JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

2. Einheit

Plan für heute

- **Univariate Statistik 1:**
  - Häufigkeiten
  - Datenmatrix und Notation

Lernziele

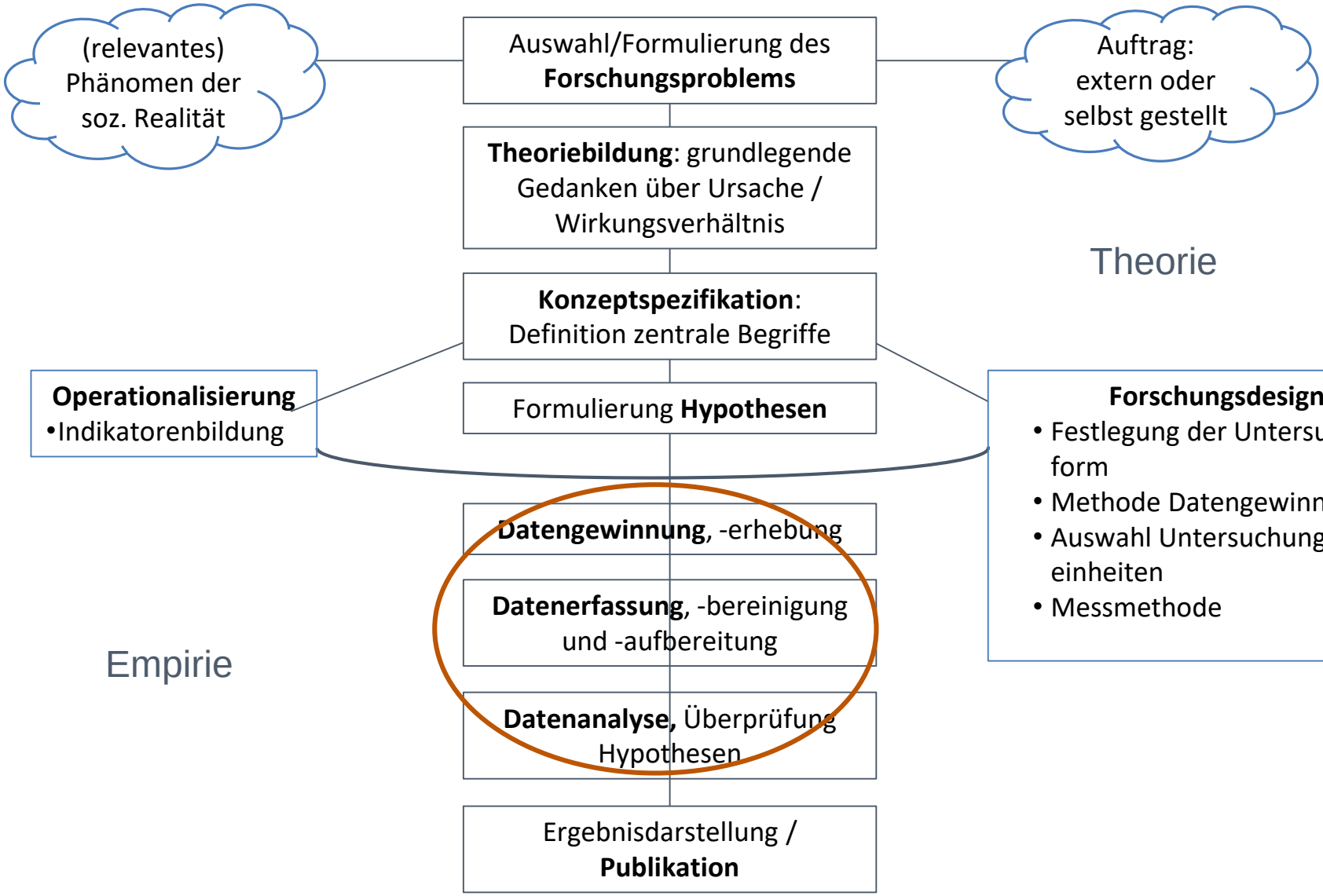
- Verständnis des Aufbaus einer Datenmatrix
- Grundlegende Kenntnis tabellarischer und grafischer Darstellungsformen von univariaten statistischen Informationen
- Kenntnis des Summenzeichens

Vom Fragebogen zur Datenmatrix

Datenerhebung:

- Z.B. wurden Individuen nach ihren Einstellungen gefragt. In unserem Beispiel liegen die Daten somit als Antworten zu den Fragen des Fragebogens vor

Forschungsprozess, „klassisch“



Vom Fragebogen zur Datenmatrix

Datenerhebung:

- Z.B. wurden Individuen nach ihren Einstellungen gefragt. Die Daten liegen somit als Antworten zu den Fragen des Fragebogens vor

Auflistung und Erfassung der Daten → Urliste /Datenmatrix

- **Urliste**:
  - Entsteht, wenn nacheinander für jede Beobachtungseinheit notiert wird, welchen Wert sie bei einer Variablen aufweist → **Rohdaten**
  - Häufig wird der Begriff Urliste auch mit der **Datenmatrix** gleichgesetzt.
  - In der Urliste sind dann die jeweiligen Merkmale eingetragen und die zugehörigen Ausprägungen Zeile für Zeile festgehalten

Vom Fragebogen zur Datenmatrix

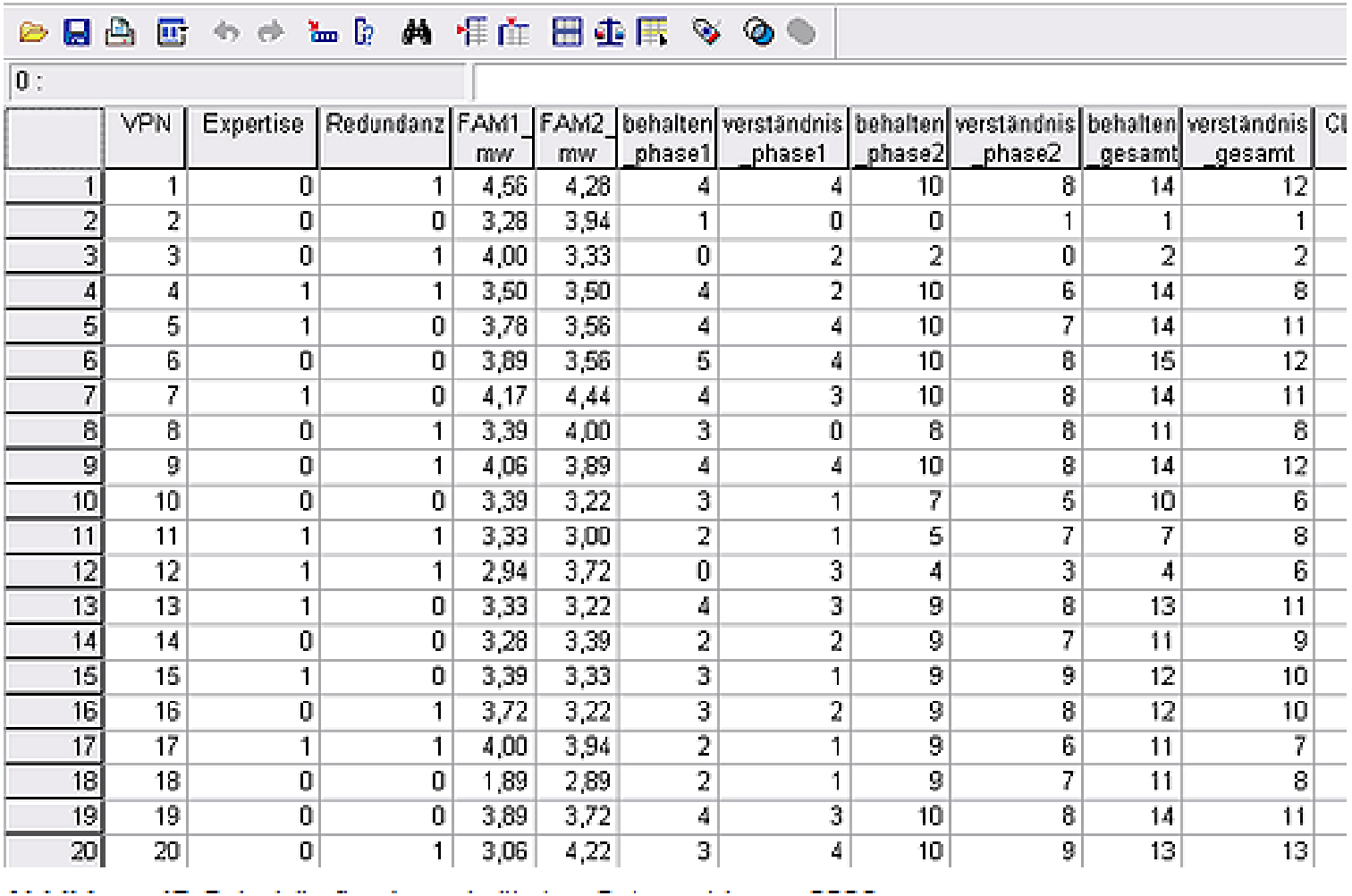
Datenerfassung in Datenmatrix (Excel, Statistikprogramm)

