Willkommen zur Vorlesung Statistik

Thema dieser Vorlesung:
Häufigkeiten und ihre Verteilung, oder:
Zusammenfassende Darstellungen einzelner Variablen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Inhaltsübersicht

- Einstieg/Wiederholung
- Häufigkeitstabellen
- Verteilungen graphisch: Kategoriale Merkmale
- Verteilungen graphisch: Merkmale mit vielen Ausprägungen

Wiederholung

1. Deskriptive Statistik vs. Inferenzstatistik
2. Urliste (Liste der Messwerte), Datenmatrix (Anordnung der Messwerte)
3. Messniveaus (Skalenniveaus):
   Kategoriale Daten: Nominalskaliert – ordinalskaliert
   Metrische Daten: Intervallskaliert – ratioskaliert
4. Diskrete/stetige Merkmale (und die ‚Relativität' dieser Unterscheidung).

Die Häufigkeitsverteilung: Numerische Darstellung absoluter und relativer Häufigkeiten

Häufigkeitsverteilung

Die Häufigkeitsverteilung stellt dar, wie oft die unterschiedlichen „Zahlenwerte“, also Merkmalsausprägungen, einer Variablen des Datensatzes vorhanden sind. Bei Merkmalen mit nicht allzu vielen Ausprägungen kann so eine erste Übersicht gegeben werden.

Die Ausprägungen einer Variablen (eines Merkmals) bezeichnen wir im Folgenden als a_j.

Ein Beispiel

Antworten auf die Frage:

Ein Integral beschreibt
(1) die Differenz zwischen zwei Punkten auf einer Kurve
(2) die Steigung einer Geraden
(3) den Flächeninhalt unter einer Kurve
(4) eine Rechenmethode zur Bestimmung der Kreisfläche
(5) den Schnittpunkt der Kurve mit der y-Achse

„Urliste“: 2 5 4 5 2 3 1 1 3 3 3 3 3 3 1 1 1 1 3 5 3 -9 -9 -9 -9 -9 -9

(-9 heißt: Befragte/r hatte nichts angekreuzt.)

Absolute Häufigkeiten

Die absoluten Häufigkeiten h_j geben an, wie viele Fälle einer bestimmten Ausprägung a_j (j=1, . . . ,k) entsprechen.

Nichts angekreuzt	|||||	→ 6
Differenz	|||||	→ 6
Steigung	||	→ 2
Flächeninhalt	||||||||	→ 9
Kreisfläche	|	→ 1
Schnittpunkt	|||	→ 3

Relative Häufigkeiten

Relative Häufigkeiten f_j: $f_j = h_j/n$

Relative Häufigkeiten f_j als Prozentwerte: $f_j = (h_j/n) \cdot 100$

mit h_j: absolute Häufigkeiten und n: Zahl der Messwerte

	h_j	Prozent
Nichts angekreuzt	6	22,2
Differenz	6	22,2
Steigung	2	7,4
Flächeninhalt	9	33,3
Kreisfläche	1	3,7
Schnittpunkt	3	11,1
n	27	

Hinweis: Die Prozentwerte sollten i. d. R. auf ganze Stellen gerundet werden!

Relative Häufigkeiten: Fehlende Werte

Weist eine Variable fehlende Werte auf, muss man entscheiden, ob man die relativen Häufigkeiten auf alle Fälle oder nur jene mit gültigen Werten bezieht. Software gibt häufig Tabellen aus, die beide Werte wiedergeben:

Integral ist ...	absolut	% alle	% nur gültige
Differenz	6	22,2	28,6
Steigung	2	7,4	9,5
Flächeninhalt	9	33,3	42,9
Kreisfläche	1	3,7	4,8
Schnittpunkt	3	11,1	14,3
Gesamt (nur gültige)	21	77,8	100
Nichts angekreuzt	6	22,2	
Gesamt (alle)	27	100	

Hinweis: M. E. sollte man die Gesamt-Prozent immer als 100 angeben (alle Fälle = 100 Prozent). Gelegentlich findet man aber auch als Gesamtprozent die Summe der einzelnen Prozentwerte

(Relative) kumulierte Häufigkeiten

$$F(x) = f_1 + \ldots f_j = \sum_{i=1}^{j} f_i$$

Sinnvoll nur bei mindestens ordinalskalierten Merkmalen. Geben an, wie groß der Anteil der Fälle ist, deren Ausprägung kleiner oder gleich a_j ist. Beispiel: Mathematiknote der Befragten im Abiturzeugnis.

