Rohdaten

Urliste oder **Rohdaten** sind die auszuwertenden Daten in der Form, wie sie nach der Datenerhebung vorliegen. Dimensionen der Urliste sind die **Fälle**, **Merkmale** und ihre **Ausprägungen.**

Voraussetzung für statistische Auswertung: jeder Fall besitzt in bezug auf jedes Merkmal genau eine Ausprägung

Beispiele für Rohdaten: Fälle, Merkmale, Ausprägungen

Fälle	Merkmal	Ausprägungen
Ausgefüllte Fragebögen	Frage im Fragebogen	angekreuzte Antwort
mündl. interviewte Personen	Frage des Interviewers	Antwort-Kategorien
Texte	Beurteilungskriterien: z.B. Inhalt, Länge, Stil	Art des Inhalts, Zahl der Worte, ...
Beobachtete Objekte, z.B. Menschen im Lokal	Beobachtete Aktivitäten, z.B. trinken, sich unterhalten	Intensität der Handlung, z.B. Anzahl Biere, Anzahl der Gesprächspartner, ...

Häufigkeitsverteilung

Die **Häufigkeitsverteilung (auch kurz „Verteilung“)** eines Merkmals ist die Darstellung seiner Ausprägungen im Verhältnis ihres Auftretens in den Fällen.

20 Antworten auf: „Wie weit stimmen Sie mit der folgenden Meinung überein: Soziale Unterschiede sind gerecht“:

4,3,3,2,3,1,4,4,3,3,3,2,2,4,3,4,2,3,2,4

(1 = stimme voll zu, 2 = stimme eher zu, 3 = stimme eher nicht zu, 4 = stimme gar nicht zu)

Messwertklassen, gruppierte Daten

Einkommen der 20 Befragten:
900,1600,800,2300,1300,2700,2100,2500,4000
1700,3300,1400,1900,1480,2900,1200,1150,
600,4100,2700

Messwertklassen sind die Zusammenfassungen von Ausprägungen zu Gruppen. Die sich daraus ergebenden Daten heißen **gruppierte Daten.**

Relative und prozentuale Häufigkeit

Sei $f(X_k)$ die **absolute Häufigkeit** der Ausprägung k des Merkmals X bei N Fällen, dann ist

$p(X_k) = f(X_k) /N$ die **relative Häufigkeit** oder **Anteil** und

$Proz(X_k) = p(X_k)*100$ die **prozentuale Häufigkeit**.

Dann gilt:

$$\sum_{alle\_k} \mathbf{f(X_k) = N}, \quad \sum_{alle\_k} \mathbf{p(X_k) = 1.0},$$

$$\sum_{alle\_k} \mathbf{Proz(X_k) = 100.0}$$

Aufgaben der Statistik

Aufgaben der Statistik, mit Häufigkeitsverteilungen formuliert:

- Die Häufigkeitsverteilung mit mathematisch durch Formeln erzeugten Verteilungen zu vergleichen, die sich aus einer Theorie über die Daten ergeben.
- Die Häufigkeitsverteilung durch möglichst wenige Kennzahlen, sog. Parameter, ausreichend zu beschreiben.
- Die Häufigkeitsverteilungen zweier oder mehrerer Merkmale – u.a. mit Hilfe ihrer Kennzahlen - daraufhin zu vergleichen, ob sie in irgendeiner Weise miteinander verkoppelt sind.

