

Einführung

Statistik I

Sommersemester 2009

Formalia

Veranstaltungsform

Literatur

Übersicht

Einführung

Was ist „Statistik“?

Warum soll ich das lernen?

Public Policy

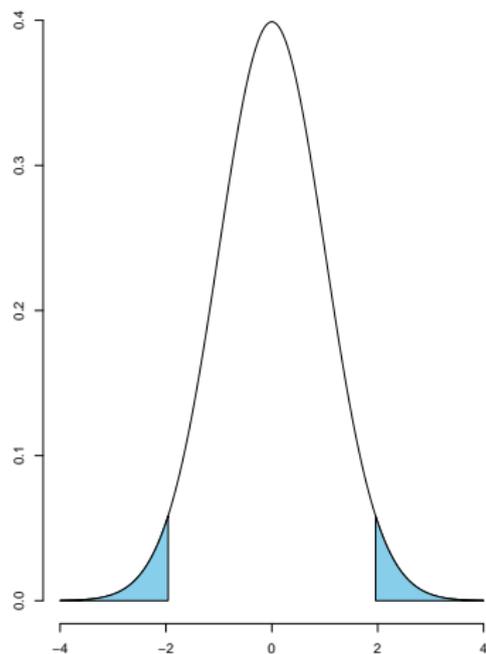
Bedeutung für Politikwissenschaft

Politik, Medien und Statistik

Eigennutz

Grundbegriffe

Zusammenfassung



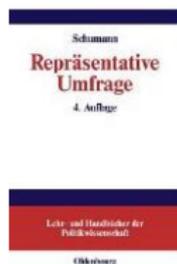
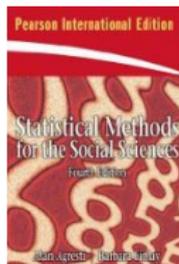
Worum geht es hier?

- ▶ Statistik I ist eine *Pflicht*vorlesung in den BA-Studiengängen
- ▶ Die Inhalte sind Bestandteil der Modulprüfung
- ▶ Das heißt nicht, daß man hier nichts fürs Leben lernen kann
- ▶ Wie kann ich das schaffen?
 - ▶ Regelmäßige (!) Vorbereitung/Nachbereitung
 - ▶ Taschenrechner
 - ▶ Besuch der Übung
- ▶ Folien etc.: <http://www.kai-arzheimer.com/Lehre-Deutschland/Statistik-I/>

Was soll ich lesen/kaufen?

Eines anschaffen

- ▶ Agresti, A./Finlay, B.: Statistical Methods for the Social Sciences. Upper Saddle River, NJ 2009
- ▶ Gehring, U./Weins C.: Grundkurs Statistik für Politologen. Wiesbaden 2006
- ▶ Schumann, S.: Repräsentative Umfrage: Praxisorientierte Einführung in empirische Methoden und statistische Analyseverfahren. München 2006



Was erwartet mich?

Vorlesung	Inhalt
1	Einführung
2	Daten und graphische Darstellungen
3	Lage- und Streuungsmaße, z-Transformation
4	Zusammenhangsmaße I
5	Zusammenhangsmaße II
6	Lineare Regression I
7	Lineare Regression II

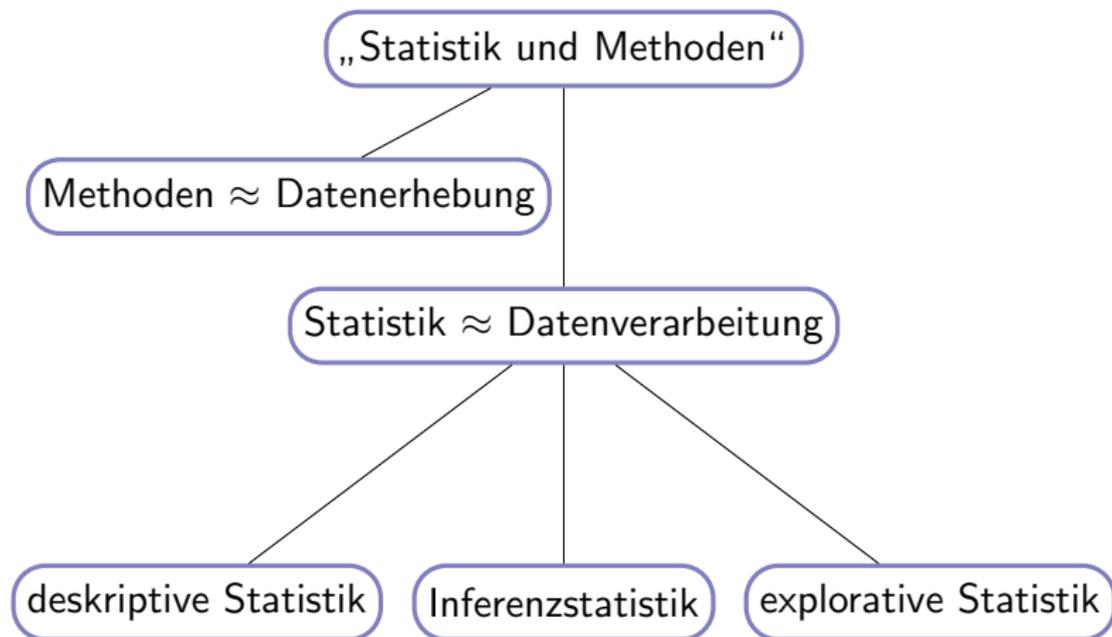
Was erwartet mich? II

Vorlesung	Inhalt
8	Nicht-parametrische Statistik
9	Grundbegriffe der Wahrscheinlichkeitsrechnung
10	Zentraler Grenzwertsatz und Konfidenzintervalle
11	Hypothesenprüfung I
12	Hypothesenprüfung II
13	Anwendungsbeispiele
14	Wiederholung

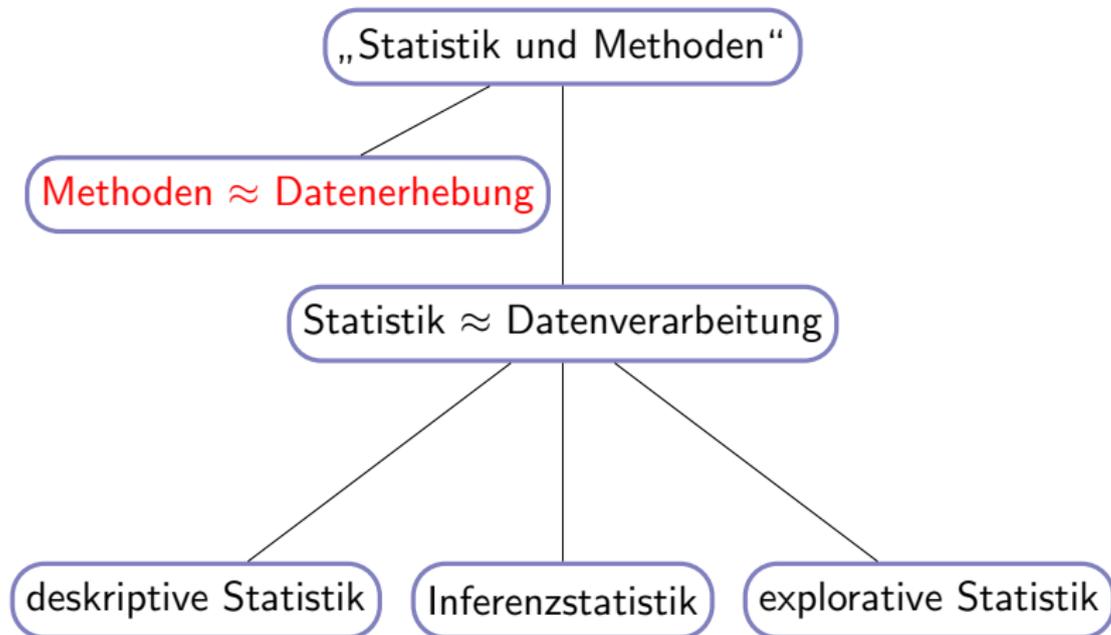
Was macht Statistik?