- Die Daten aus allen Fragebogen werden in Form einer **Datenmatrix** aufbereitet:
  - Für jedes erfragte **Merkmal** wird dazu eine **Spalte** benutzt und das Merkmal mit einer Kurzbezeichnung (z.B. V1) charakterisiert
  - Merkmalsausprägungen werden als Zahlenwerte erfasst (kodiert)
  - Die **Antworten** der Befragten werden in je einer **Zeile** festgehalten; die Rohdaten einer Datenerhebung werden kodiert, d.h. die Ausprägungen der Merkmale werden als Zahlen dargestellt
  - Index: Zur Kennzeichnung der Zeilen wird dem Datensatz eine Indexspalte vorangestellt (laufende Nummer)

Datenmatrix, Beispiel

V1	V2 (Geschlecht)	V2 (Alter)	V3 (Schulabschluss)	V4 (Lebens-zufriedenheit)
001	1	24	1	2
002	2	34	3	6
003	2	45	2	1
004	1	67	2	7

Beispiel Datenmatrix (Excel, SPSS)



0 :

	VPN	Expertise	Redundanz	FAM1_mw	FAM2_mw	behalten_phase1	verständnis_phase1	behalten_phase2	verständnis_phase2	behalten_gesamt	verständnis_gesamt	Cl
1	1	0	1	4,56	4,28	4	4	10	8	14	12	
2	2	0	0	3,28	3,94	1	0	0	1	1	1	
3	3	0	1	4,00	3,33	0	2	2	0	2	2	
4	4	1	1	3,50	3,50	4	2	10	6	14	8	
5	5	1	0	3,78	3,56	4	4	10	7	14	11	
6	6	0	0	3,89	3,56	5	4	10	8	15	12	
7	7	1	0	4,17	4,44	4	3	10	8	14	11	
8	8	0	1	3,39	4,00	3	0	8	8	11	8	
9	9	0	1	4,06	3,89	4	4	10	8	14	12	
10	10	0	0	3,39	3,22	3	1	7	5	10	6	
11	11	1	1	3,33	3,00	2	1	5	7	7	8	
12	12	1	1	2,94	3,72	0	3	4	3	4	6	
13	13	1	0	3,33	3,22	4	3	9	8	13	11	
14	14	0	0	3,28	3,39	2	2	9	7	11	9	
15	15	1	0	3,39	3,33	3	1	9	9	12	10	
16	16	0	1	3,72	3,22	3	2	9	8	12	10	
17	17	1	1	4,00	3,94	2	1	9	6	11	7	
18	18	0	0	1,89	2,89	2	1	9	7	11	8	
19	19	0	0	3,89	3,72	4	3	10	8	14	11	
20	20	0	1	3,06	4,22	3	4	10	9	13	13	

Datenmatrix – Wie werden Daten erfasst?

Erhobene Variablen und gemessene Ausprägungen werden in einer **Datenmatrix** organisiert: Tabelle aller erhobenen Merkmale für alle Beobachtungseinheiten

	Variable 1	Variable 2
Fall 1	Wert von Fall 1 auf Variable 1	Wert von Fall 1 auf Variable 2
Fall 2	Wert von Fall 2 auf Variable 1	Wert von Fall 2 auf Variable 2
Fall 3	Wert von Fall 3 auf Variable 1	Wert von Fall 3 auf Variable 2
Fall 4	Wert von Fall 4 auf Variable 1	Wert von Fall 4 auf Variable 2

Beispiel mit Übung: Urliste (n=160)

ID	Studiengang	Semesterzahl	Abinote	Studienzufriedenheit
1	BA SocSc (1)	1	2,0	8
2	BA SocSc (1)	3	3,1	7
3	Medizin (3)	4	1,1	7
4	BA SocSc (1)	2	1,7	9
5	BA Psycho (4)	6	1,5	6
6	Jura (2)	5	2,4	8
…	…	…	…	…

- Wie viele Variablen sehen Sie? Wie viele Beobachtungseinheiten hat die vollständige Datenerhebung?

Beispiel mit Übung: Urliste (n=160)

ID	Studiengang	Semesterzahl	Abinote	Studienzufriedenheit
1	BA SocSc	1	2,0	8
2	BA SocSc	3	3,1	7
3	Medizin	4	1,1	7
4	BA SocSc	2	1,7	9
5	BA Psycho	6	1,5	6
6	Jura	5	2,4	8
…	…	…	…	…

- Wie viele Variablen sehen Sie? (5)
- Wie viele Beobachtungseinheiten hat die vollständige Datenerhebung? (n=160)

Von der Datenmatrix zur statistischen Auswertung

- Urliste bzw. Datenmatrix als Basis der statistischen Auswertung
- Aber: Aus Datenmatrix lässt sich nur eingeschränkt erkennen, wie sich die Beobachtungseinheiten auf die verschiedenen Merkmalsausprägungen verteilen
- Soll z.B. nur untersucht werden, wie zufrieden die Studierenden mit dem Studium sind, interessiert nur eine Merkmalsdimension

→ **univariate Auswertung**: **eine** interessierende Variable wird in ihrem Auftreten in unseren Daten betrachtet

Univariate Datenanalyse

- Wie verteilen sich die Befragten auf die Studiengänge?

ID	Studiengang	Semesterzahl	Abinote	Studienzufriedenheit
1	BA SocSc	1	2,0	8
2	BA SocSc	3	3,1	7
3	Medizin	4	1,1	7
4	BA SocSc	2	1,7	9
5	BA Psycho	6	1,5	6
6	Jura	5	2,4	8
…	…	…	…	…

Bivariate Datenanalyse

- Zusammenhänge zwischen 2 Variablen (ab der 6. Einheit geplant), z.B. Hängt die Studienzufriedenheit mit der Semesterzahl zusammen?

ID	Studiengang	Semesterzahl	Abinote	Studienzufriedenheit
1	BA SocSc	1	2,0	8
2	BA SocSc	3	3,1	7
3	Medizin	4	1,1	7
4	BA SocSc	2	1,7	9
5	BA Psycho	6	1,5	6
6	Jura	5	2,4	8
…	…	…	…	…

Univariate Datenanalyse

- Wertet einzelne Variablen aus
- **1. Schritt: Häufigkeitsverteilung der einzelnen Ausprägungen (tabellarisch und grafisch)**
- 2. Schritt: Informationsmenge vieler Beobachtungen auf wenige Kennzahlen verdichten
  - Dabei lassen sich Lage-, Streuungs- und Formmaße unterscheiden (ab nächste Einheit)

Univariate Häufigkeitsverteilung

- Eine univariate **Häufigkeitsverteilung** ist eine Methode zur (statistischen) Beschreibung einer Variablen:
  - wie verteilen sich die Beobachtungseinheiten auf die Merkmalsausprägungen des Merkmals?
  - Oder: Wo „häufen“ sich die Beobachtungseinheiten auf die Merkmalsausprägungen eines Merkmals?
- Univariate Datenanalyse in tabellarischer oder grafischer Form