Note Mathematik	absolut	Prozent	Kum.Prozent
1	3	11,1	11,1
2	7	25,9	37,0
3	9	33,3	70,4
4	8	29,6	100,0
Gesamt	27	100	100

Es können auch absolute kumulierte Häufigkeiten berechnet werden.

Merkmale mit vielen Ausprägungen

Metrische Variablen (Alter, Einkommen, Zahl der Überstunden, Berufserfahrung ...) haben oft viele Ausprägungen. Beispiel: Alter von Berufstätigen.

21	22	24	24	25	26	27	27	27	27
28	28	28	28	29	31	31	31	31	32
32	32	33	33	33	34	34	35	35	35
35	35	36	36	36	36	36	37	37	37
37	37	37	37	38	38	39	39	39	39
40	40	41	41	41	41	41	42	42	42
43	43	44	45	45	45	46	47	47	47
47	48	48	48	49	49	49	52	52	52
53	53	53	54	54	55	55	55	56	57
58	60	60	61	64	65	65	65	65	65

Quelle: Simulierte Daten, $n = 100$

Merkmale mit vielen Ausprägungen

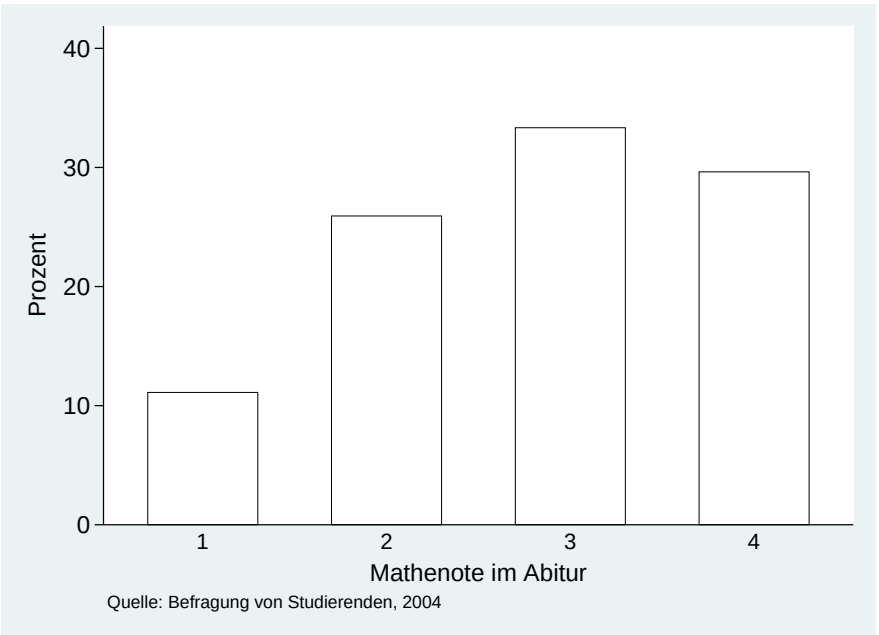
Eine gruppierte oder klassierte Häufigkeitstabelle ist eine einfache Möglichkeit, sich einen Überblick zu verschaffen, sie ist jedoch mit Informationsverlust verbunden. Die Klassenbreite kann gleich oder unterschiedlich sein (letzteres, um etwa unterschiedlichen Häufigkeiten gerecht zu werden).

Alter in Jahren	Prozent
20 bis 24	4
25 bis 29	11
30 bis 34	12
35 bis 39	23
40 bis 44	13
45 bis 49	14
50 bis 54	8
55 bis 59	6
60 bis 64	4
65 bis 69	5
n	100

Graphische Darstellung von Verteilungen, Teil I: Kategoriale Merkmale

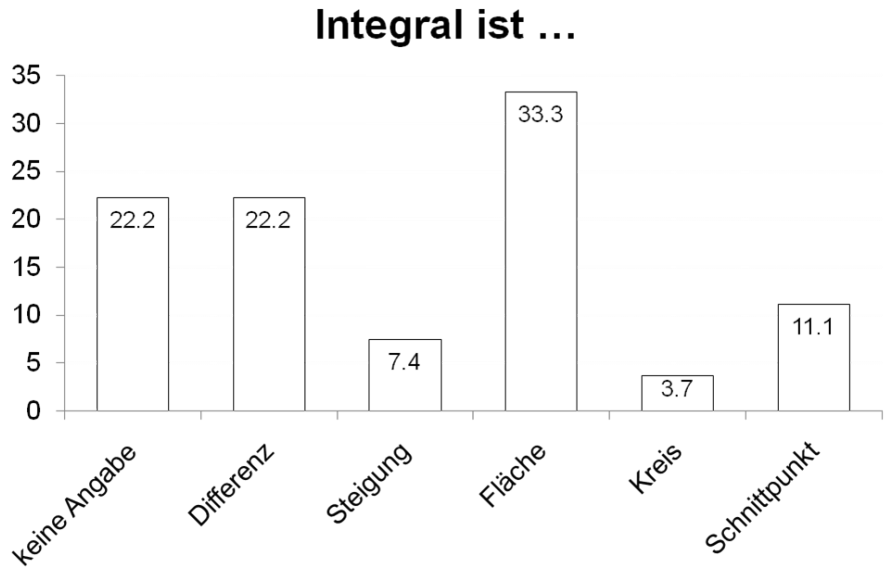
Kategoriale Merkmale, Möglichkeit I: Säulendiagramm