- ▶ Mathematische Statistik
 - ▶ Teilgebiet der Mathematik
 - ▶ Befaßt sich mit zufälligen Ereignissen/Variablen im weitesten Sinn
 - ▶ Liefert mathematische Modelle für sozialwissenschaftliche Anwendungen
 - ▶ Kommt hier nur ganz am Rande vor
- ▶ Angewandte/sozialwissenschaftliche/politikwissenschaftliche Statistik
 - ▶ Teilgebiet der sozialwissenschaftlichen Methodenlehre
 - ▶ Befaßt sich mit der *Anwendung* von Analyseverfahren auf politikwissenschaftliche Daten
 - ▶ (Mathematische Statistik liefert Grundlagen)
 - ▶ Dient der Beschreibung und Auswertung von Daten
 - ▶ Schnittstelle zwischen politikwissenschaftlichen Theorien und empirischer Forschung

Was macht Statistik? II



Was macht Statistik? II



Wozu Methoden?

- ▶ Methoden: Wie erhält man Daten?
- ▶ Erhebungsmethoden
 - ▶ Beobachtung
 - ▶ Inhaltsanalyse
 - ▶ Bevölkerungsumfrage
 - ▶ Expertenbefragung
 - ▶ Prozeßdaten
 - ▶ ...
- ▶ Forschungsdesigns
 - ▶ Experimentell
 - ▶ Quasi-experimentell
 - ▶ Ex-post-facto (Survey)
 - ▶ ...

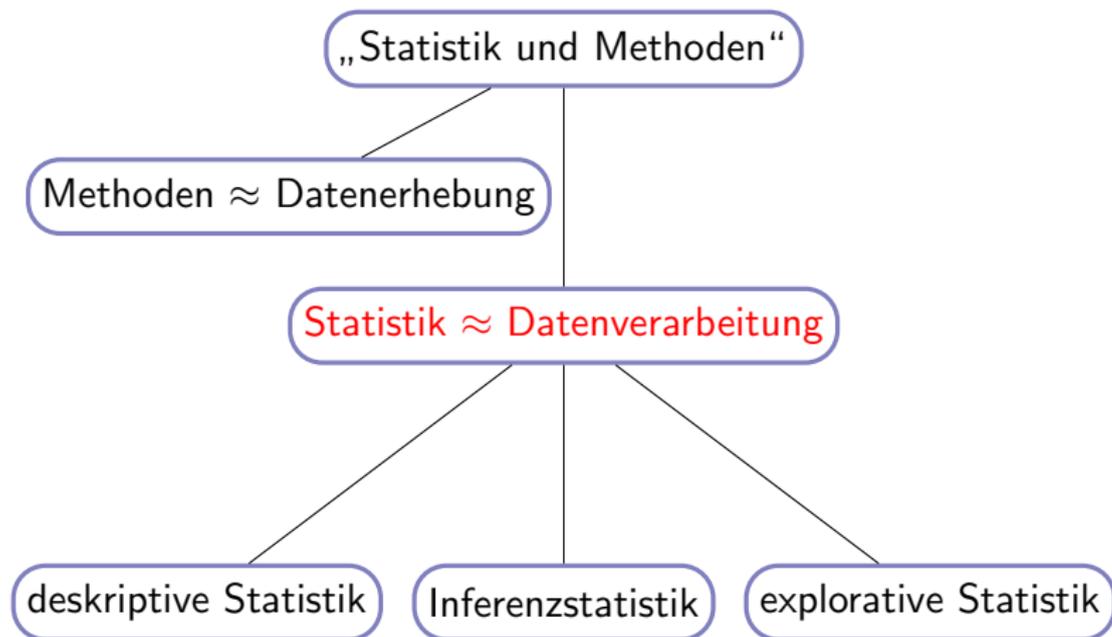
Was sind Daten?

- ▶ Empirische Beobachtungen/Messungen an einem politischen Objekt (z.B. Person, Staat, Parteiprogramm, Parlament)
- ▶ Die quantifiziert und einem Computer oder auf Papier gespeichert werden
- ▶ Verschiedene „Skalenniveaus“
 1. Ratio-Skalen
 2. Intervall-Skalen
 3. Ordinal-Skalen
 4. Nominal-Skalen

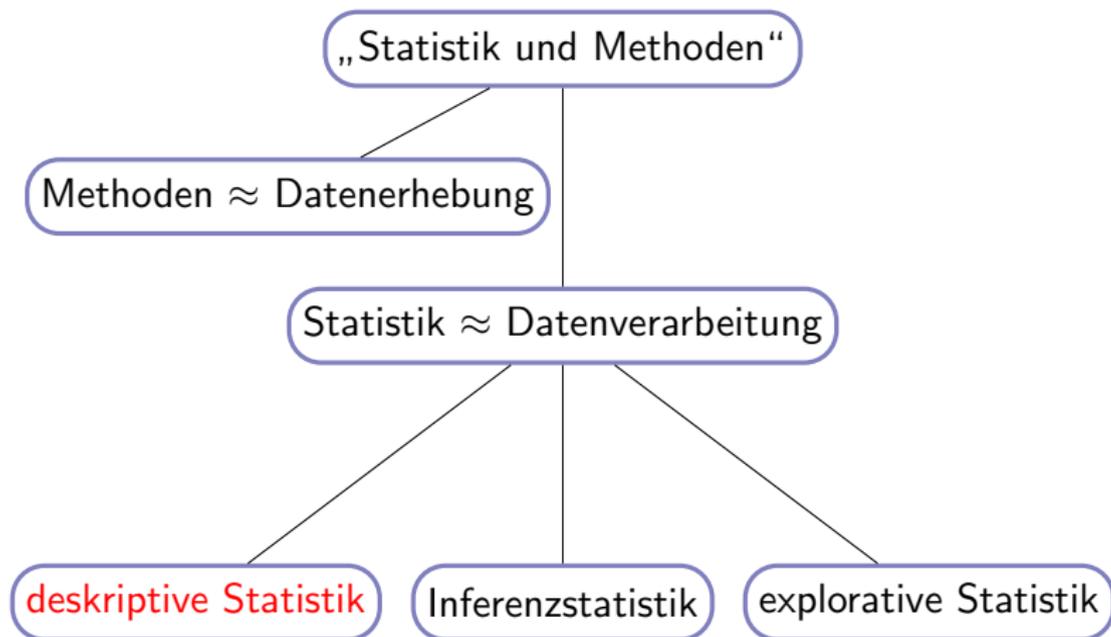
Wie war das mit den Skalenniveaus?

