Statistik I - Sitzung 3

Bernd Schlipphak

Institut für Politikwissenschaft

Woche 3

1. Univariate Darstellung
   - Häufigkeitsverteilungen
   - Grafische Darstellungen
   - Gruppierung

Zusammenfassung Skalierungsniveaus

Merksatz Skalierungsniveau

Jedes Skalenniveau weist neben seinen spezifischen Eigenschaften auch alle Eigenschaften der niedrigeren Skalenniveaus auf $\Rightarrow$ metrisch skalierte Variablen können auch als nominal oder ordinal skaliert behandelt werden, aber nicht umgekehrt!

Zusammenfassung Skalierungsniveaus

- Wozu brauchen wir dieses Wissen über die Skalenniveaus von Variablen(-ausprägungen)?
- Sie spielen eine wichtige Rolle dafür, für welche Variablen überhaupt bestimmte Maße berechnet und sinnvoll interpretiert werden können
- Das beginnt bereits bei der univariaten Darstellung!

Darstellung univariater Verteilungen

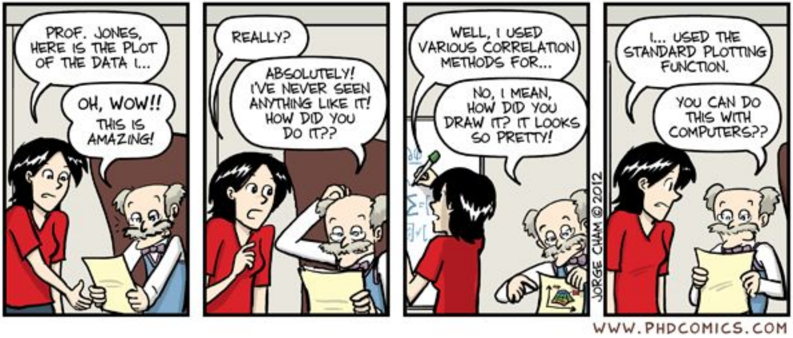


Abbildung: Comic von www.phdcomics.com

Darstellung univariater Verteilungen

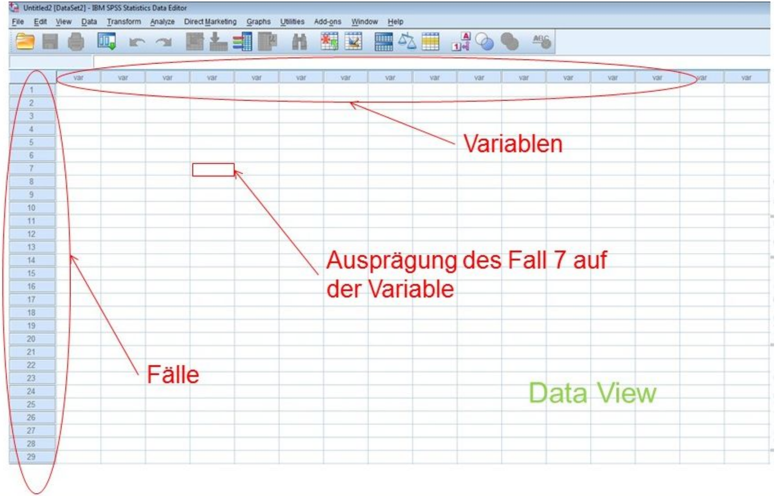
Merksatz Univariate Darstellung

Die univariate Darstellung ist die beschreibende Darstellung einer einzigen *Variablen* im Hinblick auf die Verteilung ihrer *Ausprägungen* über die untersuchten *Fälle* hinweg!

Die Urliste

- **Urliste** = Datenmenge, welche die Ausprägungen einer Variablen für eine Menge von Objekten enthält (auch: = Rohdaten, Datensatz)
- In der Urliste wird einem Objekt (= **Fall**) der Wert (= **Ausprägung**) einer **Variablen** zugeordnet
- Die Urliste besteht also aus einer Tabelle, in welcher die Zeilen Fällen (= Objekten) entsprechen, die Spalten Variablen und die einzelnen Zellen den Ausprägungen der Fälle auf den Variablen

Die Urliste


Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Variablen
Ausprägung des Fall 7 auf der Variable
Fälle
Data View

Die Urliste

*DATA_Q71G_HDI_COMBINED_FINAL.sav [DataSet1] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Direktmarketing Grafik Extras Erweiterungen Fenster Hilfe

