheimer.com/Statistik-II/stata/
http://www.kai-arzheimer.com/Statistik-II/stata/
Kursmaterial Statistik II

PACKAGES you could -net describe-:
  statistik2      Datensätze und Kommando (*.do)-Dateien
```

Interaktivität

- ▶ Datensätze und Kommandodateien stehen über die Homepage zur Verfügung
- ▶ Zugriff ist aus Stata heraus möglich
- ▶ `net from`
`http://www.kai-arzheimer.com/Statistik-II/stata/`
- ▶ Per Klicken oder per `net get allbus2008` kann das Paket `allbus2008` geladen und lokal installiert werden
- ▶ Das Paket enthält eine Version des ALLBUS 2008 sowie eine `*.do` Datei mit Kommandos

Aufbau eines Stata-Befehls

- ▶ Viele optionale Komponenten, aber einfache Struktur

```
[by varlist:] command [varlist] [in range] [if exp]  
[filename], [options]
```

- ▶ `by...`: – Befehl für Subgruppen wiederholen
- ▶ `varlist` – Befehl auf diese Variablen anwenden
- ▶ `range` – Befehl auf Gruppe von Fällen anwenden (Reihenfolge)
- ▶ `if` – Befehl auf Gruppe von Fällen anwenden (logische Bedingung)
- ▶ `filename` – Dateinamen (oft mit `using` eingeleitet)
- ▶ `,` `options` – Optionen für Befehl festlegen

```
use allbus2008, clear
```

Was ist der ALLBUS 2008?

- ▶ „Allgemeine Bevölkerungsumfrage der Sozialwissenschaften“ (+ISSP)
- ▶ „Goldstandard“
- ▶ Mehr-Themen-Befragung, seit 1980 alle zwei Jahre mit neuen Befragten wiederholt
- ▶ Regelmäßige Replikation einzelner Fragenblöcke
- ▶ Design?
- ▶ Themen 2008 u. a.
 - ▶ Politische Einstellungen, politische Partizipation, politische Entfremdung und Protest
 - ▶ Staatsbürgerschaft, Herkunft, Nationalstolz
 - ▶ Medien
 - ▶ Soziales Kapital und Freizeit
 - ▶ Wirtschaft

Häufigkeitstabellen

- ▶ Wie häufig kommen einzelne Kategorien einer Variablen vor?
- ▶ Z. B. Straftäter härter bestrafen?

Allbus Fragebogen

[F015_a]

⇒ Antwortschema zu Frage 15 : Aussagen bitte vorlesen! Liste 15 liegt vor!

	Stimme voll und ganz zu	Stimme eher zu	Weder noch	Stimme eher nicht zu	Stimme überhaupt nicht zu	Weiß nicht	KA
Einwanderer sollten verpflichtet werden, sich den deutschen Sitten und Gebräuchen anzupassen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Die Politik sollte sich aus der Wirtschaft heraushalten.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Zum Schutz der Umwelt sollten härtere Maßnahmen getroffen werden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Gleichgeschlechtliche Ehen sollten gesetzlich anerkannt werden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Frauen und Männer sollten bei Bewerbungen und Beförderungen gleich behandelt werden.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Straftäter sollten härter bestraft werden als bisher.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

⇒ Fortsetzung nächste Seite!

Stimme voll und ganz zu

Stimme eher zu

Weder noch

Stimme eher nicht zu

Stimme überhaupt nicht zu

Häufigkeitstabellen

- ▶ Wie häufig kommen einzelne Kategorien einer Variablen vor?
- ▶ Z. B. Straftäter härter bestrafen?
- ▶ Wie findet man die passende Variable?
- ▶ `describe`

describe

The screenshot shows the Stata 10.1 interface with the following content:

Command window:

```

1  use v11
2  use http://www.kai-archheiner.com... 190
3  use http://www.kai-archheiner.com...
4  list* 100
5  label v13
6  label v13
7  nonmissing
8
9  d v1
10 d v11
11 lab v13
12 help describe
13 d
14 describe
15 help describe
16 describe
    
```

Output window:

```

*-----*
*       data from http://www.kai-archheiner.com/statistik-ii/stata/wib2008.dta
*-----*
obs:      3,469          ALBUCompact 2008, Stat-Termberson
vars:     400          25 Aug 2009 16:25
total size: 1,873,861 (63.4% of memory free)
    