1. Ratio-Skalen
 - ▶ Stimmenanteil eines Kandidaten, Lebensalter in Jahren
 - ▶ Nullpunkt, Verhältnis zwischen Meßwerten interpretierbar
 2. Intervall-Skalen
 - ▶ Links-Rechts-Selbsteinstufung, Sympathiethermometer
 - ▶ Kein Nullpunkt, aber konstante Abstände
 3. Ordinal-Skalen
 - ▶ Regimetypologie: Totalitarismus, Autoritarismus, Demokratie
 - ▶ Abstände nicht konstant, aber klare Reihenfolge
 4. Nominal-Skalen
 - ▶ Parteipräferenz;
Zugehörigkeit zu: NATO, Warschauer Pakt, Blockfreie Bewegung
 - ▶ Keine Rangfolge der Kategorien
- ▶ Skalenniveau → zulässige Verfahren

Was macht Statistik? II



Was macht Statistik? II



Was bedeutet „deskriptive Statistik“?

- ▶ Variable = (numerische) Eigenschaft eines Objektes
- ▶ Deskriptive Statistik = „beschreibende Statistik“
- ▶ Information über eine oder mehrere Variablen wird verdichtet
→ statt mehrerer hundert/tausend Zahlen eine Zahl

Ein Beispiel

Stata/IC 10.1 - U:\Work\Lehre\2008-2009\SoSe-2009\Statistik-I\data-demo\sub-ess.dta - [Results]

File Edit Data Graphics Statistics User Window Help

Review

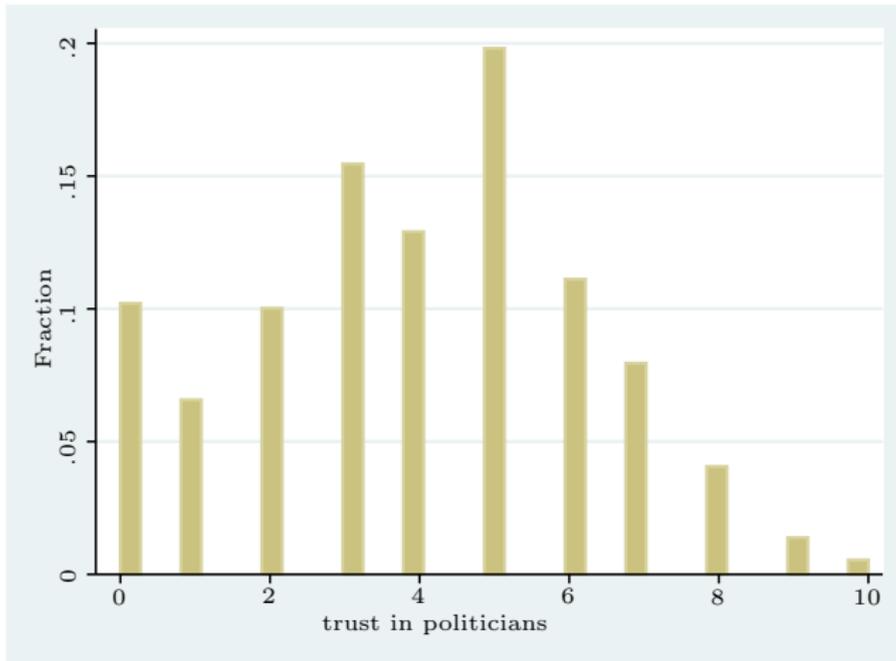
Command	_rc
1 pwd	
2 dir	
3 lookfor left	
4 edit	
5 list cntry idno lrscale	1
6 sort idno	
7 list cntry idno lrscale	1
8 gen hilfe=invnorm(uniform())	
9 sort hilfe	
10 list cntry idno lrscale	1
11 list cntry idno lrscale trstplt	
12 list cntry idno lrscale trstplt trstplc	1
13 list cntry idno lrscale trstplt trstplc vote	1
14 listtex cntry idno lrscale trstplt trstplc v...	498
15 listtex cntry idno lrscale trstplt trstplc v...	1
16 listtex cntry idno lrscale trstplt trstplc v...	
17 listtex cntry idno lrscale trstplt trstplc v...	
18 listtex cntry idno lrscale trstplt trstplc v...	1
19 listtex cntry idno lrscale trstplt trstplc v...	
20 listtex cntry idno lrscale trstplt trstplc v...	
21 list cntry idno lrscale trstplt	1
22 list cntry idno lrscale trstplt trstplc vote	

```
. list cntry idno lrscale trstplt trstplc vote
```

	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
10.	AT	3276	4	2	2	yes
11.	SE	201748	5	no trust	2	no
12.	SE	101183	4	7	6	yes
13.	AT	1932	7	2	4	yes
14.	AT	824	.	1	3	yes
15.	AT	2009	.	3	8	yes
16.	SE	100122	2	6	2	not elig
17.	SE	202939	8	5	4	yes
18.	AT	2924	5	no trust	9	yes
19.	IT	3048800	5	3	2	yes
20.	AT	807	5	3	8	yes

European Social Survey 2002/2003, Daten für Österreich, Italien, Schweden; 5 463 Beobachtungen (Fälle)

Graphische Darstellung



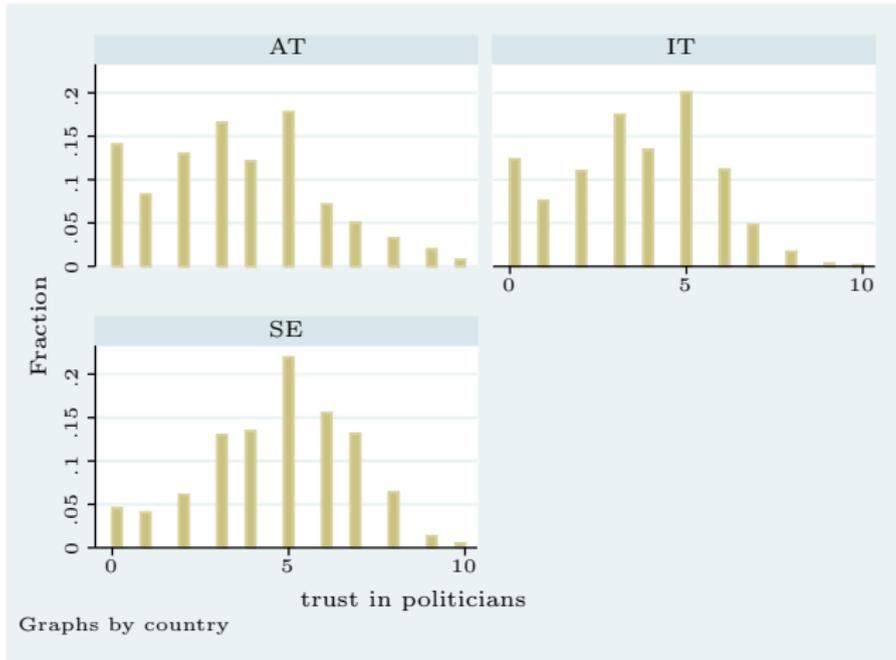
Lage-und Streuungsparameter

```
. summ trustplt,det
```

```
trust in politicians
```