Anmerkungen zur Notation

- Eine Variable bzw. ein Merkmal wird per Konvention mit einem **X** gekennzeichnet. Beispiel: *Studiengang*
- X kann wiederum verschiedene **Ausprägungen** annehmen, die als $\boldsymbol{x_k}$ gekennzeichnet werden. Beispiel: *BA Social Sciences (1), Jura, …*
- x_k wird als **Laufindex** bezeichnet, er reicht von Ausprägung 1 bis zur maximalen Ausprägung *m*

Häufigkeitstabelle

Häufigkeitstabelle enthält (meist) mindestens die folgenden Angaben:

- **Absolute Häufigkeiten**: Anzahl der Beobachtungseinheiten, bei der die jeweilige Kategorie auftritt, $f x_k \; bzw. H x_k$
- **Relative Häufigkeiten**: Häufigkeit in Bezug zur Gesamtzahl der Fälle n, Anteilswerte $hx_k = \frac{Hx_k}{n} = \frac{fx_k}{n}$
- **Prozentuale Häufigkeiten**: Multiplikation der relativen Häufigkeiten mit 100 ergibt die prozentualen Häufigkeiten einer Ausprägung, $hx_k \times 100$

Beispiel Häufigkeitstabelle

Studiengang x_k	**Absolute Häufigkeit** fx_k *bzw.* Hx_k	**Relative Häufigkeit** $hx_k = \frac{Hx_k}{n} = \frac{fx_k}{n}$	**Prozentuale Häufigkeit** $hx_k \times 100$
BA Social Sciences (1)	80	0,5	50
Jura (2)	10	0,0625	6,25
Medizin (3)	40	0,25	25
BA Psycho (4)	30	0,1875	18,75
Summe	*160*	*1*	*100*

Häufigkeitstabelle

Gruppierte Häufigkeitstabelle:

- Metrische Variablen haben häufig sehr viele Ausprägungen (z.B. Alter, Abiturnote, Einkommen) → Darstellung in einer „normalen“ Häufigkeitstabelle sehr unübersichtlich
- Lösung: gruppierte Häufigkeitstabelle, Merkmale werden in „Klassen“ eingeteilt (z.B. Altersgruppen, Einkommensgruppen)
- Nachteil: Informationsverlust, dafür aber anschaulich

Häufigkeitstabelle

Gruppierte Häufigkeitstabelle:

- Beispiel: Häufigkeiten der Altersgruppen in einer großen Bevölkerungsumfrage
- Klassengrenzen dürfen nicht überlappen
- Klassenbreiten können, müssen aber nicht gleich groß sein
- Klassen sollten lückenlos aufeinander folgen

		Häufigkeit	Prozent
Gültig	15 - 24 years	1902	8,6
	25 - 39 years	4713	21,4
	40 - 54 years	5610	25,5
	55 years and older	9802	44,5
	Gesamtsumme	22027	100,0

Kumulierte relative Häufigkeiten

- Werden meistens ebenfalls in einer Häufigkeitstabelle angegeben
- Zeigen, wie häufig eine bestimmte Ausprägung und alle niedrigeren Ausprägungen eines Merkmals beobachtet wurden (Prozenträngе)
- Bei ordinal skalierten Merkmalen bieten die kumulierten Prozentsätze eine anschauliche Interpretationsmöglichkeit

Kumulierte relative Häufigkeiten

Mögliche Anwendungsfragen:

- „Wie viel Prozent der Befragten sind unter 40 Jahre alt?“
- „Wie viel Prozent der Schüler*innen haben mindestens die Note ‚gut‘ erhalten?“
- „Welcher Anteil der Befragten hat ein Einkommen von weniger als 1500€?“
- „Wie viel Prozent aller Bewerber*innen haben mindestens einen Realschulabschluss erworben?“

Kumulierte relative Häufigkeiten

- Beispiel Altersgruppen in einer Befragung dargestellt in SPSS

		Häufigkeit	Prozent	Gültige Prozent	Kumulative Prozente
Gültig	15 - 24 years	1902	8,6	8,6	8,6
	25 - 39 years	4713	21,4	21,4	30,0
	40 - 54 years	5610	25,5	25,5	55,5
	55 years and older	9802	44,5	44,5	100,0
	Gesamtsumme	22027	100,0	100,0	

Kumulierte relative Häufigkeiten

- Beispiel politisches Interesse

Kategorie	absolute Häufigkeit	relative Häufigkeit	prozentuale Häufigkeit	kumulierte prozentuale Häufigkeit
sehr stark	425	0,122	12,2	12,2
stark	877	0,251	25,1	37,3
mittel	1437	0,412	41,2	78,5
wenig	564	0,162	16,2	94,7
überhaupt nicht	186	0,053	5,3	100,0
Gesamt	3490	1,000	100,0	

Daten: ALLBUS 2016. Eigene Berechnungen

Kumulierte relative Häufigkeiten

- Beispiel politisches Interesse

Kategorie	absolute Häufigkeit	relative Häufigkeit	prozentuale Häufigkeit	kumulierte prozentuale Häufigkeit
sehr stark	425	0,122	12,2	12,2
stark	877	0,251	25,1	37,3
mittel	1437	0,412	41,2	78,5
wenig	564	0,162	16,2	94,7
überhaupt nicht	186	0,053	5,3	100,0
Gesamt	3490	1,000	100,0	

Daten: ALLBUS 2016. Eigene Berechnungen

Interpretation: 37,3% der Befragten sind mind. „stark“ politisch interessiert

Kumulierte relative Häufigkeiten

Berechnung:

- geben an, wie groß der relative Anteil der Fälle kleiner oder gleich der Merkmalsausprägung x_k ist
- Der kumulierte Prozentsatz summiert zeilenweise die prozentuale Häufigkeit der Fälle auf

→ schrittweise Addition (Kumulation) der Prozentsätze der Merkmalsausprägungen

Beispiel: Kumulierte relative Häufigkeiten

Wie hoch ist der prozentuale Anteil der Personen, die **höchstens** 39 Jahre alt sind?

		Häufigkeit	Prozent	Gültige Prozent	Kumulative Prozente
Gültig	15 - 24 years	1902	8,6	8,6	8,6
	25 - 39 years	4713	21,4	21,4	30,0
	40 - 54 years	5610	25,5	25,5	55,5
	55 years and older	9802	44,5	44,5	100,0
	Gesamtsumme	22027	100,0	100,0	

Beispiel: Kumulierte relative Häufigkeiten

Wie hoch ist der prozentuale Anteil der Personen, die **älter als** 39 Jahre sind? (100% - 30% = 70%)

		Häufigkeit	Prozent	Gültige Prozent	Kumulative Prozente
Gültig	15 - 24 years	1902	8,6	8,6	8,6
	25 - 39 years	4713	21,4	21,4	30,0
	40 - 54 years	5610	25,5	25,5	55,5
	55 years and older	9802	44,5	44,5	100,0
	Gesamtsumme	22027	100,0	100,0	

Beispiel: Kumulierte relative Häufigkeiten

Wie hoch ist der prozentuale Anteil der Personen, die **älter** als 24 Jahre, aber **höchstens** 54 Jahre alt sind?

		Häufigkeit	Prozent	Gültige Prozent	Kumulative Prozente
Gültig	15 - 24 years	1902	8,6	8,6	8,6
	25 - 39 years	4713	21,4	21,4	30,0
	40 - 54 years	5610	25,5	25,5	55,5
	55 years and older	9802	44,5	44,5	100,0
	Gesamtsumme	22027	100,0	100,0	

Übung: (Kumulierte) relative Häufigkeiten

1) Bitte ergänzen Sie die Tabelle.

2) Wie viel Prozent aller Schüler*innen hat mindestens die Note „befriedigend" erreicht?

3) Wie hoch ist der prozentuale Anteil derjenigen Schüler*innen, die eine schlechtere Note als „gut“ erreicht haben?