```

variable name	type	format	label	variable label
v1	xx	%10.0g	v1	STUDENUMMER
v2	ix	%10.0g	v2	IDENTIFIKATIONSNUMMER DES BEFRAGTEN
v3	ix	%10.0g	v3	EMERGENCYEXIT - SCHWELZET IN MEIST - GOST
v4	byte	%10.0g	v4	DEUTSCHE STAATSBÄHNERREISE
v5	byte	%10.0g	v5	ATTRAKTIVITÄT DES BEFR., INTERVIEWSTART
v6	byte	%10.0g	v6	WIRTSCHAFTSLAGE IN DER BRD HEUTE
v7	byte	%10.0g	v7	WIRTSCHAFTSLAGE, BEFR. HEUTE
v8	byte	%10.0g	v8	WIRTSCHAFTSLAGE DER BRD IN 1 JAHR
v9	byte	%10.0g	v9	WIRTSCHAFTSLAGE, BEFR. IN 1 JAHR
v10	byte	%10.0g	v10	REGIERUNGSBILANZIERUNG WIRTSCHAFTSLAGE
v11	double	%10.0g	v11	HAUFIGKEIT VON FERIENEN PRO WOCHE
v12	xx	%10.0g	v12	FERIENGESAMTDauer PRO TAG IN MINUTEN
v13	byte	%10.0g	v13	FERIENGESAMTDauer PRO TAG, KAT.
v14	byte	%10.0g	v14	NACHFRAGENKOMMUM - GEFÜHLT ALS TV
v15	double	%10.0g	v15	KONGUMHAELFKG. - NACHRICHTEN OFF. TV
v16	byte	%10.0g	v16	NACHRICHTENKOMMUM - PRIVATES TV
v17	double	%10.0g	v17	KONGUMHAELFKG. - NACHRICHTEN PRIVATES TV
v18	double	%10.0g	v18	HAUFIGKEIT TAGESZEITUNG LEBEN PRO WOCHE
v19	byte	%10.0g	v19	INTERNET FLER INFORMATION UEBER POLITIK
v20	byte	%10.0g	v20	HAUFIGKEIT INTERNET FLER POLITIK-INFO
v21	byte	%10.0g	v21	ERWARTENDE ZU FERIENEN-VERPFLICHTUNG
v22	byte	%10.0g	v22	POLITIK ALS WIRTSCHAFT HERAUSHALTEN
v23	byte	%10.0g	v23	HAERTERE MASSNAHMEN FLER UMWELTSCHUTZ
v24	byte	%10.0g	v24	ANERKENNUNG GLEICHGESCHLECHTLICHER EBEN
v25	byte	%10.0g	v25	HAERTERE STRAFEN FLER STRAFRECHT
v26	byte	%10.0g	v26	HAERTERE STRAFEN FLER STRAFRECHT
v27	byte	%10.0g	v27	REGIERUNG, SOZ. SICHERUNG AN WICHTIGSTEN
v28	byte	%10.0g	v28	UMVERTEILUNG ZU GUNSTEN EINFACHER LEUTE
v29	byte	%10.0g	v29	ERWARTENDE GUT F. DEUTSCHE WIRTSCHAFT
v30	byte	%10.0g	v30	DEUTSCHES MILITÄR F. KRIEG GEGEN TERROR
v31	byte	%10.0g	v31	UNTER ALLEN UMWÄNDLUNGEN FREIHEIT ACHTEN
v32	byte	%10.0g	v32	MULTIPLIKATIONSPUNKT BEI WICHL. ALLES?
v33	byte	%10.0g	v33	GGPS, AN OFFENEN POLITIK, NEHMEN SACHEN
v34	byte	%10.0g	v34	GGPS, MICH AN WAHLEN BETEILIGEN
v35	byte	%10.0g	v35	GGPS, AN OFF. DEKLARATIONEN BETEILIGEN
v36	byte	%10.0g	v36	GGPS, IN EINER BILDGEGENSTÄNDLICHE METERS.

Variable labels window:

Name	Label
v1	STUDENUMMER
v2	IDENTIFIKATIONSNUMMER DES BEFRAGTEN
v3	EMERGENCYEXIT - SCHWELZET IN MEIST - GOST
v4	DEUTSCHE STAATSBÄHNERREISE
v5	ATTRAKTIVITÄT DES BEFR., INTERVIEWSTART
v6	WIRTSCHAFTSLAGE IN DER BRD HEUTE
v7	WIRTSCHAFTSLAGE, BEFR. HEUTE
v8	WIRTSCHAFTSLAGE DER BRD IN 1 JAHR
v9	WIRTSCHAFTSLAGE, BEFR. IN 1 JAHR
v10	REGIERUNGSBILANZIERUNG WIRTSCHAFTSLAGE
v11	HAUFIGKEIT VON FERIENEN PRO WOCHE
v12	FERIENGESAMTDauer PRO TAG IN MINUTEN
v13	FERIENGESAMTDauer PRO TAG, KAT.
v14	NACHFRAGENKOMMUM - GEFÜHLT ALS TV
v15	KONGUMHAELFKG. - NACHRICHTEN OFF. TV
v16	NACHFRAGENKOMMUM - PRIVATES TV
v17	KONGUMHAELFKG. - NACHRICHTEN PRIVATES TV
v18	HAUFIGKEIT TAGESZEITUNG LEBEN PRO WOCHE
v19	INTERNET FLER INFORMATION UEBER POLITIK
v20	HAUFIGKEIT INTERNET FLER POLITIK-INFO
v21	ERWARTENDE ZU FERIENEN-VERPFLICHTUNG
v22	POLITIK ALS WIRTSCHAFT HERAUSHALTEN
v23	HAERTERE MASSNAHMEN FLER UMWELTSCHUTZ
v24	ANERKENNUNG GLEICHGESCHLECHTLICHER EBEN
v25	HAERTERE STRAFEN FLER STRAFRECHT
v26	HAERTERE STRAFEN FLER STRAFRECHT
v27	REGIERUNG, SOZ. SICHERUNG AN WICHTIGSTEN
v28	UMVERTEILUNG ZU GUNSTEN EINFACHER LEUTE
v29	ERWARTENDE GUT F. DEUTSCHE WIRTSCHAFT
v30	DEUTSCHES MILITÄR F. KRIEG GEGEN TERROR
v31	UNTER ALLEN UMWÄNDLUNGEN FREIHEIT ACHTEN

Häufigkeitstabellen

- ▶ Wie häufig kommen einzelne Kategorien einer Variablen vor?
- ▶ Z. B. Straftäter härter bestrafen?
- ▶ Wie findet man die passende Variable?
- ▶ describe
- ▶ Besser: lookfor straf

lookfor straf

Stata/IC 10.1 - http://www.kal-arzhelms.com/Statistik-II/stata/wilbus2008.dta - [Results]

File Edit Data Graphics Statistics User Window Help

```

1  *-----*
2  use http://www.kal-arzhelms.com... 190
3  use http://www.kal-arzhelms.com...
4  label var 190
5  label var 100
6  label var v33
7  nonmissing
8  tab v33
9  d v33
10 tab v33
11 tab v23
12 help label
13 d
14 describe
15 help describe
16 describe
17 lookfor straf
    