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	obs	5371
25%	2	0	Sum of wgt.	5371
50%	4		Mean	3.962391
		Largest	Std. Dev.	2.307643
75%	6	10	Variance	5.325215
90%	7	10	Skewness	.0049219
95%	8	10	Kurtosis	2.383984
99%	9	10		

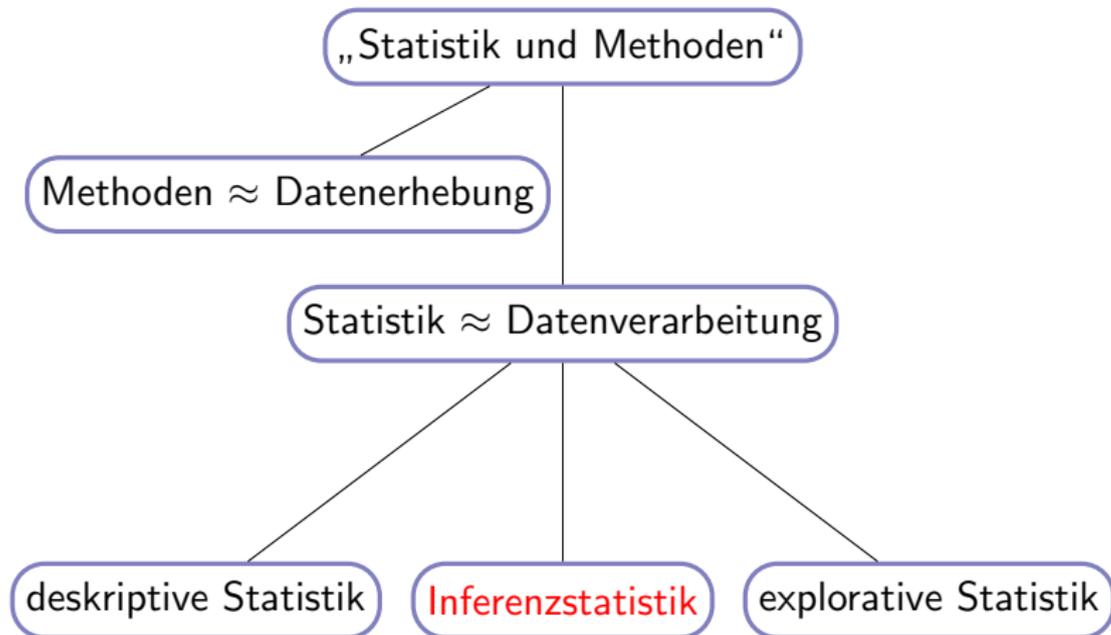
Zwei Dimensionen



Was bedeutet „deskriptive Statistik“?

- ▶ Variable = (numerische) Eigenschaft eines Objektes
- ▶ Deskriptive Statistik = „beschreibende Statistik“
- ▶ Information über eine oder mehrere Variablen wird verdichtet
→ statt mehrerer hundert/tausend Zahlen eine Zahl
- ▶ Zahlen für
 - ▶ Verteilungen (eine Variable)
 - ▶ Zusammenhänge (zwei Variablen)
 - ▶ Modelle (zwei und mehr Variablen)

Was macht Statistik? II



Was bedeutet „Inferenzstatistik“?

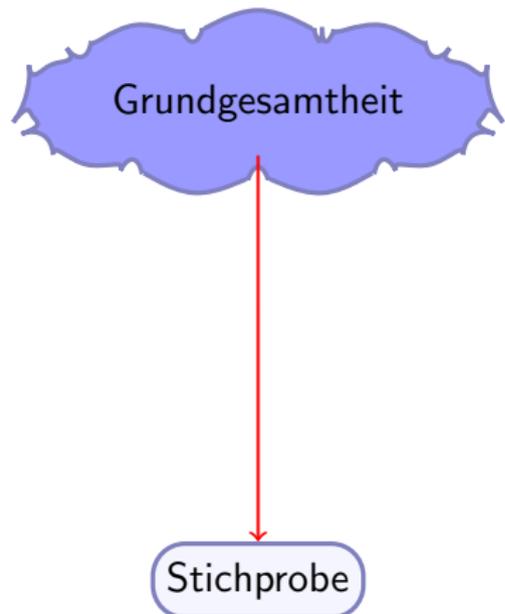
- ▶ Inferenzstatistik = „schließende“ Statistik
- ▶ Häufig können nicht alle relevanten Objekte untersucht werden
 - ▶ 8 Million Österreicher, 60 Millionen Italiener, 9 Millionen Schweden
 - ▶ 2 257 österreichische, 1 207 italienische, 1 999 schwedische Befragte



Stichprobe

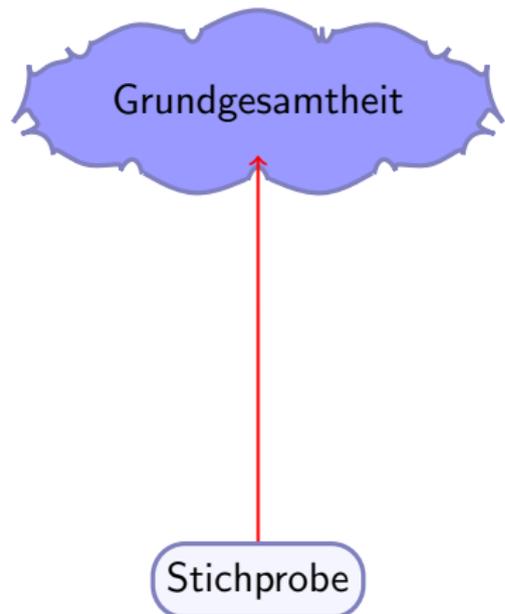
Was bedeutet „Inferenzstatistik“?

- ▶ Inferenzstatistik = „schließende“ Statistik
- ▶ Häufig können nicht alle relevanten Objekte untersucht werden
 - ▶ 8 Million Österreicher, 60 Millionen Italiener, 9 Millionen Schweden
 - ▶ 2 257 österreichische, 1 207 italienische, 1 999 schwedische Befragte
- ▶ Wie bekomme ich „gute“ Stichproben?



Was bedeutet „Inferenzstatistik“?