Schulnote	Absolute Häufigkeiten	Relativer Anteil (%)	Kumulierte relative Häufigkeit (%)
„sehr gut“	150		
„gut“	230		
„befriedigend“	400		
„ausreichend“	190		
„mangelhaft“	25		
„ungenügend“	5		
Gesamt	1000		

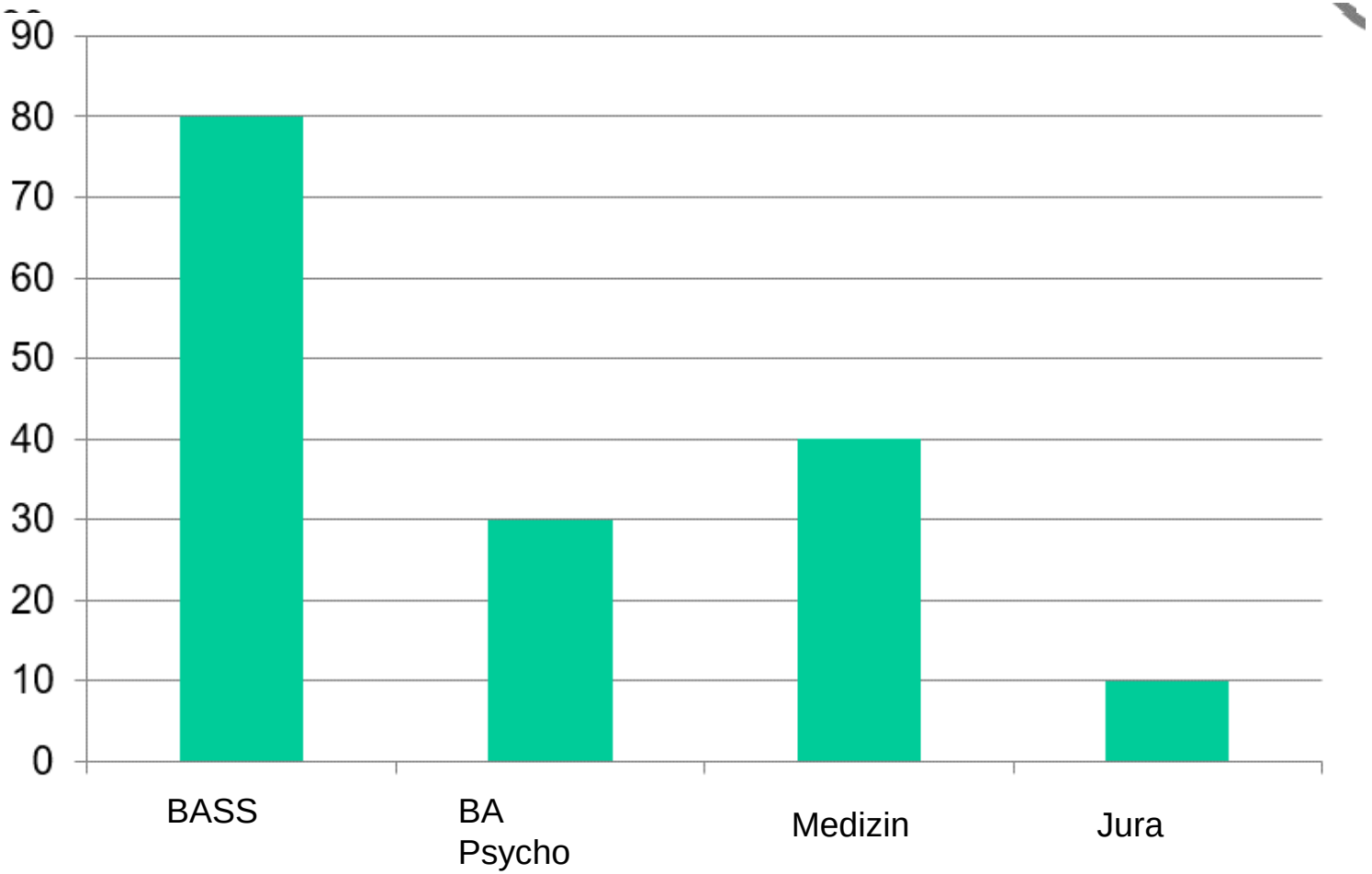
Grafische Darstellungen

- Können die Verteilung von Ausprägungen eines Merkmals auf einen Blick illustrieren
- Unterstützen dabei die tabellarische Darstellung
- Verschiedene Möglichkeiten der univariaten grafischen Darstellung
- Am Häufigsten genutzt: Säulen-/Balkendiagramm, Kreisdiagramm, Histogramm

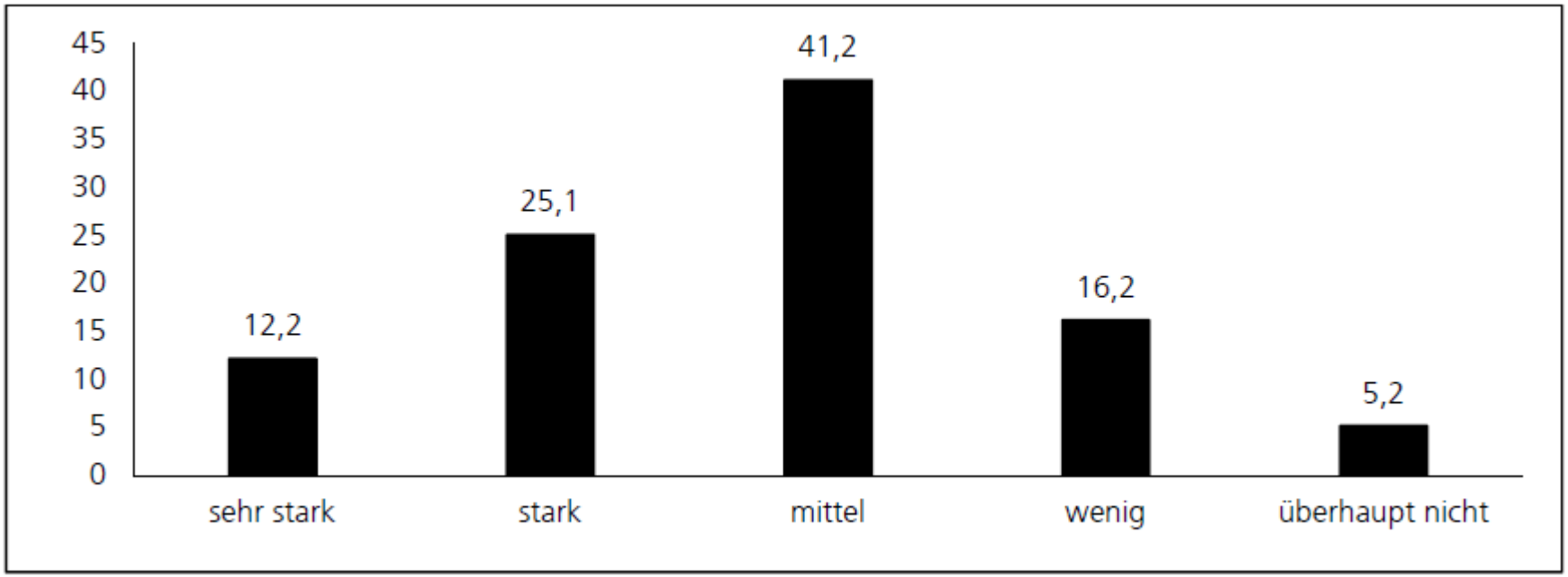
Säulendiagramm

- Die Häufigkeiten der Merkmalsausprägungen werden durch **Säulen** dargestellt (auch Linien- oder Stabdiagramm)
- Höhe der Säulen spiegelt die Anzahl von Beobachtungen (absolute Häufigkeiten) oder den prozentualen Anteil der Beobachtungen (relative Häufigkeiten) wider
- sowohl für **nominal- wie ordinalskalierte** Merkmale geeignet
- Für die Ausprägungen auf der X-Achse muss eine Reihenfolge festgelegt werden
- Variante: **Balkendiagramm** (Säulen waagrecht angeordnet)