```

lookfor: Name Label

Name	Label
v1	STUDBUCHNUMMER
v2	IDENTIFIZIERT KONTAKTNUMMER DES BEFRAGTEN
v3	BEFRAGNGEGEBET "GROEBEREIT" = WEIST -LOST
v4	DEUTSCHE STAATSBURGHERSCHAFT
v5	ATTRAKTIVITAET DES BEFR. -POSTENSTABAT
v6	WIRTSCHAFTSLAGE PRO TAG DER HEUTE
v7	WIRTSCHAFTSLAGE BEFR. HEUTE
v8	WIRTSCHAFTSLAGE DES BEFR. IN 1 JAHR
v9	WIRTSCHAFTSLAGE, BEFR. IN 1 JAHR
v10	REGIERUNGSEMPFANGUNG WIRTSCHAFTSLAGE
v11	HANDLUNGSZEIT FÜR PERSEKUT. PRO WOCHE
v12	PERSEKUT. DURCHLAUF PRO TAG IN MINUTEN
v13	PERSEKUT. DURCHLAUF PRO TAG, KAT.
v14	NACHRICHTENKONSUM: OBERFLÄCHLICHE TV
v15	KONSUM: TIEFENGEH. - NACHRICHTEN: OEFF. TV
v16	NACHRICHTENKONSUM: PRIVATE TV
v17	KONSUM: TIEFENGEH. - NACHRICHTEN: PRIVATE TV
v18	HANDLUNGSZEIT TAGS: UNTERRICHTEN PRO WOCHE
v19	INTEREST FÜR INFORMATION ÜBER KULTUR
v20	HANDLUNGSZEIT: INTERNET FÜR POLITIKINFO
v21	ENKWAERDER: ZUM ANWANDERUNG VERFUHRER
v22	POLITIK AUS WIRTSCHAFT HERAUSZUHALTEN
v23	HANDTERR. MASSNAHMEN VOR WIRTSCHAFTSLEIST.
v24	ANERKENNUNG GLEICHBERECHTIGTER ERHEB.
v25	BEISPI. GLEICHBERECHTIGUNG V. MANN UND FRAU
v26	HANDTERR. STRAFEN FÜR STRAFKAT.?
v27	REGIERUNG: SOZ. SICHERUNG AN WICHTIGSTEN?
v28	UMFERTILGUNG ZU KANTON ERWERBENS LEIST.
v29	ENKWAERDER GUT F. DEUTSCHE WIRTSCHAFT
v30	DEUTSCHE MÜLTER F. BERU. GEGEN TEROR.
v31	UNTER ALLEN UMFRAGEN FREIHEIT ACHTEN?

variable name	type	format	label	var label
v26	byte	%10.0g	v26	HAERTERE STRAFEN FÜR STRAFKAT.?

lookfor straf

2: http://www.kal-arzhelms.com/Statistik-II/stata/wilbus2008.dta - [Statafile-1]

Häufigkeitstabellen

- ▶ Wie häufig kommen einzelne Kategorien einer Variablen vor?
- ▶ Z. B. Straftäter härter bestrafen?
- ▶ Wie findet man die passende Variable?
- ▶ `describe`
- ▶ Besser: `lookfor straf`
- ▶ Label = Etikett für interne Werte
 - ▶ Variablennamen
 - ▶ Numerische Werte von Variablen
- ▶ Gesuchte Variable: `v26`
- ▶ Häufigkeitstabelle: `tabulate v26` oder `tab v26`

Häufigkeiten: Härtere Strafen

```
. tab v26
```

HAERTERE STRAFEN FUER STRAFSTAETER?	Freq.	Percent	Cum.
STIMME VOLL ZU	1,882	54.25	54.25
STIMME EHER ZU	781	22.51	76.77
WEDER NOCH	429	12.37	89.13
STIMME EHER NICHT ZU	210	6.05	95.19
STIMME GAR NICHT ZU	91	2.62	97.81
WEISS NICHT	62	1.79	99.60
KEINE ANGABE	14	0.40	100.00
Total	3,469	100.00	

Häufigkeiten: Härtere Strafen

```
. lab list v26
```

```
v26:
```

1. STIMME VOLL ZU
2. STIMME EHER ZU
3. WEDER NOCH
4. STIMME EHER NICHT ZU
5. STIMME GAR NICHT ZU
8. WEISS NICHT
9. KEINE ANGABE

```
. tab v26 if v26<8
```

HAERTERE STRAFEN FUER STRAFTAETER?	Freq.	Percent	Cum.
1. STIMME VOLL ZU	1,882	55.47	55.47
2. STIMME EHER ZU	781	23.02	78.49
3. WEDER NOCH	429	12.64	91.13
4. STIMME EHER NICHT ZU	210	6.19	97.32
5. STIMME GAR NICHT ZU	91	2.68	100.00
Total	3,393	100.00	

Gibt es Unterschiede zwischen den Geschlechtern?

Häufigkeiten: Härtere Strafen

```
. lookfor geschl
```

variable name	storage type	display format	value label	variable label
v24	byte	%10.0g	v24	ANERKENNUNG GLEICHGESCHLECHTLICHER EHEN?
v151	byte	%10.0g	v151	GESCHLECHT, BEFRAGTE<R>

```
. tab v151
```

GESCHLECHT, BEFRAGTE<R>	Freq.	Percent	Cum.
MANN	1,712	49.35	49.35
FRAU	1,757	50.65	100.00
Total	3,469	100.00	

```
. tab v26 v151
```

HAERTERE STRAFEN FUER STRAFSTAETER?	GESCHLECHT, BEFRAGTE<R>		Total
	MANN	FRAU	
STIMME VOLL ZU	876	1,006	1,882
STIMME EHER ZU	405	376	781
WEDER NOCH	233	196	429
STIMME EHER NICHT ZU	113	97	210
STIMME GAR NICHT ZU	53	38	91
WEISS NICHT	23	39	62
KEINE ANGABE	9	5	14
Total	1,712	1,757	3,469

Häufigkeiten: Härtere Strafen

```
. tab v26 v151 if v26<8 ,col
```

Key
<i>frequency</i>
<i>column percentage</i>

HAERTERE STRAFEN FUER STRAFSTAETER?	GESCHLECHT, BEFRAGTE<R>		Total
	MANN	FRAU	
STIMME VOLL ZU	876 52.14	1,006 58.73	1,882 55.47
STIMME EHER ZU	405 24.11	376 21.95	781 23.02
WEDER NOCH	233 13.87	196 11.44	429 12.64
STIMME EHER NICHT ZU	113 6.73	97 5.66	210 6.19
STIMME GAR NICHT ZU	53 3.15	38 2.22	91 2.68
Total	1,680 100.00	1,713 100.00	3,393 100.00

Häufigkeiten: Härtere Strafen