Die einzelnen Ausprägungen werden in einer Säule dargestellt, deren Höhe proportional den relativen Häufigkeiten ist. Beispiel: Mathematik-Noten der Studierendenbefragung.



Säulendiagramm: Alternative

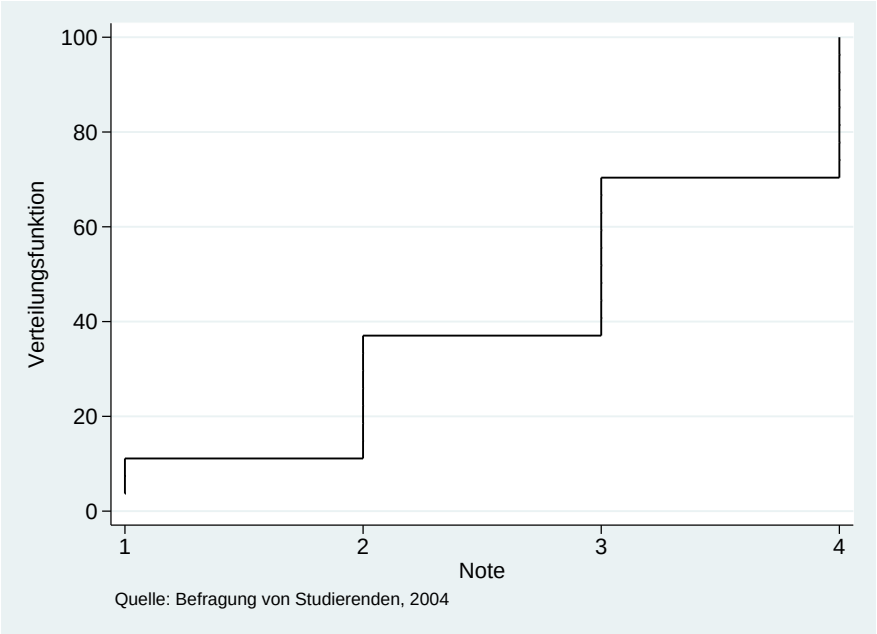
Lässt man die die Gitternetzlinien der Graphik weg, kann es zweckvoll sein, die relativen Häufigkeiten in der Graphik anzugeben.



Angaben in Prozent. Quelle: Befragung von Studierenden, 2004
Achtung: Aufgrund technischer Probleme erscheinen hier teilweise oder überwiegend leider doch Gitternetzlinien. Sorry! In der nächsten Version wird es vielleicht besser ...

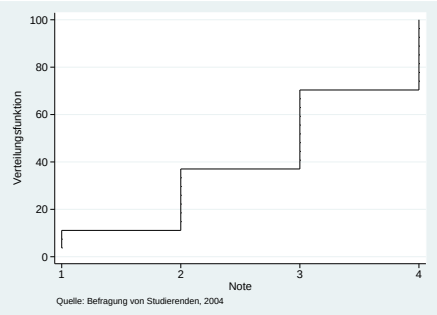
Empirische Verteilungsfunktion I

Die relativen kumulierten Häufigkeiten werden auch als empirische Verteilungsfunktion bezeichnet. Sie können ebenfalls graphisch dargestellt werden (Beispiel Mathematiknoten; Erläuterungen nächste Seite!):



Die Verteilungsfunktion endet bei 100 (bzw. bei 1 in nicht-prozentualer Darstellung)