Beispiel Säulendiagramm

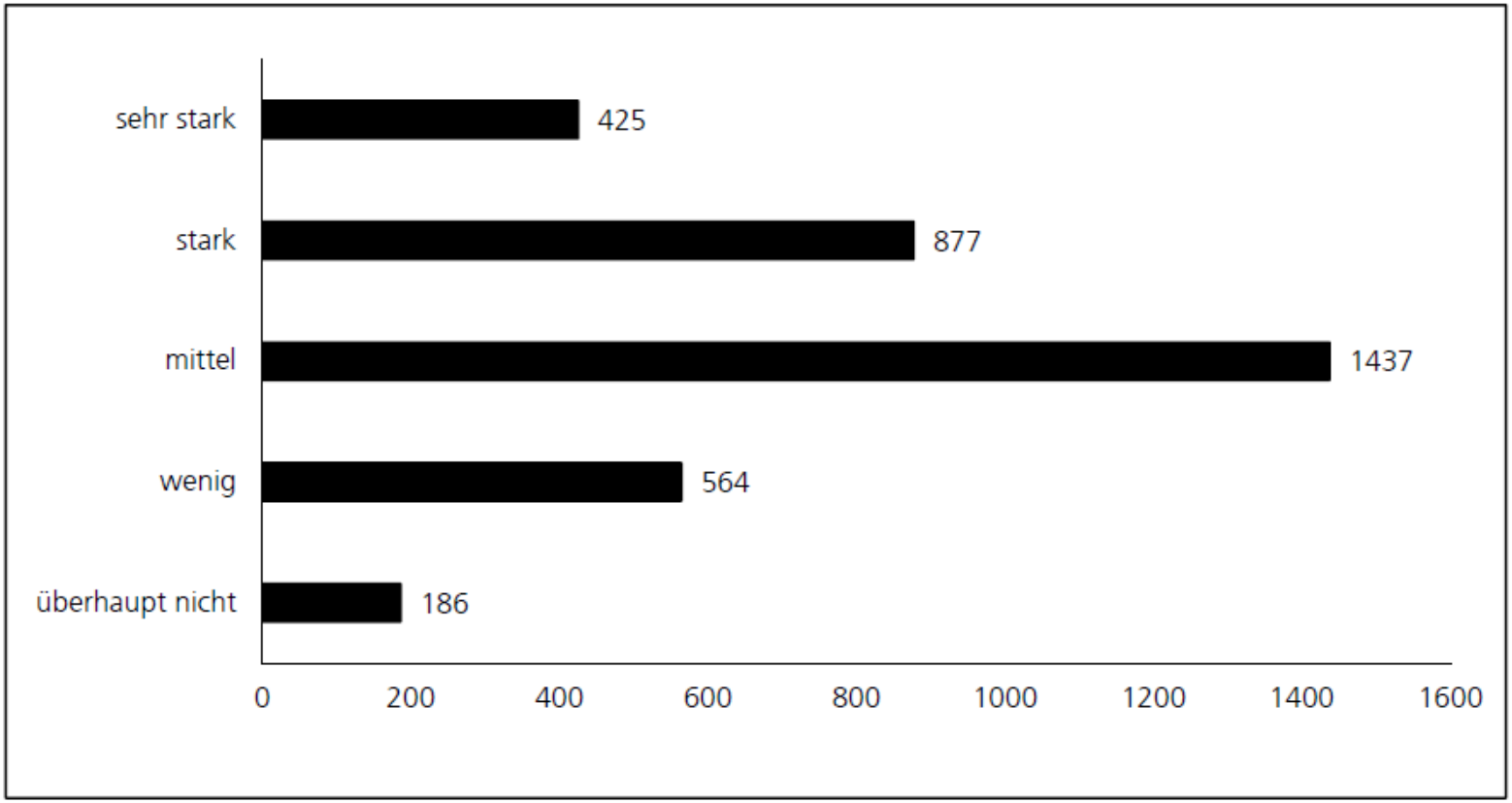


Bsp. Säulendiagramm Politisches Interesse



Daten: ALLBUS 2016. Eigene Berechnungen

Bsp. Balkendiagramm Politisches Interesse

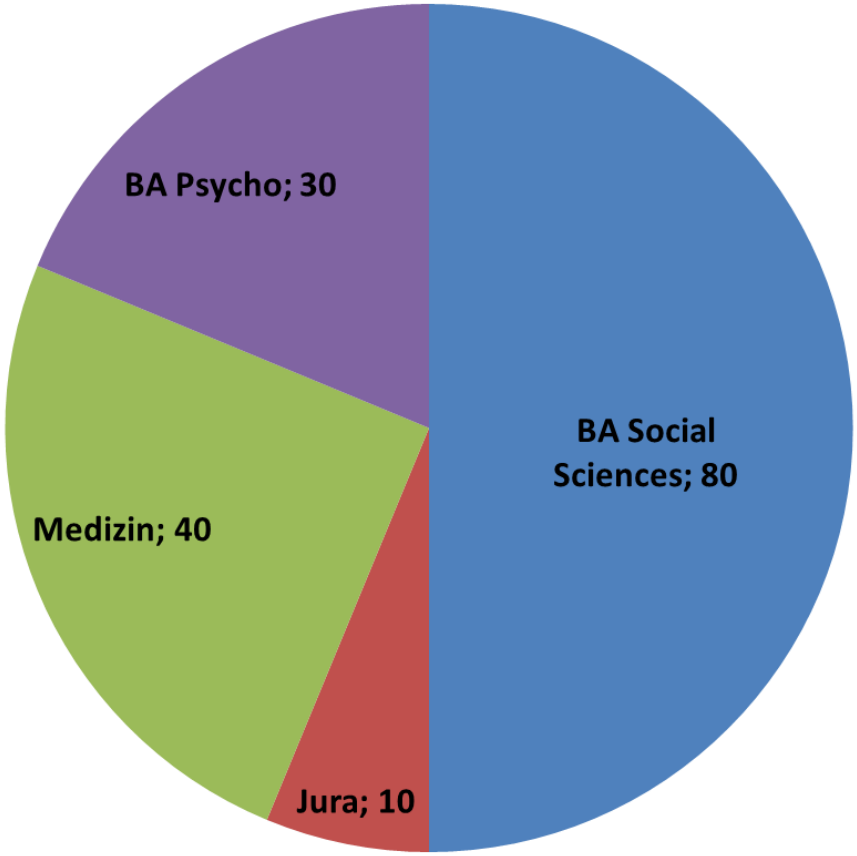


Daten: ALLBUS 2016. Eigene Berechnungen

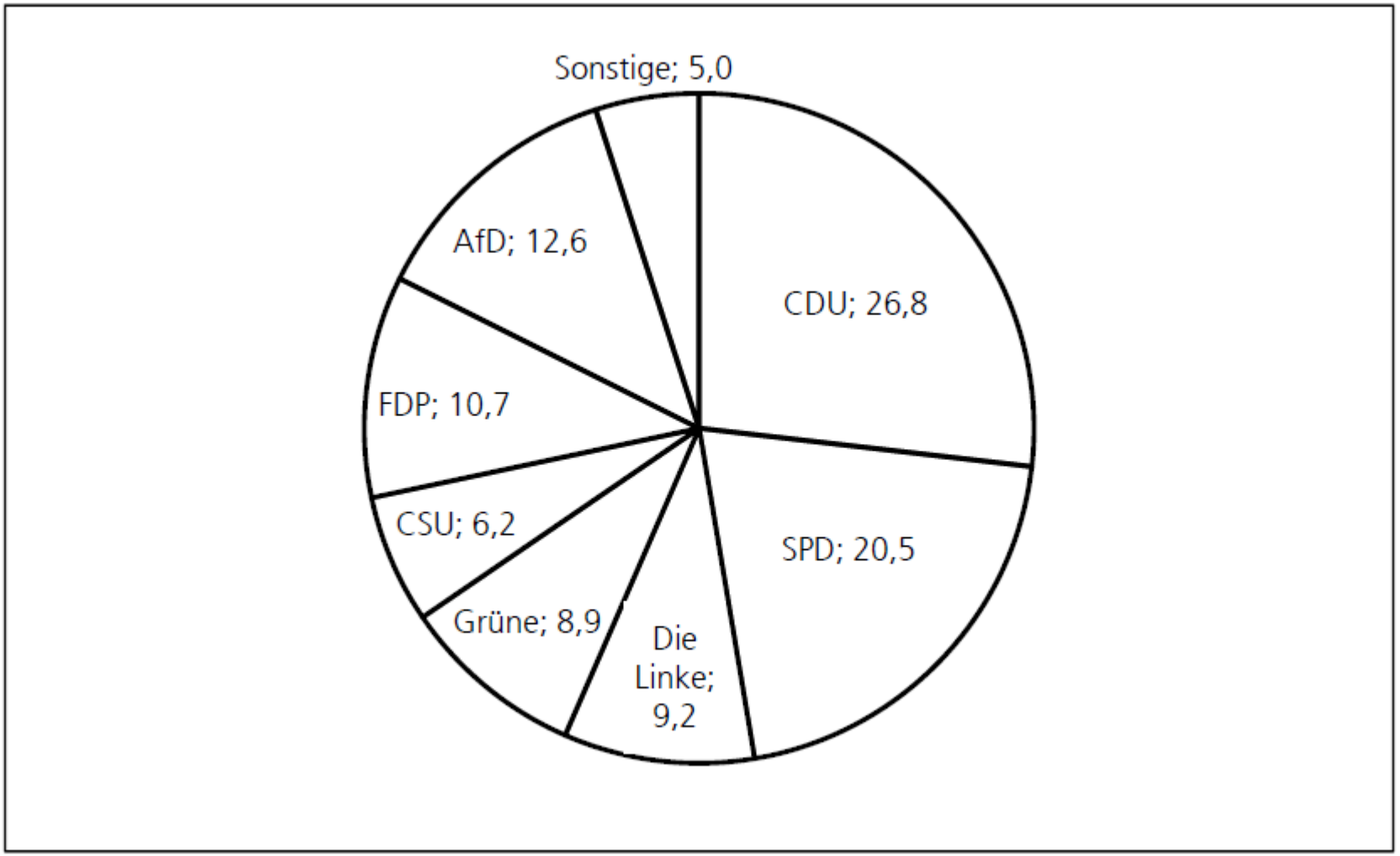
Kreisdiagramm

- Auch Tortendiagramm oder Pie-Chart
- Ein Kreis wird so in Kreissektoren unterteilt, dass die Flächen der Kreissektoren zu den beobachteten Häufigkeiten der einzelnen Ausprägungen proportional sind
- vor allem für **nominale** Daten

Beispiel 1 Kreisdiagramm, absolute Häufigkeit



Beispiel 2 Kreisdiagramm, Zweitstimmen 2017 in %

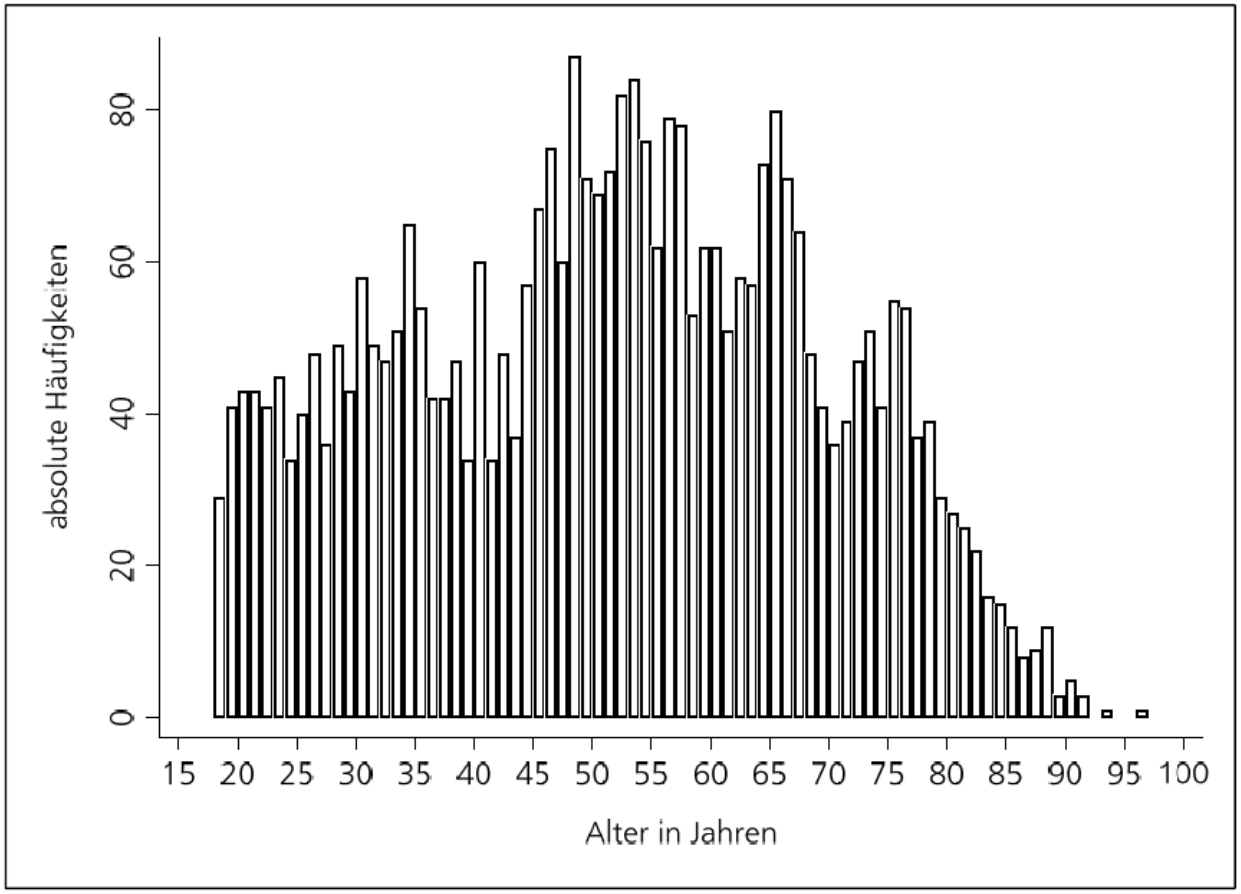


Quelle: Bundeswahlleiter (https://bundeswahlleiter.de/bundestagswahlen/2017/ergebnisse.html)