```
. tab v26 v151 if v26<8 ,row
```

Key
<i>frequency</i>
<i>row percentage</i>

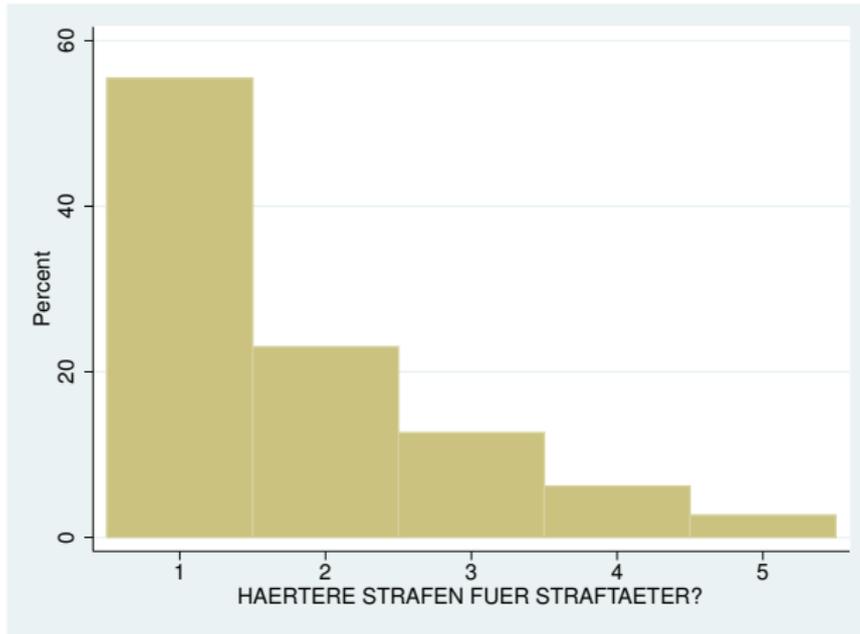
HAERTERE STRAFEN FUER STRAFSTAETER?	GESCHLECHT, BEFRAGTE<R>		Total
	MANN	FRAU	
STIMME VOLL ZU	876 46.55	1,006 53.45	1,882 100.00
STIMME EHER ZU	405 51.86	376 48.14	781 100.00
WEDER NOCH	233 54.31	196 45.69	429 100.00
STIMME EHER NICHT ZU	113 53.81	97 46.19	210 100.00
STIMME GAR NICHT ZU	53 58.24	38 41.76	91 100.00
Total	1,680 49.51	1,713 50.49	3,393 100.00

Grafikbefehle

- ▶ Aus historischen Gründen verschiedene Interfaces
- ▶ Teils sehr komplexe Optionen
- ▶ Handbücher, Kohler/Kreuter, Visual Guide
- ▶ Kategoriale Variable, eine Dimension – warum Histogramm eigentlich nicht geeignet?

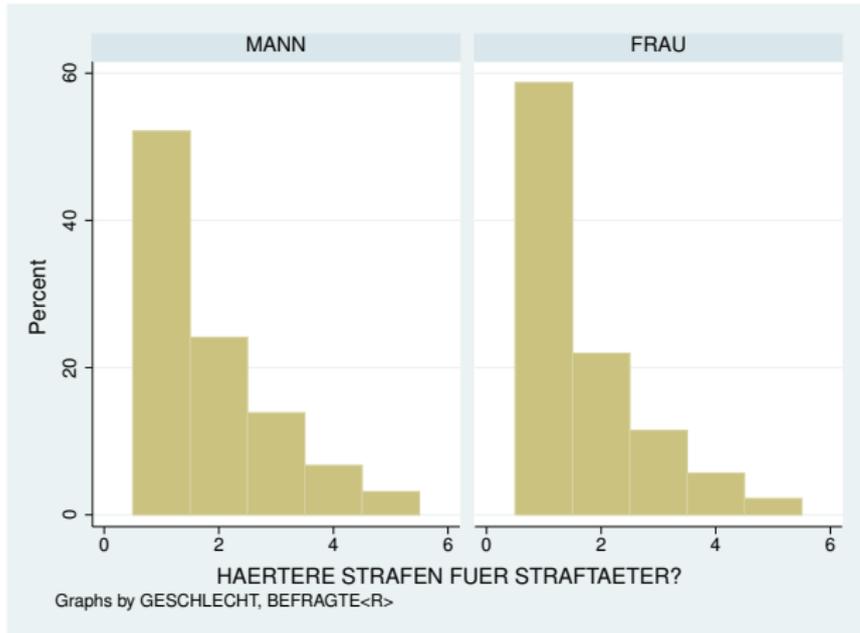
Barplot/Histogramm

```
. graph twoway hist v26 if v26<8,disc percent
```



Barplot/Histogramm nach Geschlecht

```
. graph twoway hist v26 if v26<8, disc percent by(v151)
```



„Echte“ Balkendiagramme

- ▶ Erfordern einen Trick:
- ▶ Kategoriale Variablen müssen in eine Serie binärer Variablen (0/1 kodiert = „Dummies“) zerlegt werden
- ▶ Jede Variable mit k Kategorien kann durch $k - 1$ Dummies ersetzt werden (andere Kodierungen möglich)
- ▶ Beispiel Konfession („katholisch“, „protestantisch“, „andere“)
 - ▶ Dummies für katholisch/protestantisch (andere= weder katholisch noch protestantisch)
 - ▶ Dummies für andere/protestantisch (katholisch= weder andere noch protestantisch)
 - ▶ Dummies für andere/katholisch (protestantisch= weder andere noch katholisch)
- ▶ Dritter Dummy redundant (wichtig für Regression)

Wie generiert man Dummies?

- ▶ Befehle um Variablen zu kopieren, erzeugen, verändern