Kumulierte Häufigkeitsverteilung

Die **kumulierte Häufigkeitsverteilung F** gibt zu jedem Wert a des Merkmals X an, wie viele Fälle kleiner oder gleich diesem Wert a sind (als relative Häufigkeit).

$$F(a) = \sum_{X \le a} p(X)$$

Verteilungsfunktion eines Merkmals heißt die Funktion F der kumulierten relativen Häufigkeiten. Sie hat 2 Eigenschaften:

1. Ihre Werte liegen zwischen 0 und 1.
2. Sie wächst monoton von 0 auf 1.

Beispiel Häufigkeitsverteilung

BRD 1998

Soziale Unterschiede sind gerecht

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	STIMME VOLL ZU	125	6,4	6,7	6,7
	STIMME EHER ZU	483	24,7	25,8	32,5
	ST.EHER NICHT ZU	746	38,2	39,9	72,4
	ST.GAR NICHT ZU	517	26,5	27,6	100,0
	Gesamt	1871	95,9	100,0	
	Missing	81	4,1		
Gesamt		1952	100,0		

USA 1993

SOCIAL DIFFERENCES ARE ACCEPTABLE

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	STRONGLY AGREE	120	,4	8,6	8,6
	SOMEWHAT AGREE	631	1,9	45,4	54,1
	SOMEWHT DISAGREE	474	1,5	34,1	88,2
	STRNGLY DISAGREE	164	,5	11,8	100,0
	Gesamt	1389	4,3	100,0	
Fehlend	NAP	30907	95,5		
	NO OPINION	58	,2		
	NA	26	,1		
	Gesamt	30991	95,7		
Gesamt		32380	100,0		

BRD 2014

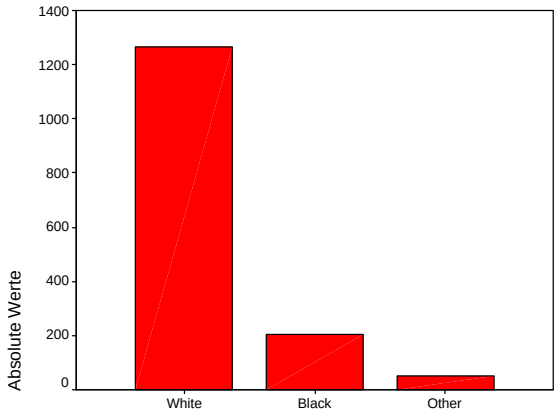
V201 SOZIALE UNTERSCHIEDE SIND GERECHT

		Häufigkeit	Prozent	Gültige Prozent	Kumulative Prozente
Gültig	1 STIMME VOLL ZU	179	5,2	5,3	5,3
	2 STIMME EHER ZU	905	26,1	26,6	31,8
	3 STIMME EHER NICHT ZU	1593	45,9	46,8	78,6
	4 STIMME GAR NICHT ZU	727	20,9	21,4	100,0
	Gesamtsumme	3404	98,1	100,0	
Fehlend	8 WEISS NICHT	48	1,4		
	9 KEINE ANGABE	19	,5		
	Gesamtsumme	67	1,9		
Gesamtsumme		3471	100,0		

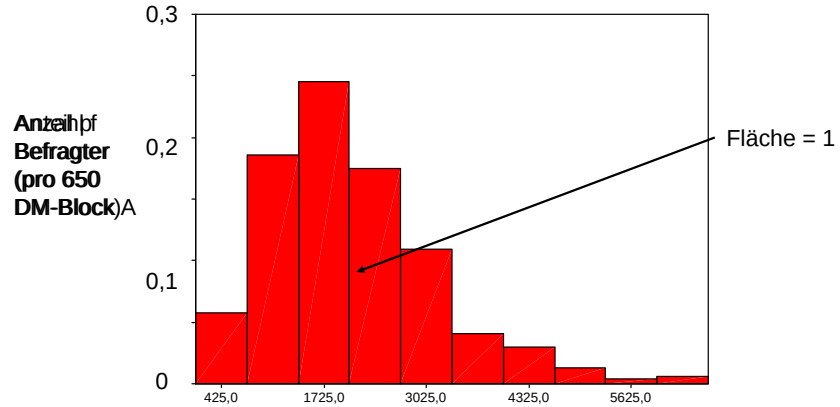
Grundprinzipien bei Grafiken

- ausreichend gekennzeichnet
- mathematisch genaue Umsetzung von Zahlen in grafische Objekte
- die Ausprägungen des Merkmals stehen auf der Abszisse (horizontale Achse, X-Achse)
- die Häufigkeiten jeder Ausprägung stehen auf der Ordinate (vertikale Achse, Y-Achse)