Histogramm

- Auch Flächendiagramm
- Der auffälligste Unterschied zu Säulen- und Balkendiagrammen ist, dass die Säulen eines Histogramms unmittelbar aneinander angrenzen
- Für **metrisch skalierte** Merkmale → „Ausprägungskontinuum“
- Ausprägungskategorien schließen nahtlos aneinander an

Beispiel Histogramm Alter



Daten: ALLBUS 2016. Eigene Berechnungen

Grafische Darstellung

„Faustregeln“ zur Auswahl des Grafiktyps:

- **nominales, ordinales Skalenniveau**
  - Kreisdiagramm (bis zu 6 Ausprägungen)
  - Säulen-/Balkendiagramm (bis zu 10 Ausprägungen)
- **metrische Skala**
  - Histogramm

Verteilungsformen

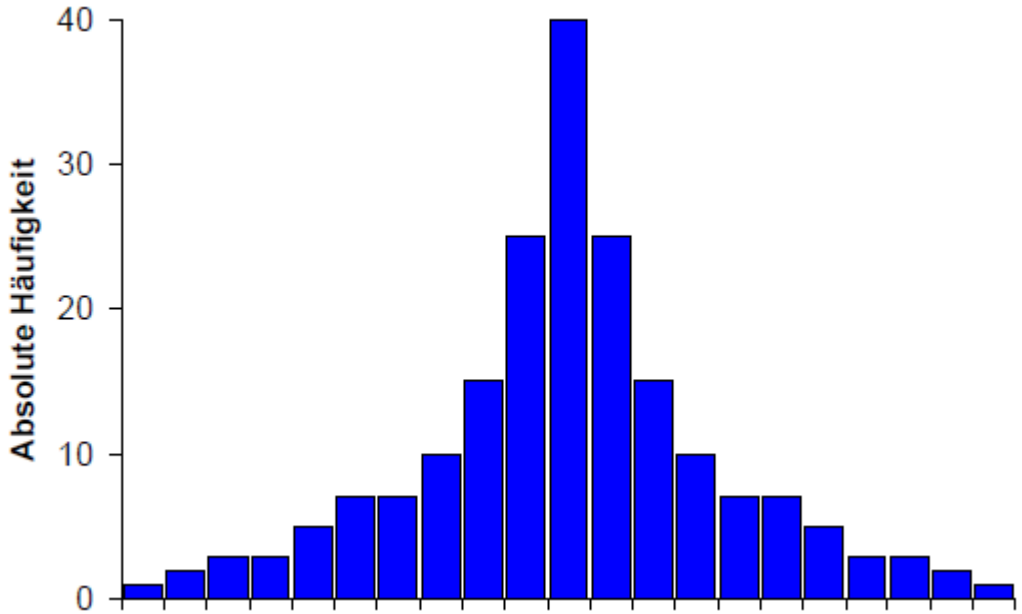
- Empirisch erhobene Daten können anhand von Häufigkeitsverteilungen tabellarisch zusammengefasst oder grafisch dargestellt werden