```
. generate dummy1=0  
. replace dummy1=1 if v26==1  
(1882 real changes made)
```

Wie generiert man Dummies?

- Befehle um Variablen zu kopieren, erzeugen, verändern

```
. generate dummy2=v26==2  
. tab dummy1 dummy2
```

dummy1	dummy2		Total
	0	1	
0	806	781	1,587
1	1,882	0	1,882
Total	2,688	781	3,469

```
. tab v26 dummy1
```

HAERTERE STRAFEN FUER STRAFTAETER?	dummy1		Total
	0	1	
STIMME VOLL ZU	0	1,882	1,882
STIMME EHER ZU	781	0	781
WEDER NOCH	429	0	429
STIMME EHER NICHT ZU	210	0	210
STIMME GAR NICHT ZU	91	0	91
WEISS NICHT	62	0	62
KEINE ANGABE	14	0	14
Total	1,587	1,882	3,469

v26 umkodieren/fehlende Werte löschen?

- ▶ Werte größer 5 sollten automatisch ignoriert werden (missing)

```
. gen strafe = v26 if v26<8  
(76 missing values generated)
```

```
.
```

```
. tab strafe
```

strafe	Freq.	Percent	Cum.
1	1,882	55.47	55.47
2	781	23.02	78.49
3	429	12.64	91.13
4	210	6.19	97.32
5	91	2.68	100.00
Total	3,393	100.00	

```
.
```

v26 umkodieren/fehlende Werte löschen?

- ▶ Werte größer 5 sollten automatisch ignoriert werden (missing)
- ▶ Hohe Werte = Ablehnung: kontraintuitiv

```
. replace strafe = 6 - strafe  
(2964 real changes made)
```

```
. tab strafe
```

strafe	Freq.	Percent	Cum.
1	91	2.68	2.68
2	210	6.19	8.87
3	429	12.64	21.51
4	781	23.02	44.53
5	1,882	55.47	100.00
Total	3,393	100.00	

Aus der neuen Variable fünf Dummies erzeugen

- ▶ Mehrere Hilfsmittel
- ▶ Sehr schnell:

Aus der neuen Variable fünf Dummies erzeugen

- ▶ Mehrere Hilfsmittel
- ▶ Sehr schnell:

```
. tab strafe,gen(mehrstraf)
```

strafe	Freq.	Percent	Cum.
1	1,882	55.47	55.47
2	781	23.02	78.49
3	429	12.64	91.13
4	210	6.19	97.32
5	91	2.68	100.00
Total	3,393	100.00	

```
. d mehrst*
```

variable name	storage type	display format	value label	variable label
mehrstraf1	byte	%8.0g	strafe== 1.0000	
mehrstraf2	byte	%8.0g	strafe== 2.0000	
mehrstraf3	byte	%8.0g	strafe== 3.0000	
mehrstraf4	byte	%8.0g	strafe== 4.0000	
mehrstraf5	byte	%8.0g	strafe== 5.0000	

Aus der neuen Variable fünf Dummies erzeugen

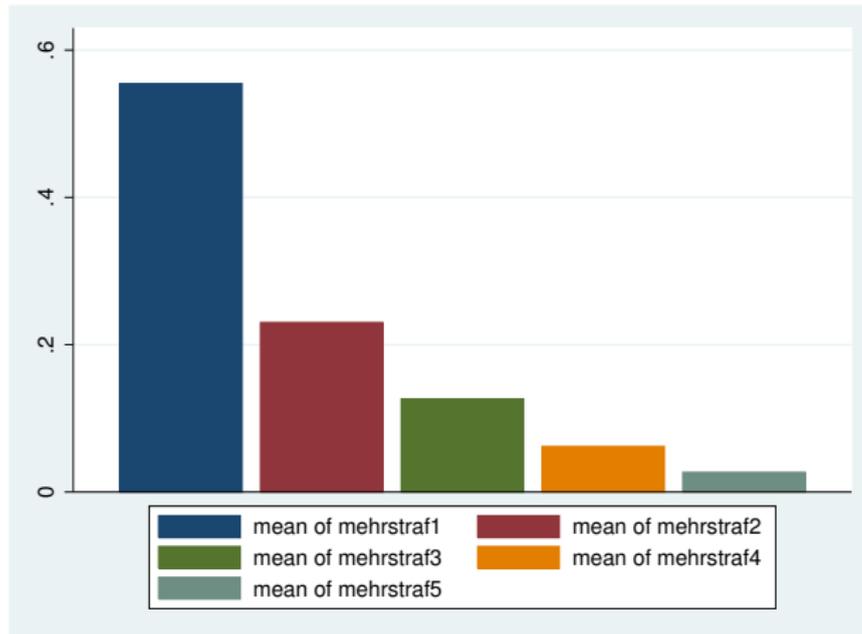
- ▶ Mehrere Hilfsmittel
- ▶ Sehr schnell:

```
. summ mehrstraf*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mehrstraf1	3393	.5546714	.4970753	0	1
mehrstraf2	3393	.2301798	.4210098	0	1
mehrstraf3	3393	.1264368	.33239	0	1
mehrstraf4	3393	.0618921	.240995	0	1
mehrstraf5	3393	.0268199	.1615807	0	1

```
. graph bar mehr*, bargap(15)
```