	PSRAID	COUNTRY	INTERNET	Q67	COMPUTER	Q148E	SMRTPHN	Q69	WOHNORT	EINKOMMEN_Z_H	BILDUNG
1	143734	2	1	1	0	2	0	2	0	0	8
2	143774	2	1	1	0	2	0	2	1	0	8
3	143776	2	1	1	0	2	0	2	1	0	12
4	143784	2	1	1	1	1	0	2	1	0	12
5	143791	2	1	1	0	2	0	2	0	0	5
6	143794	2	1	1	0	2	0	2	0	0	5
7	143820	2	1	1	0	2	0	2	0	0	15
8	143869	2	1	1	0	2	1	1	1	0	6
9	143872	2	1	1	0	2	1	1	1	0	10
10	143873	2	1	1	0	2	1	1	1	0	5
11	143879	2	1	1	0	2	1	1	1	0	10
12	143880	2	1	1	0	2	1	1	1	0	10
13	143892	2	1	1	0	2	.	.	1	0	9
14	143911	2	0	2	0	2	1	1	1	0	3
15	143925	2	1	1	0	2	0	2	1	0	12
16	143926	2	1	1	1	1	1	1	1	0	15
17	143934	2	1	1	0	2	0	2	1	0	10
18	143943	2	1	1	0	2	0	2	0	0	10
19	143946	2	1	1	0	2	0	2	0	0	12
20	143948	2	1	1	0	2	0	2	0	0	10
21	143952	2	1	1	0	2	0	2	1	0	12
22	143967	2	1	1	0	2	1	1	1	0	0
23	143975	2	1	1	1	1	1	1	1	0	0
24	143978	2	1	1	0	2	1	1	1	0	0
25	143988	2	1	1	0	2	0	2	1	0	10

Häufigkeiten

- Absolute Häufigkeit einer Variablenausprägung $= f_j$
  - Anzahl der Fälle mit dieser Ausprägung
- Relative Häufigkeit einer Variablenausprägung $= p_j$
  - Anteil der Fälle mit dieser Ausprägung an allen Fällen
  - $p_j = f_j/n$

Häufigkeitsverteilungen

- Absolute Häufigkeitsverteilung
  - Darstellung der absoluten Häufigkeiten aller Ausprägungen
- Relative Häufigkeitsverteilung
  - Darstellung der relativen Häufigkeiten aller Ausprägungen

Häufigkeitsverteilungen

- Kumulierte Häufigkeitsverteilung
  - gibt für jede Ausprägung der Variable die Summe (absolut) / den Anteil (relativ) der Fälle an, die diese oder eine 'niedrigere' Ausprägung besitzen
  - daher erst ab Ordinalskalenniveau einsetzbar!

Häufigkeitsverteilungen - Beispiel

- Verteilung von Schulnoten unter 30 Schülern
  - N (= Anzahl der Fälle) = 30
  - 6 SchülerInnen haben die Note *Sehr gut*,
  - 3 SchülerInnen haben die Note *Gut*,
  - 9 SchülerInnen haben die Note *Befriedigend*,
  - 9 SchülerInnen haben die Note *Ausreichend* und
  - 3 Schüler sind durchgefallen *(Mangelhaft)*

	Sehr gut	**Gut**	**Befriedigend**	**Ausreichend**	**Mangelhaft**
Anzahl SchülerInnen	6	3	9	9	3

Häufigkeitsverteilungen

Schulnoten	Absolute Häufigkeit	Relative Häufigkeit	Relative Häufigkeit(%)	Kumulierte Häufigkeit (%)
Sehr gut	6	$0.2\ (= \frac{f_j}{n} = \frac{6}{30})$	20	20
Gut	3	0.1	10	30
Befriedigend	9	0.3	30	60
Ausreichend	9	0.3	30	90
Mangelhaft	3	0.1	10	100
$\sum$	30(=n)	1	100	

Häufigkeitsverteilungen

Schulnoten	f_j (= Abs. Häufigkeit)	p_j (= Rel. Häufigkeit)	$p_j * 100$	$p_j cum * 100$
Sehr gut	6	0.2 $(= \frac{f_j}{n} = \frac{6}{30})$	20	20
Gut	3	0.1	10	30
Befriedigend	9	0.3	30	60
Ausreichend	9	0.3	30	90
Mangelhaft	3	0.1	10	100
$\sum$	30(=n)	1	100	

Häufigkeitsverteilungen

QA24B DEMOCRACY SATISFACTION - EUROPEAN UNION

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Very satisfied	1673	5,4	5,5	5,5
	Fairly satisfied	13420	43,7	44,4	50,0
	Not very satisfied	8239	26,8	27,3	77,2
	Not at all satisfied	2292	7,5	7,6	84,8
	DK	4591	14,9	15,2	100,0
	Gesamt	30215	98,4	100,0	
Fehlend	INAP - TCC	500	1,6		
Gesamt		30715	100,0		

Abbildung: Darstellung der EU-Demokratiezufriedenheit mit Eurobarometer-Daten

Grafische Darstellungen von Verteilungen

- Stabdiagramm
- Säulendiagramm
- Balkendiagramm
- Histogramm (für metrisch skalierte Variablen)
- Kreisdiagramm (= Kuchen-/Tortendiagramm)