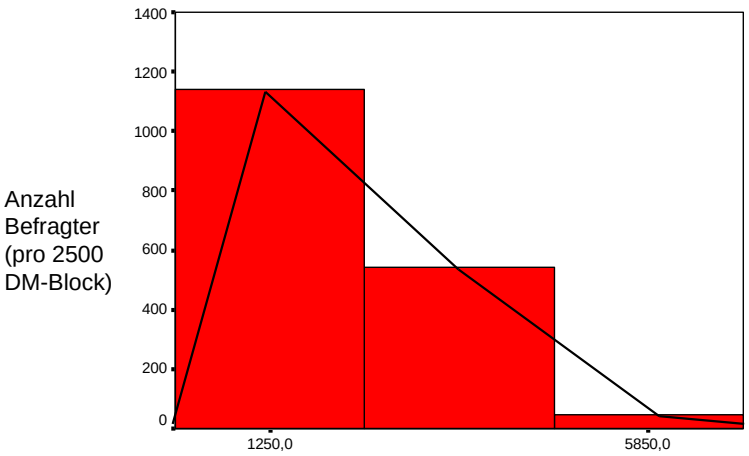
Ein **Balkendiagramm** ist eine Darstellung einer Häufigkeitsverteilung von nominalen Daten in Säulenform, wobei sich die Säulen nicht berühren.



Ein **Histogramm** ist eine Darstellung einer Häufigkeitsverteilung von ordinalen oder gruppierten Daten in Säulenform, wobei die Säulen aneinander anschließen.

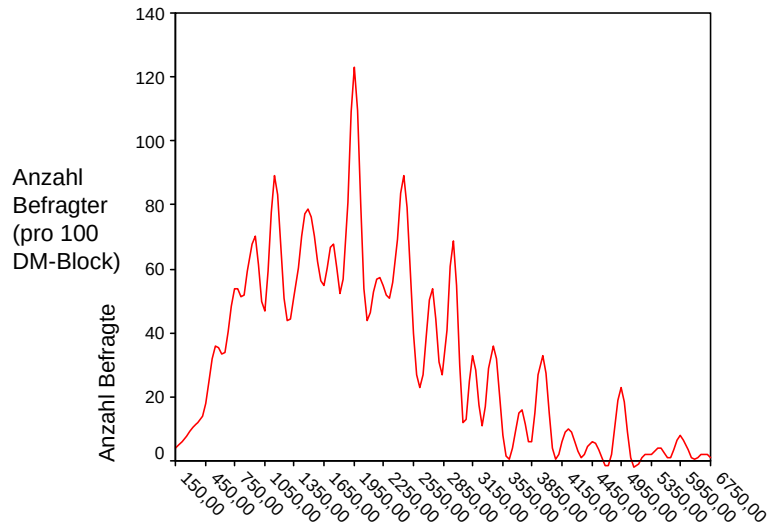


Polygonzug



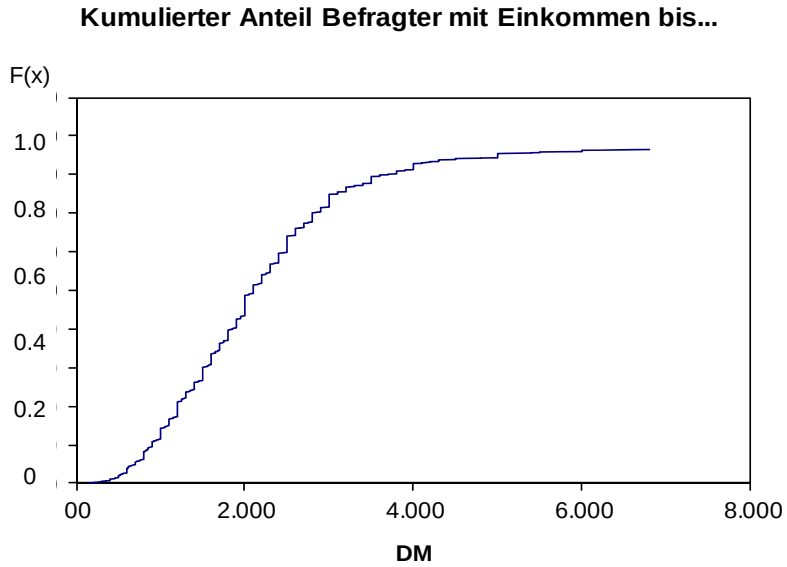
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE

Feines Histogramm: Dichtefunktion

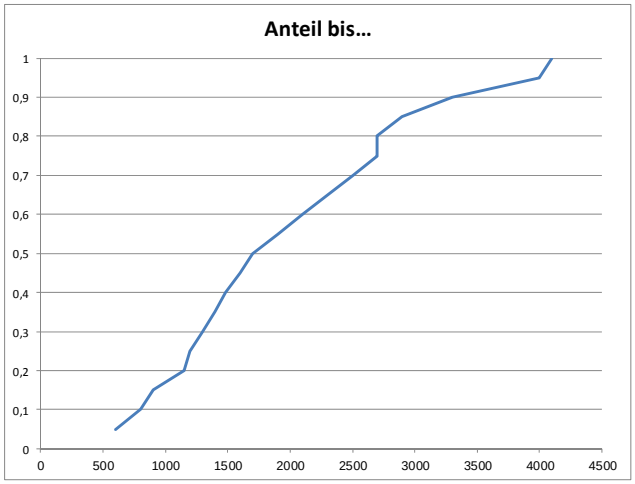


Einkommen in DM

Grafik der kumulierten Verteilung



Kumulierte Verteilun g: Beispiel



Befragter	Einkommen
1	600
2	800
3	900
4	1150
5	1200
6	1300
7	1400
8	1480
9	1600
10	1700
11	1900
12	2100
13	2300
14	2500
15	2700
16	2700
17	2900
18	3300
19	4000
20	4100

Kreuztabelle (Beziehung zweier kategorialer Merkmale)

Befragter	Einkommen	Soz.Unt. gerecht
1	900	2
2	1600	4
3	800	3
4	2300	2
5	1300	4
6	2700	2
7	2100	3
8	2500	3
9	4000	2
10	1700	3
11	3300	4
12	1400	4
13	1900	2
14	1480	3
15	2900	4
16	1200	3
17	1150	3
18	600	4
19	4100	1
20	2700	3

Kreuztabelle zweier Merkmale

	1 Stimme voll zu	2 Stimme eher zu	3 Stimme eher nicht zu	4 stimme gar nicht zu	Summe
< 1500	-	1	4	3	8
1500 - 3000	-	3	4	2	9
> 3000	1	1	-	1	3
Summe	1	5	8	6	20

Randverteilungen

In den jeweiligen Summenspalten einer Kreuztabelle stehen die beiden **Randverteilungen.** Es sind die Häufigkeitsverteilungen jedes der beiden Merkmale.

Jede **Zelle** in der Kreuztabelle enthält die Anzahl der Fälle, die die Ausprägung derjenigen beiden Merkmale besitzen, durch die die Zelle gebildet wird.

Die „Ursache“ bzw. „**unabhängige Variable**“ sollte in den Zeilen, die „Wirkung“ bzw. „**abhängige Variable**“ in den Spalten stehen.

Prozentuierungen, bedingte Verteilung

Um relative oder prozentuale Häufigkeiten zu bilden, hat man bei einer Kreuztabelle drei Möglichkeiten: man kann auf die Gesamtsumme der Fälle oder auf eine der beiden Randverteilungen prozentuieren. In den letzten beiden Fällen spricht man von **Zeilen- oder Spaltenprozenten**, je nachdem, ob die Summen der Zeilen oder die Summen der Spalten die jeweiligen 100% bilden.