Echtes Balkendiagramm

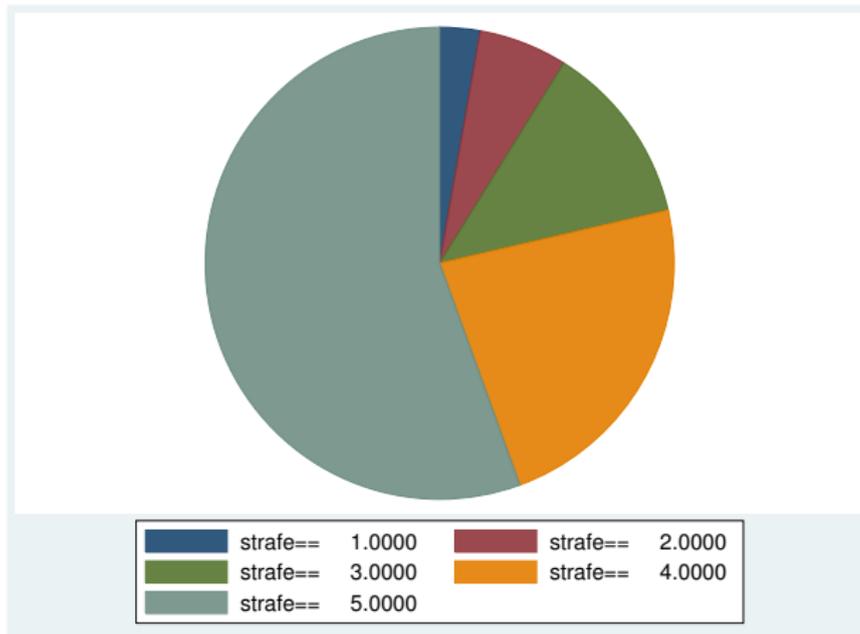


Tortendiagramme...

- ▶ ...sind unerwünscht
- ▶ ...können mit dem gleichen Trick (Dummies) erzeugt werden

```
. graph pie mehrstraf*
```

Tortendiagramme...

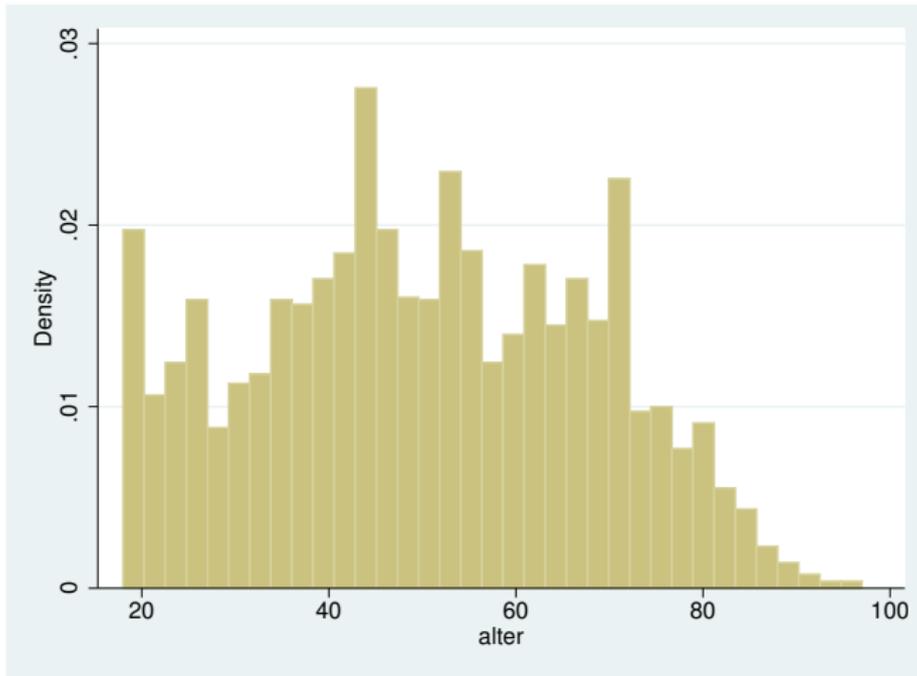


Histogramm und Dichteschätzung

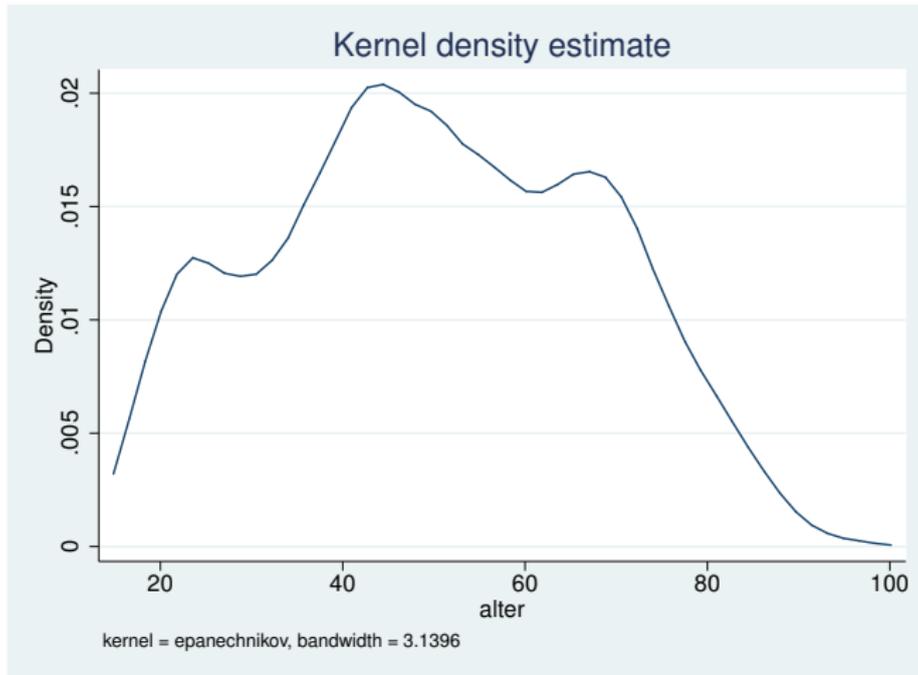
- ▶ (Konzeptuell) kontinuierliche Variablen
- ▶ Z. B. Alter (18-999)
- ▶ Alter > 120 auf missing setzen

```
. gen alter = v154 if v154<121  
(12 missing values generated)  
. hist alter  
(bin=35, start=18, width=2.2571429)  
. kdensity alter  
. kdensity alter,by(v151)  
option by() not allowed  
r(191);  
. graph twoway kdensity alter,by(v151)
```