Beispiel Kreisdiagramm

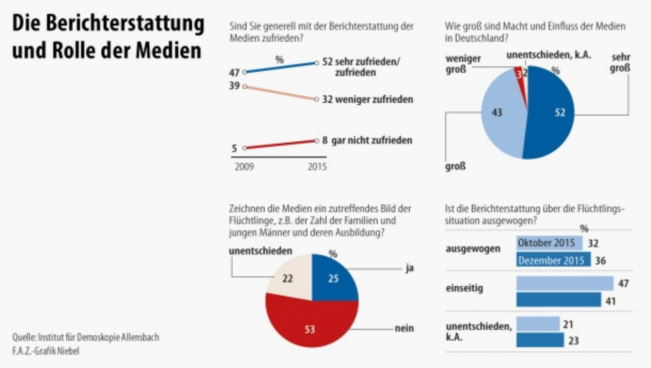


Abbildung: Darstellung Allensbach-Ergebnisse aus der FAZ (http://tinyurl.com/j6bc3oz)

Beispiel Kreisdiagramm

Merksatz Kreisdiagramm

Das Kreisdiagramm erschwert eher die Unterscheidung von Häufigkeiten. Es wird daher im wissenschaftlichen Bereich nahezu nie verwendet!

Beispiel Balken-/Stab-/Säulendiagramm

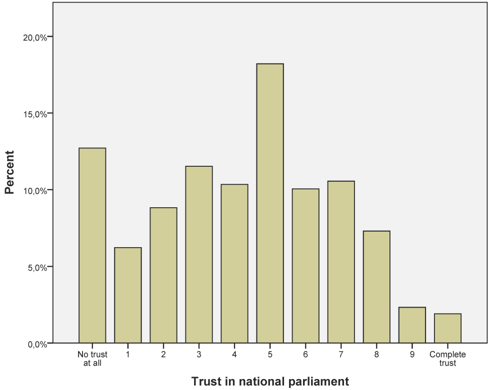


Abbildung: Darstellung des Vertrauens in Nat. Parlament mit ESS-Daten

Beispiel Histogramm

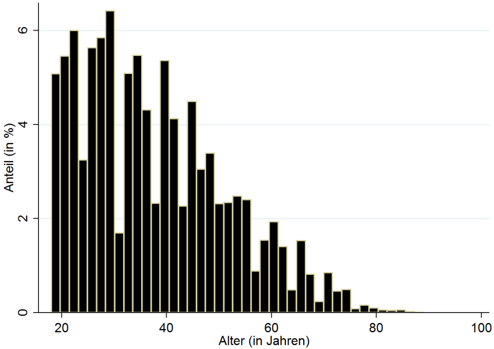
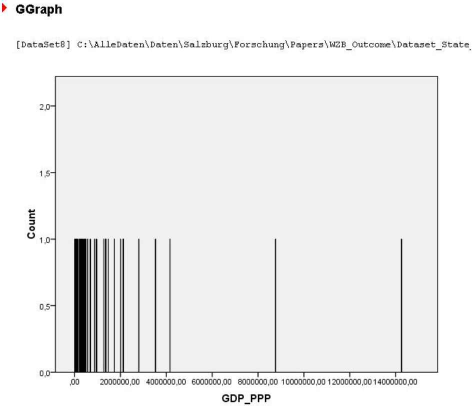


Abbildung: Darstellung der Altersverteilung im Arab Barometer (3. Welle)

Beispiel Histogramm - Probleme

- Bei metrischen und insbesondere bei absolut skalierten Variablen können manchmal alle Werte nur jeweils für einen Fall vorkommen ⇒ das führt bei der Darstellung zu absurden Grafiken!



Beispiel Histogramm - Gruppierung

- Bei metrischen und insbesondere bei absolut skalierten Variablen können manchmal alle Werte nur für jeweils einen Fall vorkommen $\Rightarrow$ das führt bei der Darstellung zu absurden Grafiken!
- Lösung hierfür: **Gruppierung von Fällen** in übersichtliche Gruppen!

Gruppierung von Fällen

- Durch eine Gruppierung 'gruppiert' man Fälle, d.h., man fasst Fälle mit ähnlichen Werten in einer größeren Gruppe zusammen
- Beispiel Alter:
  - Alte Werte = 18, 19, . . . , 35 (Jahre) $\Rightarrow$ Neuer Wert = 1 (jung)
  - Alte Werte = 36, 37, 38, . . . , 60 (Jahre) $\Rightarrow$ Neuer Wert = 2 (mittel)
  - Alte Werte = 61, 62, . . . , k (Jahre) $\Rightarrow$ Neuer Wert = 3 (alt)

Gruppierung von Fällen

- Vorteil: Anteile der drei Gruppen junger, mittlerer und alter Menschen bzw. deren Verteilung leicht in Balkendiagramm überschaubar

Merksatz Gruppierung I

Man verliert durch eine Gruppierung immer an Information. Daher geht mit einer Gruppierung auch meist ein Abstieg der Skalenniveaus einher – im Falle des Alters verändert sich das Skalenniveau vom metrischen auf das ordinale Niveau!

Merksatz Gruppierung II

Theoretische Begründung der Gruppenbildung ist notwendig! Warum werden bestimmte Werte in eine Gruppe zusammen gelegt und nicht in andere?