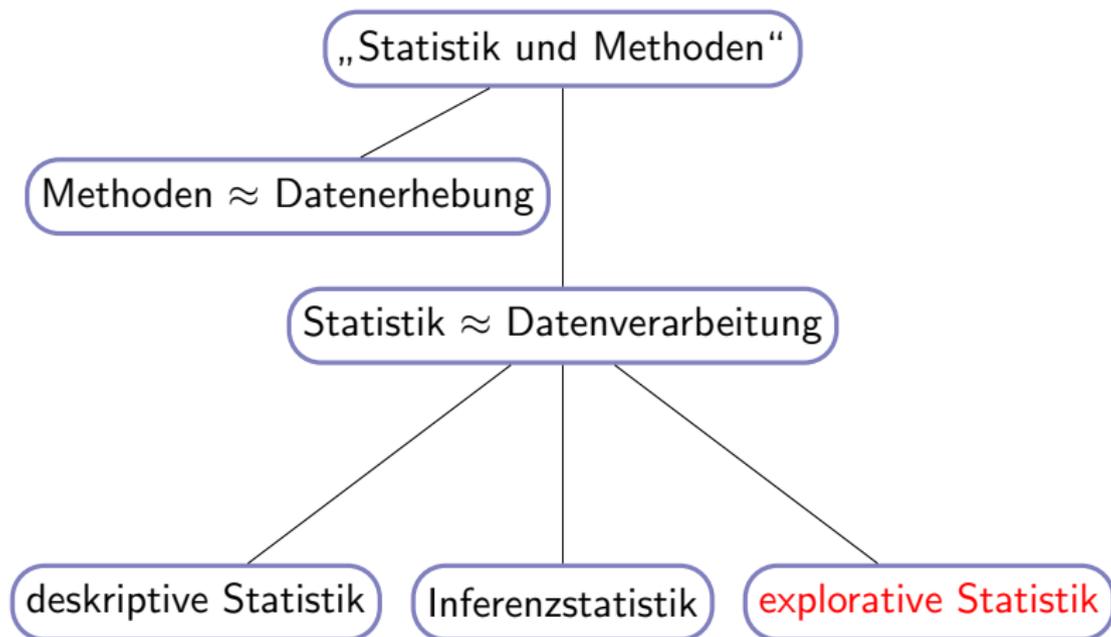
- ▶ Inferenzstatistik = „schließende“ Statistik
- ▶ Häufig können nicht alle relevanten Objekte untersucht werden
 - ▶ 8 Million Österreicher, 60 Millionen Italiener, 9 Millionen Schweden
 - ▶ 2 257 österreichische, 1 207 italienische, 1 999 schwedische Befragte
- ▶ Wie kann ich von der Stichprobe auf die Grundgesamtheit schließen?



Was bedeutet „Inferenzstatistik“? II

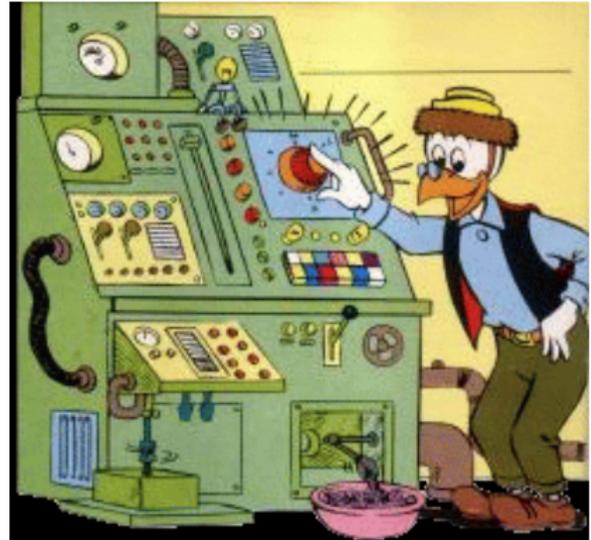
- ▶ (Setzt deskriptive Statistik voraus)
- ▶ Unter welchen Bedingungen ist der Schluß von einer Stichprobe auf eine Grundgesamtheit zulässig?
- ▶ Wie sicher kann ich mir über die Ergebnisse sein bzw.
- ▶ Wie wahrscheinlich ist ein Fehler in einer bestimmten Größenordnung?
- ▶ *Test* von Unterschieds- und Zusammenhangshypothesen, *Konfidenzbänder*
- ▶ **Praktisch alle publizierten Studien nutzen Inferenzstatistik**

Was macht Statistik? II



Was bedeutet „explorative Statistik“?

- ▶ Mittelding zwischen deskriptiver und Inferenzstatistik
- ▶ „Data Mining“, „serendipity“
- ▶ Wichtig für viele Anwendungen in der Wirtschaft
- ▶ Sehr große Datenmengen
- ▶ In den Sozialwissenschaften noch wenig bekannt/verbreitet



Warum soll ich das lernen?

Public Policy

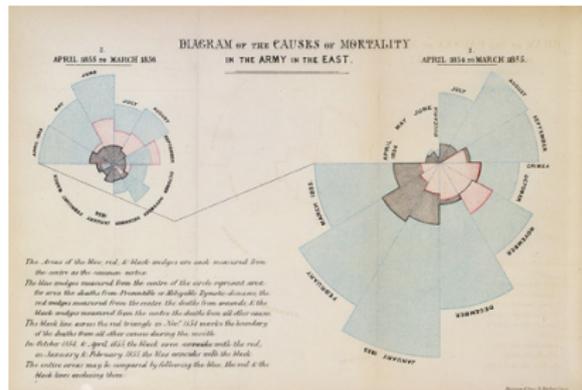
„To understand God's thoughts we must study statistics, for these are the measure of His purpose.“ (Florence Nightingale)

Warum soll ich das lernen?

Public Policy

„To understand God's thoughts we must study statistics, for these are the measure of His purpose.“ (Florence Nightingale)

- ▶ Krankenschwester während des Krim-Krieges & Begründerin der modernen Pflegewissenschaft
- ▶ Bewies Zusammenhang zwischen Hygiene und Überlebensraten
- ▶ Stellte ihre Ergebnisse so dar, daß Politiker sie verstehen konnten → Reformen



Warum soll ich das lernen? II

Politische Vierteljahresschrift 3/2008

1. Zwischen Macht und Gerechtigkeit
2. Regieren jenseits des Nationalstaates und der Mythos einer 80-Prozent-Europäisierung
3. Wirtschaftspolitischer Handlungsspielraum im Zeitalter der Globalisierung.
4. Rechtsextremistische Wahlabsicht und regionaler Kontext
5. Die Messbarkeit von Demokratiequalität als ungelöstes Theorieproblem
6. Forschung zu Protest und sozialen Bewegungen

Warum soll ich das lernen? II

Politische Vierteljahresschrift 3/2008

1. **Zwischen Macht und Gerechtigkeit**
2. Regieren jenseits des Nationalstaates und der Mythos einer 80-Prozent-Europäisierung
3. Wirtschaftspolitischer Handlungsspielraum im Zeitalter der Globalisierung.
4. Rechtsextremistische Wahlabsicht und regionaler Kontext
5. Die Messbarkeit von Demokratiequalität als ungelöstes Theorieproblem
6. Forschung zu Protest und sozialen Bewegungen

- ▶ Nur einer von sechs Hauptartikeln kommt *ohne* Statistik aus

Warum soll ich das lernen? II

Politische Vierteljahresschrift 3/2008

1. Zwischen Macht und Gerechtigkeit
2. Regieren jenseits des Nationalstaates und der Mythos einer 80-Prozent-Europäisierung
3. Wirtschaftspolitischer Handlungsspielraum im Zeitalter der Globalisierung.
4. Rechtsextremistische Wahlabsicht und regionaler Kontext
5. Die Messbarkeit von Demokratiequalität als ungelöstes Theorieproblem
6. Forschung zu Protest und sozialen Bewegungen

- ▶ Nur einer von sechs Hauptartikeln kommt *ohne* Statistik aus
- ▶ Und auch diese Feststellung ist Statistik

Warum soll ich das lernen? IIIa

Durchschnittswerte

The „average“ salary is £ 24 000, but most people earn less. Most people have an „above average“ number of feet. In six out of seven of the countries in the G7, productivity is „below average“.