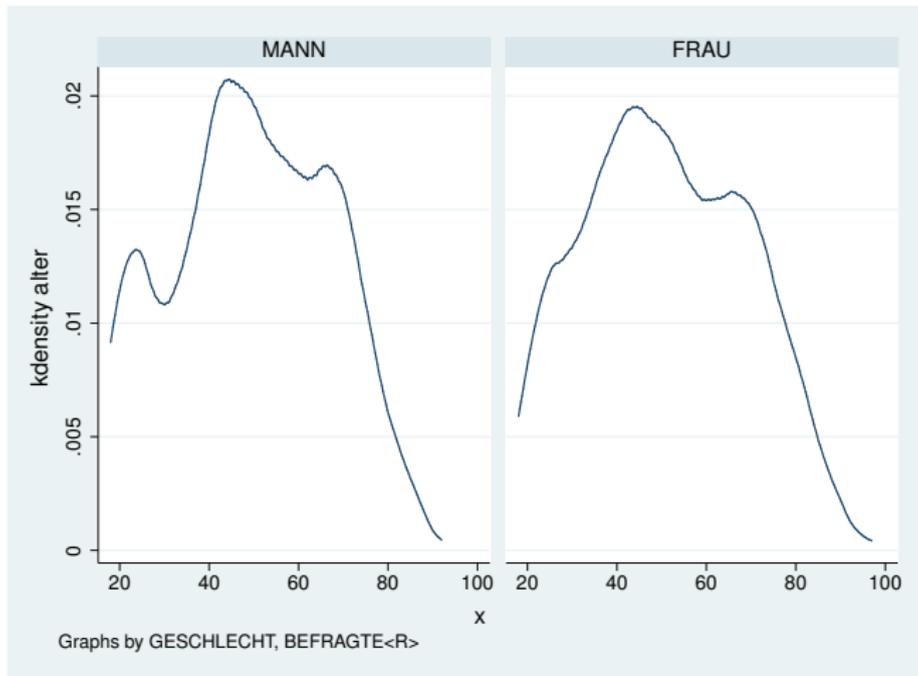
Alter: Histogramm



Alter: Dichteschätzung



Alter: Dichteschätzung nach Geschlecht



Vergleich mit theoretischer Verteilung

- ▶ Ist das Alter normalverteilt?
- ▶ Normalverteilung (Dichte)
 - ▶ Symmetrisch
 - ▶ Glockenförmig
 - ▶ Ca. 90% der Fläche ± 1.64 Standardabweichungen vom Mittelwert; 95% der Fläche ± 1.96 Standardabweichungen
- ▶ Kumulierte Normalverteilung: Integral der Dichteverteilung
- ▶ Wichtiges Modell für viele reale und theoretische Verteilungen
- ▶ Dichte und kumulierte Verteilungen sind *Funktionen*
- ▶ x - Werten zwischen $\pm\infty$ werden y -Werte zugeordnet
- ▶ Stata kann diese und viele andere Funktionswerte bestimmen, plotten, ausgeben

Normalverteilung

- ▶ Wieviel Prozent der Verteilung liegen links der Werte 1, 2, 3?

```
. display normal(1)  
.84134475  
  
. display normal(2)  
.97724987  
  
. display normal(3)  
.9986501
```

```
. display 2+2  
4  
  
. display "Hallo!"  
Hallo!
```

Normalverteilung

- ▶ Wieviel Prozent der Verteilung liegen links der Werte 1, 2, 3?
- ▶ Welche Werte trennen von links 2.5%, 10% und 50% der (Standard)Normalverteilung ab?

```
. display invnormal(.025)  
-1.959964  
. display invnormal(.10)  
-1.2815516  
. display invnormal(.50)  
0
```

Normalverteilung

- ▶ Wieviel Prozent der Verteilung liegen links der Werte 1, 2, 3?
- ▶ Welche Werte trennen von links 2.5%, 10% und 50% der (Standard)Normalverteilung ab?

```
. display invnormal(.025)
```

```
-1.959964
```

```
. display invnormal(.10)
```

```
-1.2815516
```

```
. display invnormal(.50)
```

```
0
```

- ▶ `normal`: *kumulierte* Normalverteilung

Normalverteilung

- ▶ Wieviel Prozent der Verteilung liegen links der Werte 1, 2, 3?
- ▶ Welche Werte trennen von links 2.5%, 10% und 50% der (Standard)Normalverteilung ab?

```
. display invnormal(.025)
```

```
-1.959964
```

```
. display invnormal(.10)
```

```
-1.2815516
```

```
. display invnormal(.50)
```

```
0
```

- ▶ `normal`: *kumulierte* Normalverteilung
- ▶ `invnormal`: *Umkehrfunktion* zur kumulierten Normalverteilung

Normalverteilung

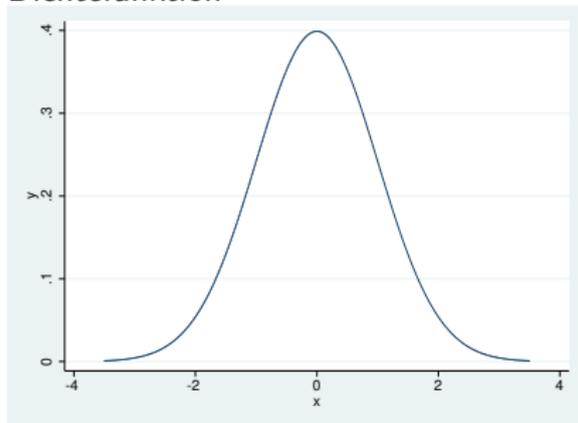
- ▶ Wieviel Prozent der Verteilung liegen links der Werte 1, 2, 3?
- ▶ Welche Werte trennen von links 2.5%, 10% und 50% der (Standard)Normalverteilung ab?

```
. display invnormal(.025)  
-1.959964  
. display invnormal(.10)  
-1.2815516  
. display invnormal(.50)  
0
```

- ▶ `normal`: *kumulierte* Normalverteilung
- ▶ `invnormal`: *Umkehrfunktion* zur kumulierten Normalverteilung
- ▶ `normalden`: *Dichtefunktion* der Normalverteilung