Die Verteilung eines Merkmals A unter der Bedingung, dass ein anderes Merkmal B eine bestimmte Ausprägung hat, heißt

bedingte Verteilung von A unter B

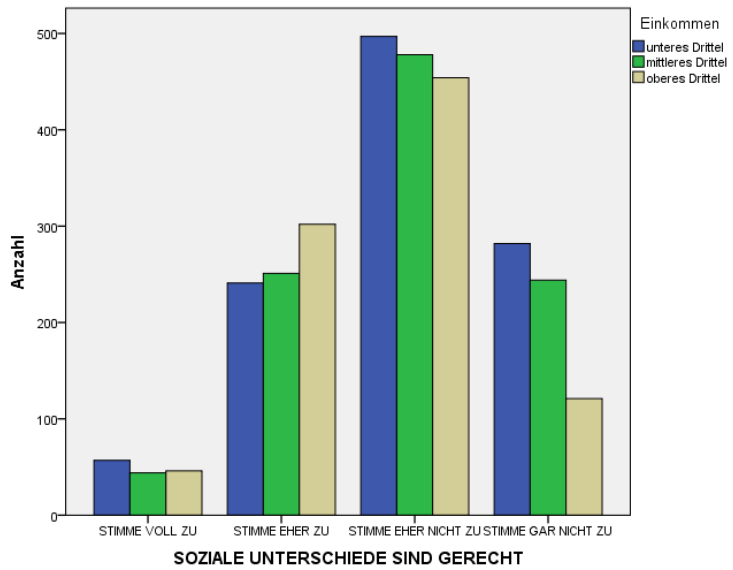
Beispiel Kreuztabelle

Kreuztabelle Einkommen*SOZIALE UNTERSCHIEDE SIND GERECHT

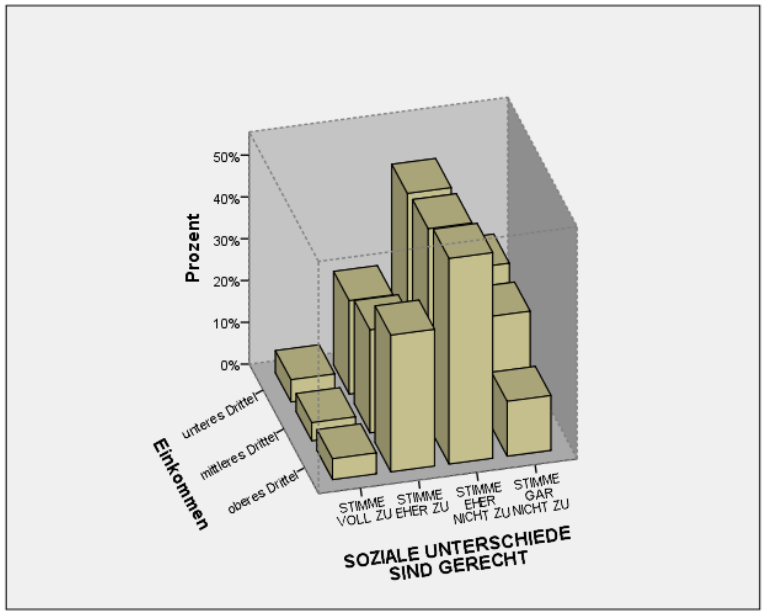
			SOZIALE UNTERSCHIEDE SIND GERECHT				Gesamtsumme
			STIMME VOLL ZU	STIMME EHER ZU	STIMME EHER NICHT ZU	STIMME GAR NICHT ZU	
Einkommen	unteres Drittel	Anzahl	57	241	497	282	1077
		% in Einkommen	5,3%	22,4%	46,1%	26,2%	100,0%
	mittleres Drittel	Anzahl	44	251	478	244	1017
		% in Einkommen	4,3%	24,7%	47,0%	24,0%	100,0%
	oberes Drittel	Anzahl	46	302	454	121	923
		% in Einkommen	5,0%	32,7%	49,2%	13,1%	100,0%
Gesamtsumme		Anzahl	147	794	1429	647	3017
		% in Einkommen	4,9%	26,3%	47,4%	21,4%	100,0%