Quelle: <http://news.bbc.co.uk/1/hi/magazine/7581120.stm>

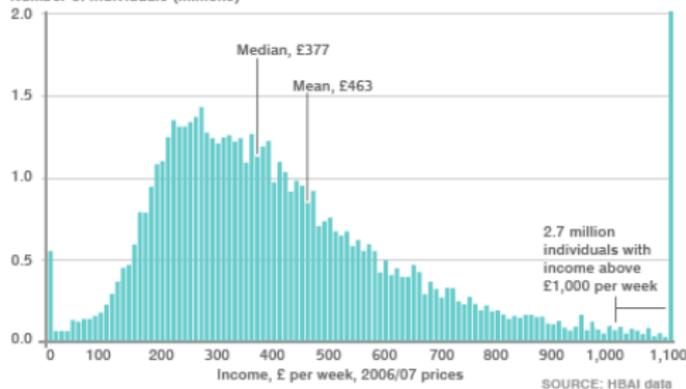
Warum soll ich das lernen? IIIa

Durchschnittswerte

The „average“ salary is £ 24 000, but most people earn less. Most people have an „above average“ number of feet. In six out of seven of the countries in the G7, productivity is „below average“.

THE UK INCOME DISTRIBUTION IN 2006 / 7

Number of individuals (millions)



Quelle: <http://news.bbc.co.uk/1/hi/magazine/7581120.stm>

Warum soll ich das lernen? IIIa

Durchschnittswerte

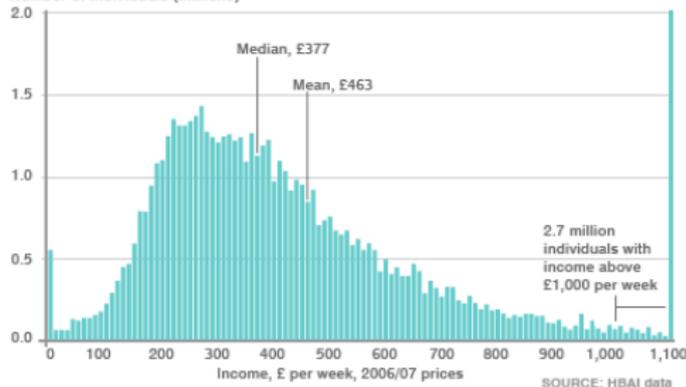
The „average“ salary is £ 24 000, but most people earn less. Most people have an „above average“ number of feet. In six out of seven of the countries in the G7, productivity is „below average“.

Problem: „average“ = „normal“?

Quelle: <http://news.bbc.co.uk/1/hi/magazine/7581120.stm>

THE UK INCOME DISTRIBUTION IN 2006 / 7

Number of individuals (millions)



Warum soll ich das lernen? IIIb

Prozentwerte

Vitamin E can kill. Supplements cause a 14% increase in mortality, said the man on telly, standing in front of a huge „14%“, propped up on a number like the bar of the local.

Quelle: <http://news.bbc.co.uk/1/hi/magazine/7568929.stm>

Warum soll ich das lernen? IIIb

Prozentwerte

Vitamin E can kill. Supplements cause a 14% increase in mortality, said the man on telly, standing in front of a huge „14%“, propped up on a number like the bar of the local.

Funny, but wasn't my risk of mortality quite high already? Like, 100%? But no, it's even worse than that. Vitamin E supplements apparently make the end more than certain.

Quelle: <http://news.bbc.co.uk/1/hi/magazine/7568929.stm>

Warum soll ich das lernen? IIIb

Prozentwerte

Vitamin E can kill. Supplements cause a 14% increase in mortality, said the man on telly, standing in front of a huge „14%“, propped up on a number like the bar of the local.

Funny, but wasn't my risk of mortality quite high already? Like, 100%? But no, it's even worse than that. Vitamin E supplements apparently make the end more than certain.

Quelle: <http://news.bbc.co.uk/1/hi/magazine/7568929.stm>

- ▶ Was ist die Ausgangsbasis?
- ▶ Mortalität \approx Wahrscheinlichkeit, innerhalb eines *Zeitraumes* zu sterben (z. B. ein, fünf, zehn Jahre)
- ▶ 3% \rightarrow 3,42%

Warum soll ich das lernen? IIIc

Stichproben

Health crisis as three million people went down last winter with Norovirus – winter vomiting disease. The flaw: The authorities don't know how many had it. They don't even know how many went to the doctor with it. All they know is the 2000 cases in the peak months that were confirmed in the lab.

They say that for every lab-confirmed case there are about 1500 in the community, a ratio of 1:1500. They haven't a clue.

Quelle: <http://news.bbc.co.uk/1/hi/magazine/7605118.stm>

- ▶ Konfidenzintervall:
1:140 oder
1:17 000
- ▶ 34 Millionen
oder 280 000?
- ▶ Erklärung für
Stichproben-
größe im Text
ist falsch

Warum soll ich das lernen? IV

Berufsperspektiven

„What did the sociologist who passed statistics say to the sociologist who failed it?”

Warum soll ich das lernen? IV

Berufsperspektiven

„What did the sociologist who passed statistics say to the sociologist who failed it?
„I'll have a Big Mac, fries, and a Coke““ (Schule des Lebens)

Warum soll ich das lernen? IV

Berufsperspektiven

„What did the sociologist who passed statistics say to the sociologist who failed it?
„I'll have a Big Mac, fries, and a Coke““ (Schule des Lebens)

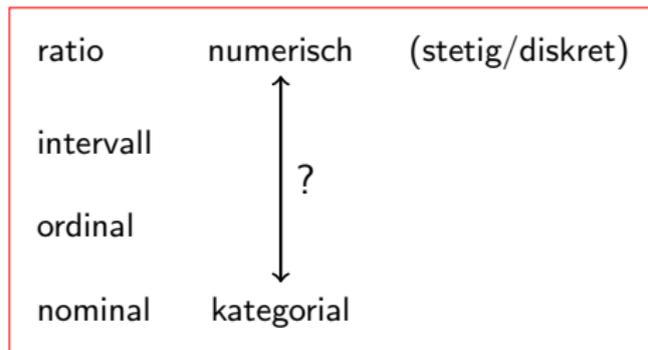
- ▶ Politik, Medien, Verbände etc. brauchen Absolventen
- ▶ Die mit grundlegenden statistischen Konzepten vertraut sind und
- ▶ Entsprechende Inhalte verstehen, bewerten, kommunizieren können
- ▶ Sozialwissenschaftler \neq Statistiker

Was sind Variablen?