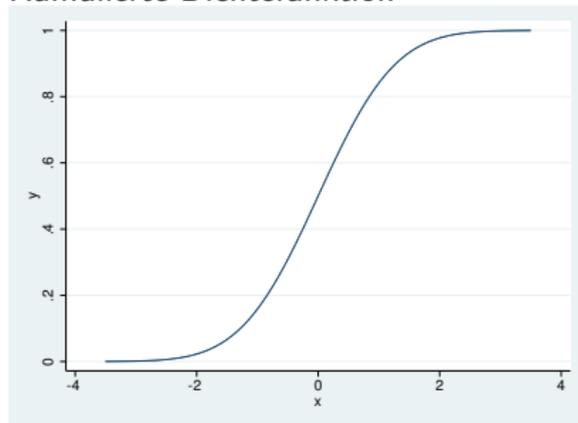
Normalverteilung graphisch

Dichtefunktion



```
. graph twoway function y=normden(x),range(-3.5 3.5)
```

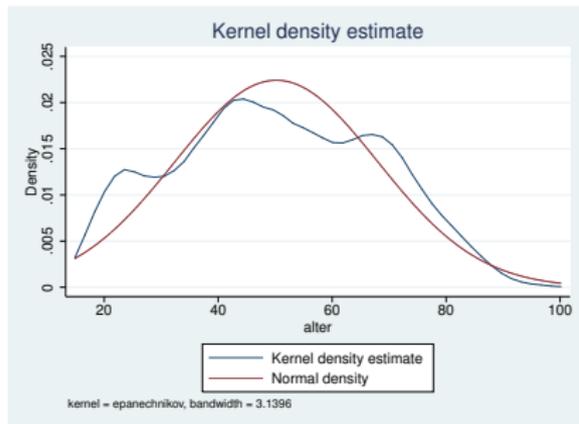
Kumulierte Dichtefunktion



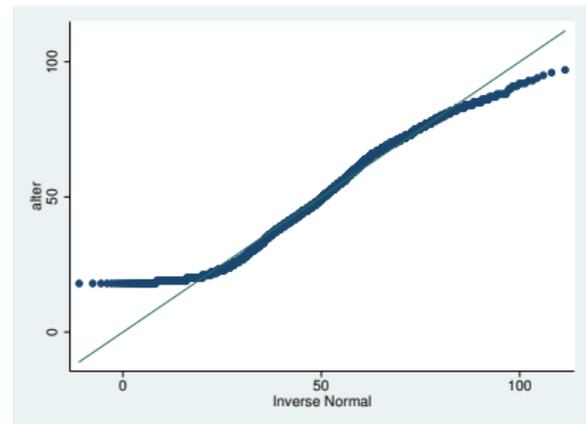
```
. graph twoway function y=normal(x),range(-3.5 3.5)
```

Alter und Normalverteilung

- ▶ Normalverteilung mit passendem Mittelwert/Varianz über Altersverteilung plotten
- ▶ Bequem: Option nutzen; genauer: Quantil-Plot



```
. kdensity alter,normal
```



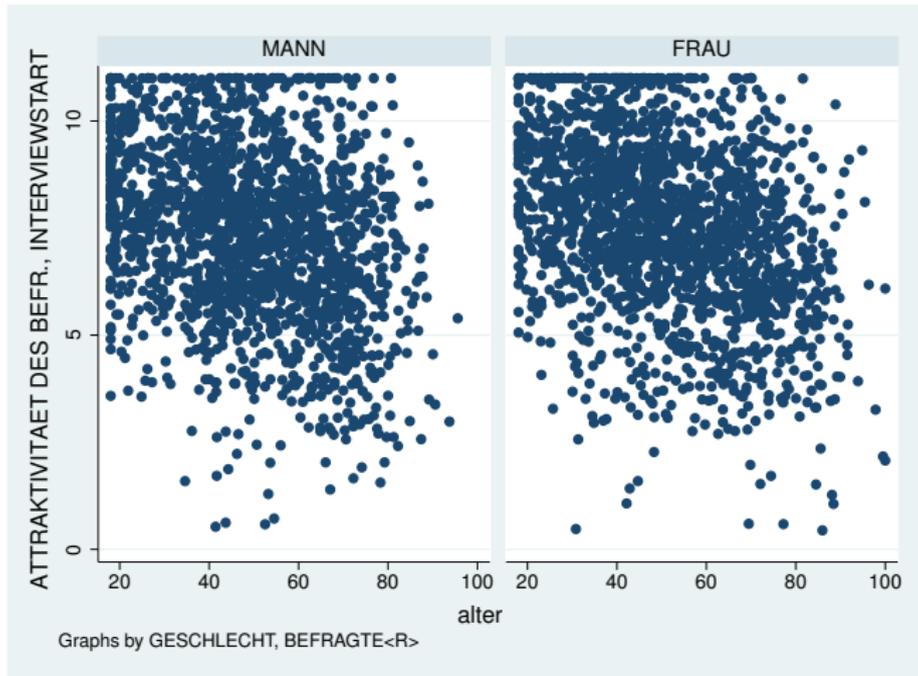
```
. qnorm alter
```

Zwei kontinuierliche Variablen: Scatterplot

- ▶ Viele Variablen nur konzeptuell kontinuierlich
- ▶ Wenige Kategorien → Punkte im Plot klumpen
- ▶ Beispiel Alter (ok) vs. Attraktivität auf den Interviewer (1-11)
- ▶ „jitter“, „random noise“

```
. graph twoway scatter v5 alter, jitter(8) by(v151)  
. graph export scatter-2.eps  
(file scatter-2.eps written in EPS format)
```

Alter und Attraktivität nach Geschlecht



Zusammenfassung

- ▶ Software essentiell für die Verwaltung, Verarbeitung und Auswertung von Daten
- ▶ Stata Programm der Wahl für die Ausbildung in der Politikwissenschaft
- ▶ Alles, was in Statistik I und hier gelernt wurde/wird, mit Stata machbar
- ▶ Wichtig: **Üben**