- Darüber hinaus lässt sich die Form einer Verteilung auch verbal beschreiben

Häufige Verteilungsformen

Symmetrische Verteilung

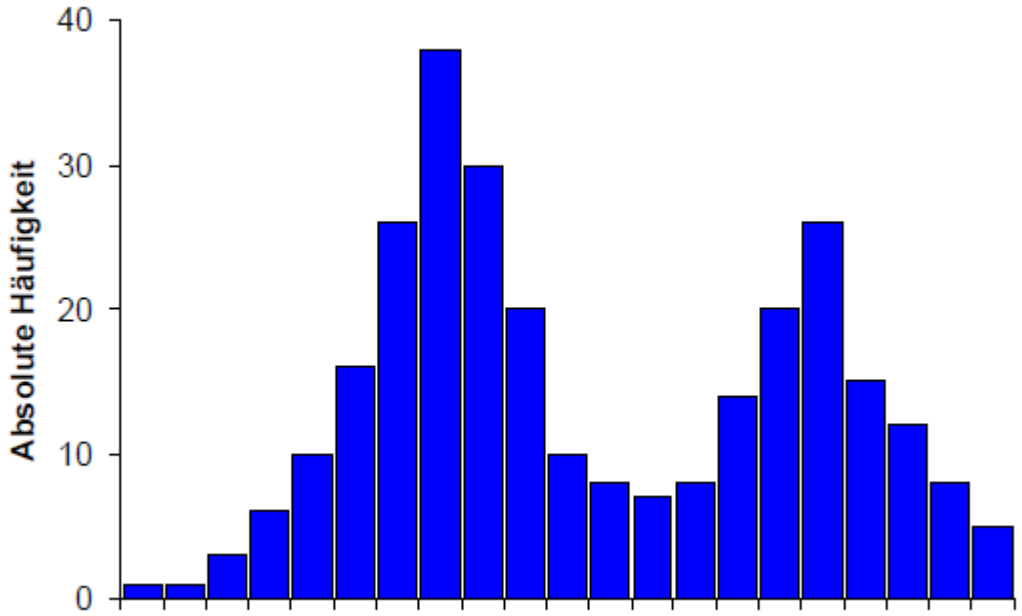
- „eingipflig“, ungefähr die gleiche Anzahl von Werten links und rechts des Gipfels



Häufige Verteilungsformen

Bimodale Verteilung

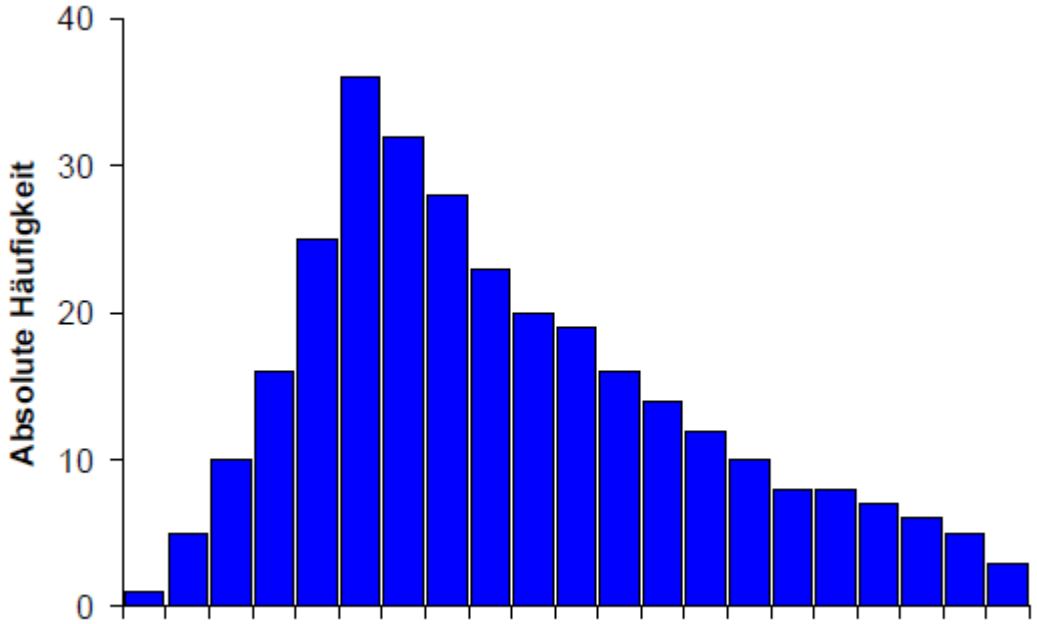
- Zwei Gipfel, evtl. 2 Subgruppen



Häufige Verteilungsformen

Linkssteile bzw. rechtsschiefe Verteilung (positive Schiefe)

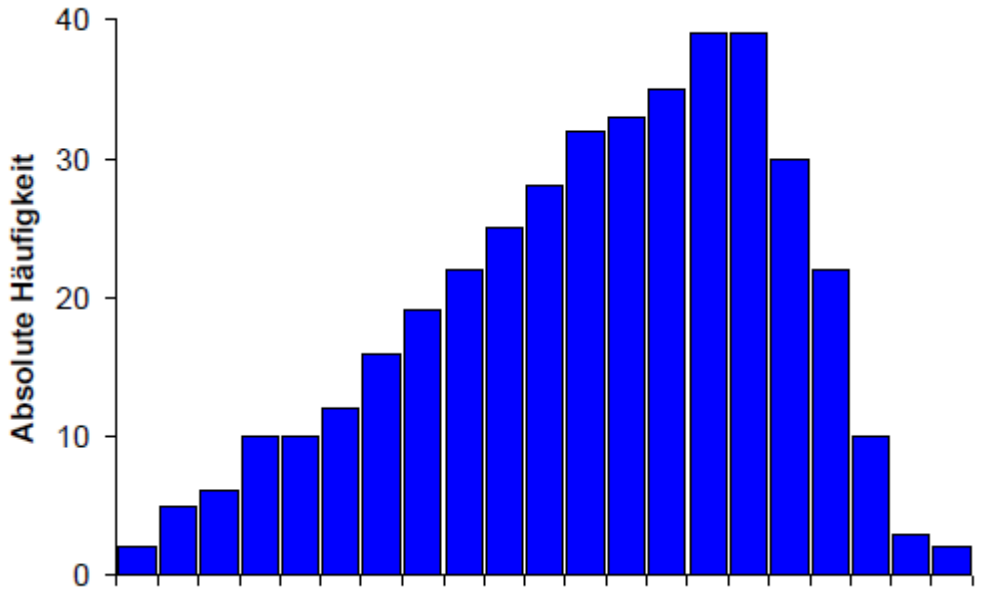
- Höchste Datenwerte recht weit vom Zentrum der Verteilung entfernt (z.B. Einkommen)



Häufige Verteilungsformen

Rechtssteile bzw. linkssschiefe Verteilung (negative Schiefe)

- Niedrigste Datenwerte relativ weit vom Zentrum der Verteilung entfernt (z.B. Sterbealter in Deutschland) **[Achtung Korrektur, dieser Satz war in der Vorlesungsaufzeichnung fehlerhaft!]**