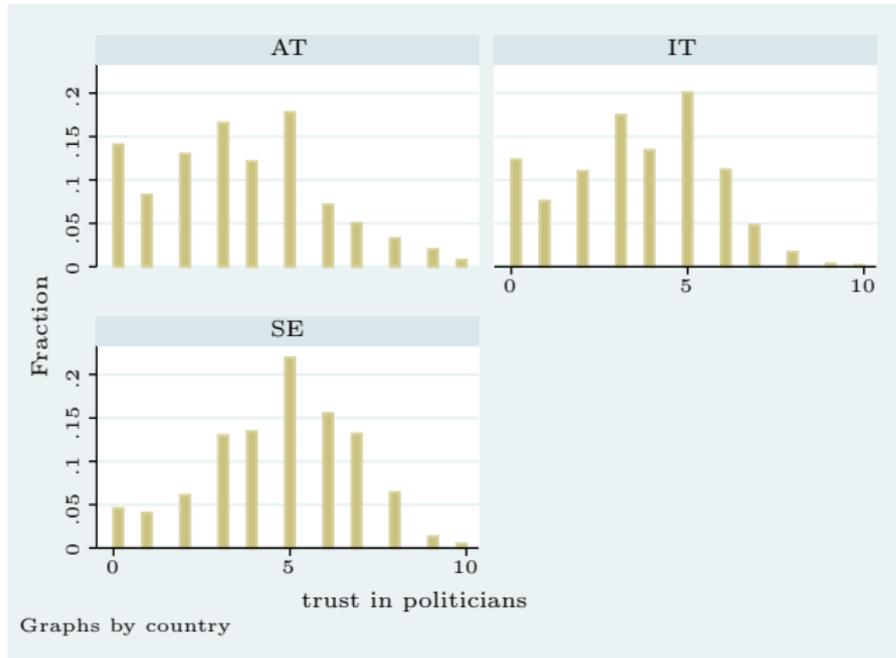
- ▶ Quantifizierbare Eigenschaften
- ▶ Von politikwissenschaftlich relevanten Objekten
- ▶ Normalerweise mehrere Variablen pro Objekt
- ▶ Abkürzung in Formeln mit mysteriösen Kleinbuchstaben ($x, y, z \dots$)
- ▶ Skalenniveau \approx Informationsgehalt

Was sind Variablen?

- ▶ Quantifizierbare Eigenschaften
- ▶ Von politikwissenschaftlich relevanten Objekten
- ▶ Normalerweise mehrere Variablen pro Objekt
- ▶ Abkürzung in Formeln mit mysteriösen Kleinbuchstaben ($x, y, z \dots$)
- ▶ Skalenniveau \approx Informationsgehalt



Zurück zum Beispiel

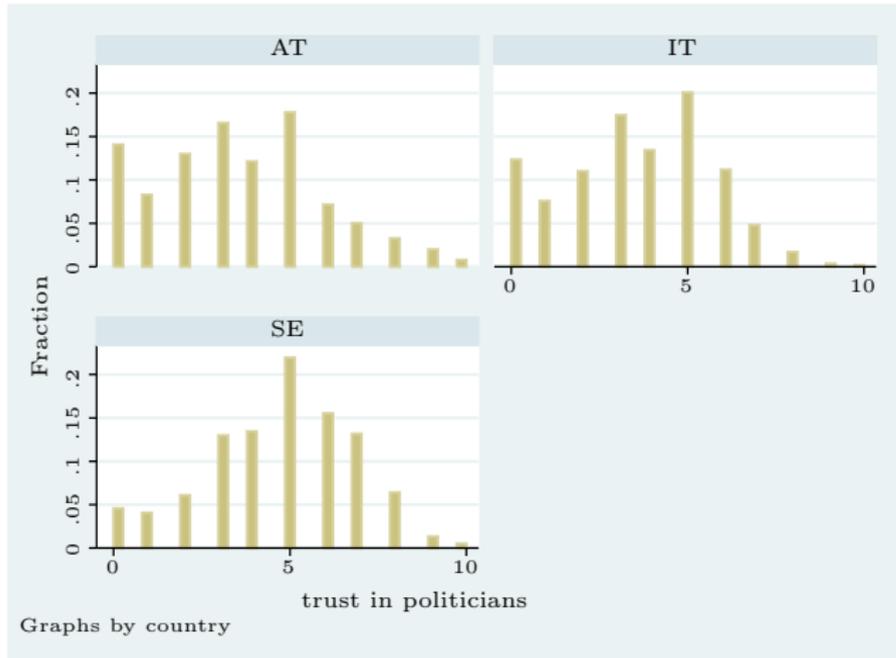


- ▶ Was sind die Objekte?
- ▶ Skalenniveau? Kontinuierlich oder kategorial?
- ▶ Deskriptive oder Inferenzstatistik?

Stichprobe vs Grundgesamtheit

- ▶ Grundgesamtheit/population \equiv Menge aller Objekte, über die etwas ausgesagt werden soll
- ▶ Stichprobe/sample \equiv Menge aller Objekte, die tatsächlich untersucht werden

Zurück zum Beispiel



- ▶ Was ist/sind die Grundgesamtheiten?
- ▶ Was ist/sind die Stichproben?

Was ist ein (statistisches) Modell?

- ▶ *Keine* 1:1-Entsprechung mit Wirklichkeit
- ▶ Reduktion der Kernaussagen einer Theorie auf Formel(n)
- ▶ Die Zusammenhänge zwischen Variablen beschreiben

Was ist ein (statistisches) Modell?

- ▶ *Keine* 1:1-Entsprechung mit Wirklichkeit
- ▶ Reduktion der Kernaussagen einer Theorie auf Formel(n)
- ▶ Die Zusammenhänge zwischen Variablen beschreiben

Churchill on Socialism

„Any man who is not a socialist before he is 40 has no heart. Any man who is still a socialist after he is 40 has no head.“

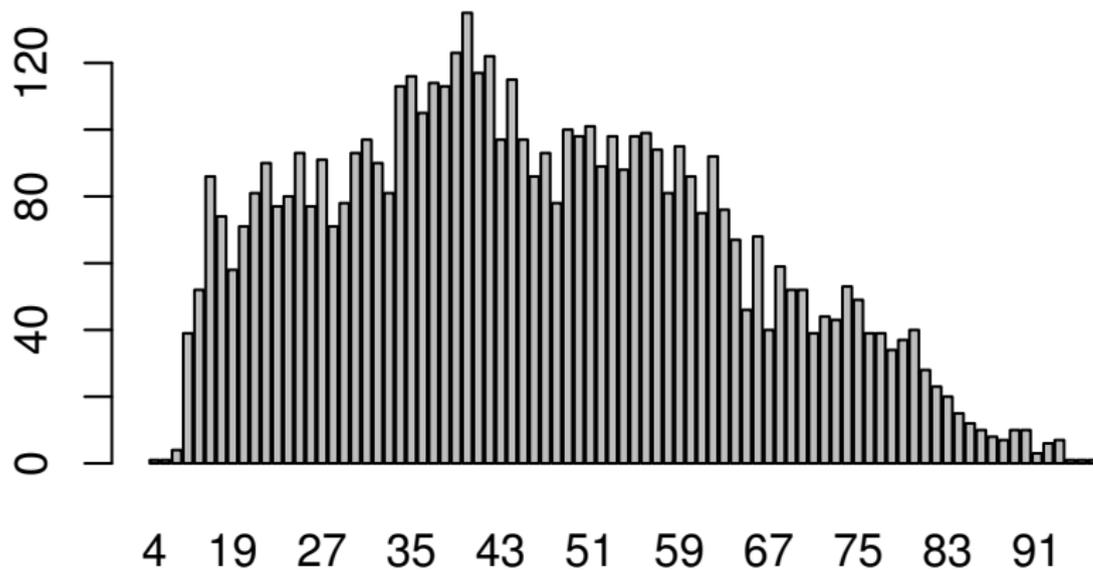
Was ist ein (statistisches) Modell?