Quelle: ALLBUS 2014

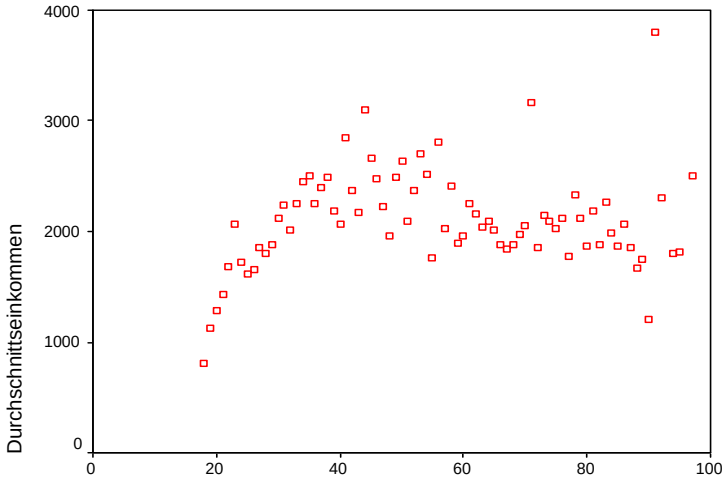
Gruppiertes Balkendiagramm



3D-Säulendiagramm



Streudiagramm (Beziehung zweier metrischer Merkmale)



3-dimensionale Kreuztabelle

Kreuztabelle Einkommen*Soziale Unterschied gerecht?

% in Einkommen

ERHEBUNGSGEBIET <WOHNGEBIET>: WEST - OST			Soziale Unterschied gerecht?		Gesamtsumme
			stimme zu	stimme nicht zu	
ALTE BUNDESLAENDER	Einkommen	unteres Drittel	35,3%	64,7%	100,0%
		mittleres Drittel	34,2%	65,8%	100,0%
		oberes Drittel	39,5%	60,5%	100,0%
	Gesamtsumme		36,5%	63,5%	100,0%
NEUE BUNDESLAENDER	Einkommen	unteres Drittel	14,9%	85,1%	100,0%
		mittleres Drittel	21,2%	78,8%	100,0%
		oberes Drittel	30,2%	69,8%	100,0%
	Gesamtsumme		20,3%	79,7%	100,0%

Hausaufgabe:

Zum Nacharbeiten der Vorlesung vor den Hausaufgaben: Müller-Benedict, Kap. 4

Sie haben folgende Daten erhoben (s.Tabelle): für 15 Befragte das monatliche Nettoeinkommen (in €) und den formalen Schulabschluss.

– Stellen Sie die *absolute und relative Häufigkeitsverteilung* (nicht die Kreuztabelle) jeweils des Schulabschlusses und des Haushaltseinkommens dar, tabellarisch und grafisch. Gruppieren Sie dafür, wenn nötig, die Daten in *geeignete Messwertklassen*. Formulieren Sie eine kurze „Zeitungsmeldung“ über das Ergebnis.
– Zeichnen Sie die *kumulierte Verteilung* des Haushaltseinkommens möglichst genau („Haushaltseinkommen“ ist ein metrisches Merkmal!) auf.
– Gibt es eine Beziehung zwischen beiden Merkmalen? Wie können Sie das darstellen?

Fall Nr.	Einkommen	Schulabschluss
1	5740	Abitur
2	4500	Abitur
3	2600	Realschule
4	2456	Realschule
5	4010	Hauptschule
6	3976	Realschule
7	7845	Abitur
8	2250	Hauptschule
9	1550	Realschule
10	1855	Hauptschule
11	3671	Abitur
12	1935	Hauptschule
13	3678	Realschule
14	1867	Realschule
15	3216	Abitur