Übung

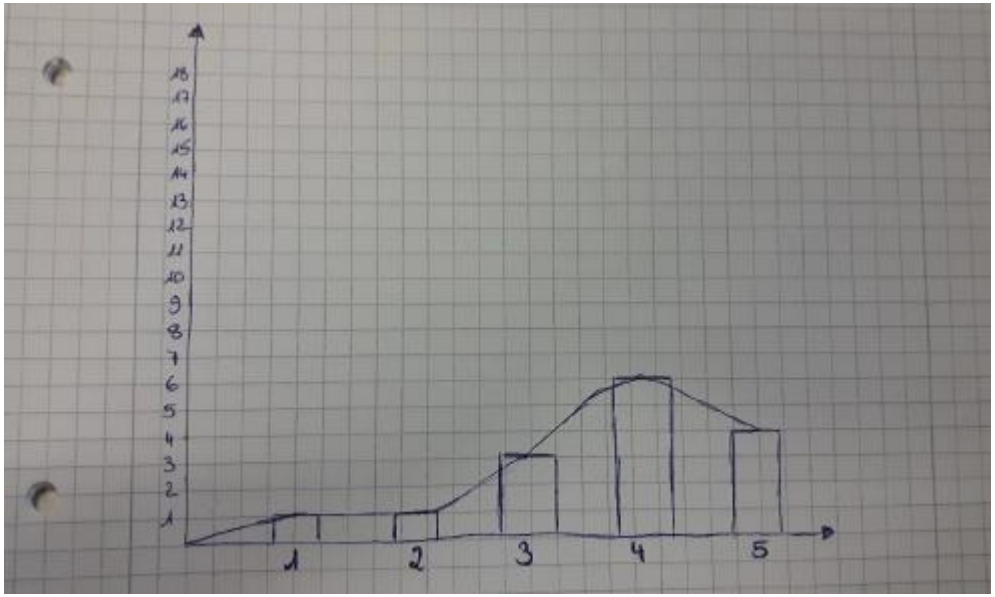
1) Bitte skizzieren Sie ein Histogramm für die absoluten Häufigkeiten der unten dargestellten Tabelle (Merkmalsausprägungen: Kein Abschluss = 1, Hauptschule = 2, Realschule = 3, Abitur = 4,Universität = 5)
2) Beschreiben Sie die Verteilungsform.

Bildungsab-schluss	$f x_k$
keiner	1
Hauptschule	1
Realschule	3
(Fach-)Abitur	6
Universität	4

Übung

1) Bitte skizzieren Sie ein Histogramm für die absoluten Häufigkeiten der unten dargestellten Tabelle (Merkmalsausprägungen: Kein Abschluss = 1, Hauptschule = 2, Realschule = 3, Abitur = 4,Universität = 5)
2) Beschreiben Sie die Verteilungsform.

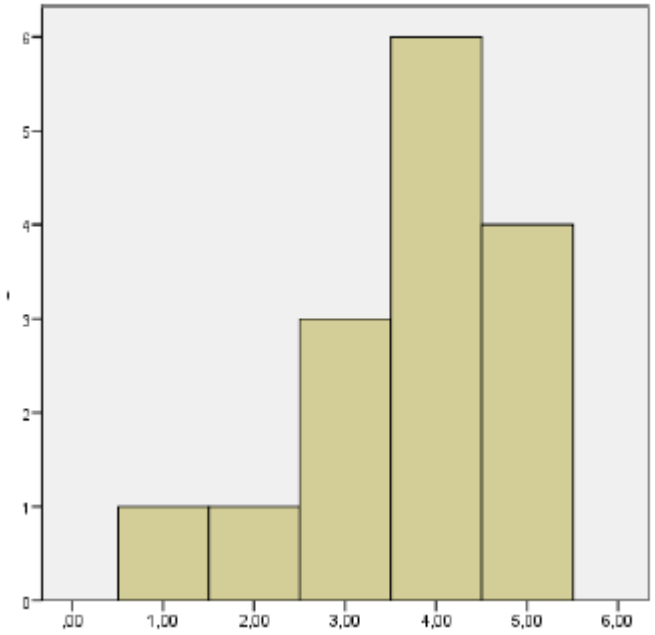
Bildungsab-schluss	$f x_k$
keiner	1
Hauptschule	1
Realschule	3
(Fach-)Abitur	6
Universität	4



Übung

1) Bitte skizzieren Sie ein Histogramm für die absoluten Häufigkeiten der unten dargestellten Tabelle (Merkmalsausprägungen: Kein Abschluss = 1, Hauptschule = 2, Realschule = 3, Abitur = 4,Universität = 5)
2) Beschreiben Sie die Verteilungsform.

Bildungsab-schluss	$f x_k$
keiner	1
Hauptschule	1
Realschule	3
(Fach-)Abitur	6
Universität	4



Notation, mathematische Formeln: Summenzeichen

Σ (‚Sigma')

- Vereinfacht die Darstellung einer Menge von Summanden zu einer Summe
- Beispiel: Was ist die Summe aller Realisationen der Variablen X (Alter)?

Person	Alter
1	21
2	20
3	21
4	20
5	23
6	25
7	20
8	23

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 = 173$$

$$= \sum_{i=1}^{n} x_i$$

Sprich: „Die Summe aller x_i-Werte für Merkmalsträger i = 1 bis Merkmalsträger n"

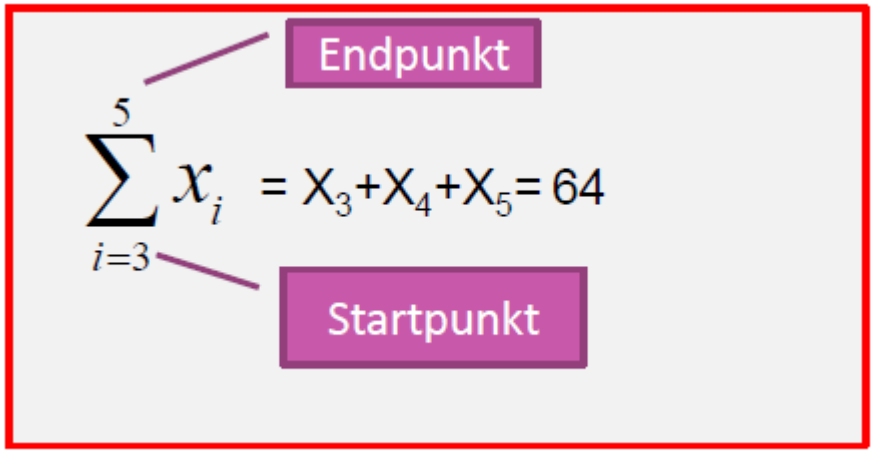
Summenzeichen

$$\sum_{i=1}^{n} x_i$$

- Unterhalb des Summenzeichens steht der Laufindex *i*
- *i* kennzeichnet konventionsgemäß einen einzelnen Fall
- Nach dem Gleichheitszeichen folgt der Wert, mit dem die Summierung beginnt (in unserem Fall der 1. Wert)
- Oberhalb der Summenzeichens steht der letzte Wert des Index, hier n
- Bei der Summenbildung wird zunächst der erste Indexwert verwendet
- Anschließend wird dieser Wert um +1 erhöht, bis der letzte Wert erreicht ist
- Hinter dem Index steht der Ausdruck, der aufsummiert werden soll – hier die Realisationen der Variablen X, also die Werte x_i

Rechnen mit Summenzeichen

Person	Alter (Jahre)
1	21
2	20
3	21
4	20
5	23
6	25
7	20
8	23



Sprich: „Die Summe aller x_i-Werte für Merkmalsträger i = 3 bis Merkmalsträger 5“

Notation, mathematische Formeln: Summenzeichen

- Vereinfachung: Wenn die Indexwerte eindeutig vorausgesetzt sind, wird oft der Start- und Endwert des Index nicht angegeben

$$\sum x_i \quad \text{anstelle von} \quad \sum_{i=1}^{n} x_i$$

Übung

Bitte bestimmen Sie die folgenden Werte für die in der nachstehenden Häufigkeitstabelle wiedergegebenen Daten:

a) n=

b) $\sum x_i$

c) $\sum x_i^2$

x_k	absolute Häufigkeit f_{x_k}
1	1
2	4
3	2
4	2
5	1

Lernziele Einheit

- Verständnis des Aufbaus einer Datenmatrix
- Grundlegende Kenntnis tabellarischer und grafischer Darstellungsformen von univariaten statistischen Informationen
- Kenntnis des Summenzeichens