- ▶ *Keine* 1:1-Entsprechung mit Wirklichkeit
- ▶ Reduktion der Kernaussagen einer Theorie auf Formel(n)
- ▶ Die Zusammenhänge zwischen Variablen beschreiben
- ▶ Ideologie =
Konstante + Faktor \times Alter +
zufällige Einflüsse

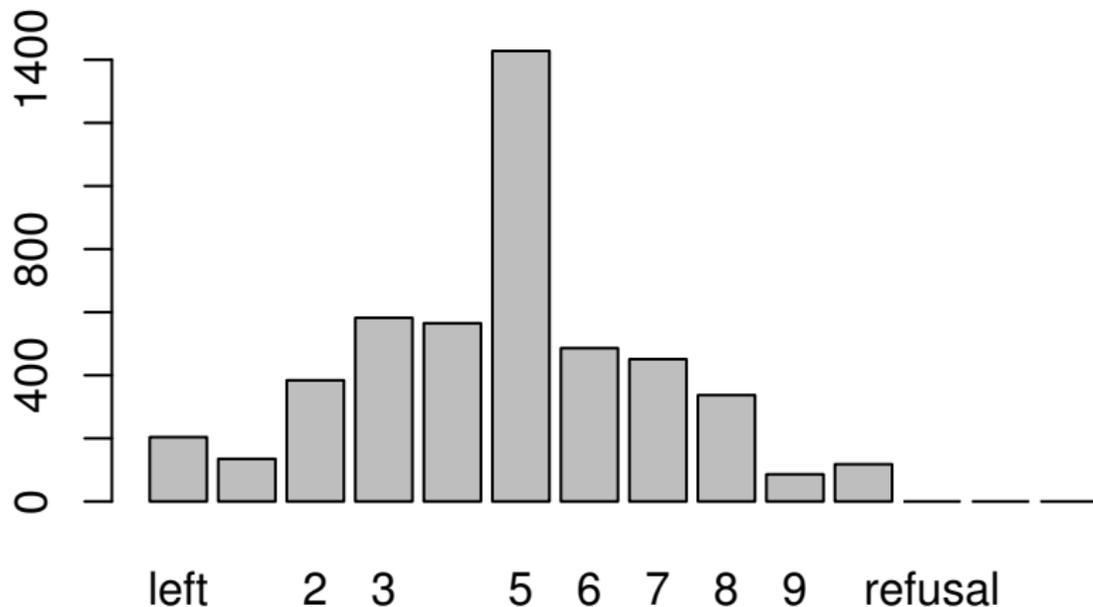
Churchill on Socialism

„Any man who is not a socialist before he is 40 has no heart. Any man who is still a socialist after he is 40 has no head.“

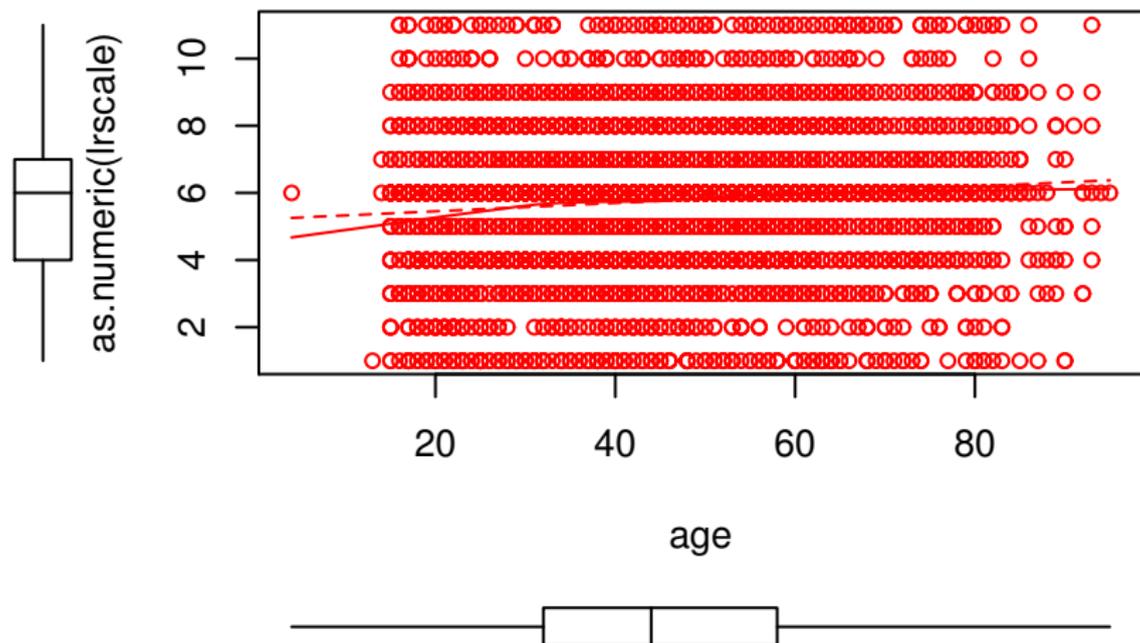
Verteilung: Alter



Verteilung: Links-Rechts



Modell Links-Rechts – Alter



$$\text{Links-Rechts} = 5,2 + 0,01 \times \text{Alter}$$

Modell Links-Rechts – Alter II

- ▶ Minimaler Zusammenhang – für jedes Lebensjahr nimmt Links-Rechts-Wert um $\frac{1}{100}$ Skaleneinheit zu
- ▶ Miserabler Fit: zufällige Einflüsse sehr viel wichtiger als systematische Komponente (Alter)
- ▶ (Schwache) Hinweise auf nicht-linearen Zusammenhang (was heißt das?)
- ▶ Alles in allem keine plausible Erklärung der empirischen Verteilung
- ▶ Miserable Theorie für drei westeuropäische Länder

Zusammenfassung

- ▶ „Statistik“ ist ein mächtiges Hilfsmittel, um numerische Daten darzustellen und auszuwerten
- ▶ Statistische Modelle fassen die Hauptaussagen politikwissenschaftlicher Theorien in mathematischer Form zusammen
- ▶ Auf diese Weise können Theorien mit der Wirklichkeit konfrontiert werden

Zusammenfassung

- ▶ „Statistik“ ist ein mächtiges Hilfsmittel, um numerische Daten darzustellen und auszuwerten
- ▶ Statistische Modelle fassen die Hauptaussagen politikwissenschaftlicher Theorien in mathematischer Form zusammen
- ▶ Auf diese Weise können Theorien mit der Wirklichkeit konfrontiert werden
- ▶ Durchschnittlich intelligente Studierende können die Grundlagen innerhalb eines Semesters erlernen und sogar Spaß daran haben