Empirische Verteilungsfunktion II



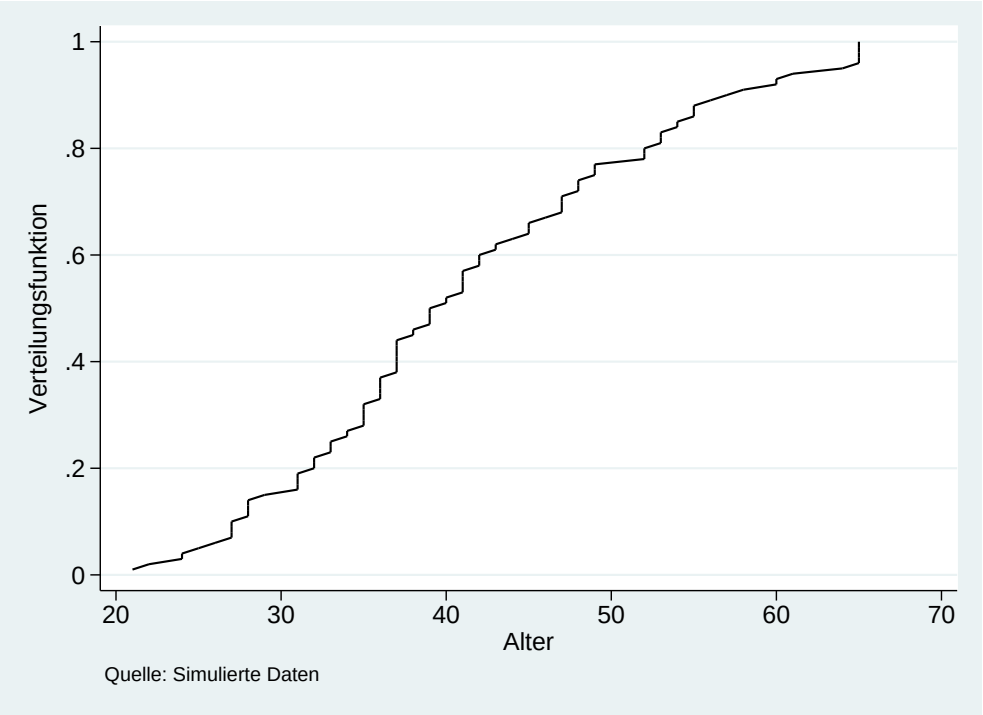
Bei Wert 1 (X-Achse) hat die Funktion einen Wert von etwas über 10 $\rightarrow$ entspricht den 11 Prozent aus der Tabelle.

Bei Wert 2 (X-Achse) steigt die Funktion auf einen Wert von etwas unter 40 $\rightarrow$ entspricht den 37 Prozent aus der Tabelle.
Gleichzeitig entspricht die Höhe der Linie beim X-Wert 2 den 26 Prozent, die auf diesen Wert entfallen.

Bei Wert 3 (X-Achse) steigt die Funktion auf einen Wert von ca. 70 $\rightarrow$ entspricht den gut 70 Prozent aus der Tabelle.
Usw.

Empirische Verteilungsfunktion III

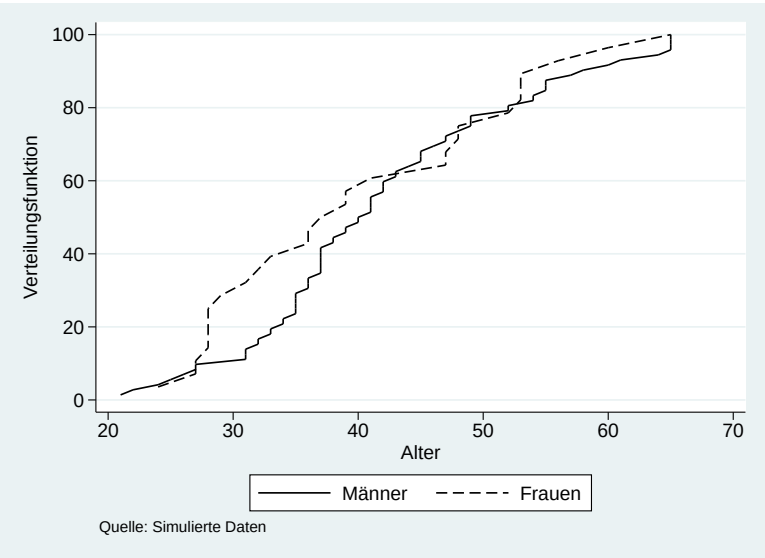
Die empirische Verteilungsfunktion kann auch bei Merkmalen mit vielen Ausprägungen verwendet werden (Beispiel: Alter der Befragten, siehe Tabelle weiter vorne).



Quelle: Simulierte Daten

Empirische Verteilungsfunktion IV

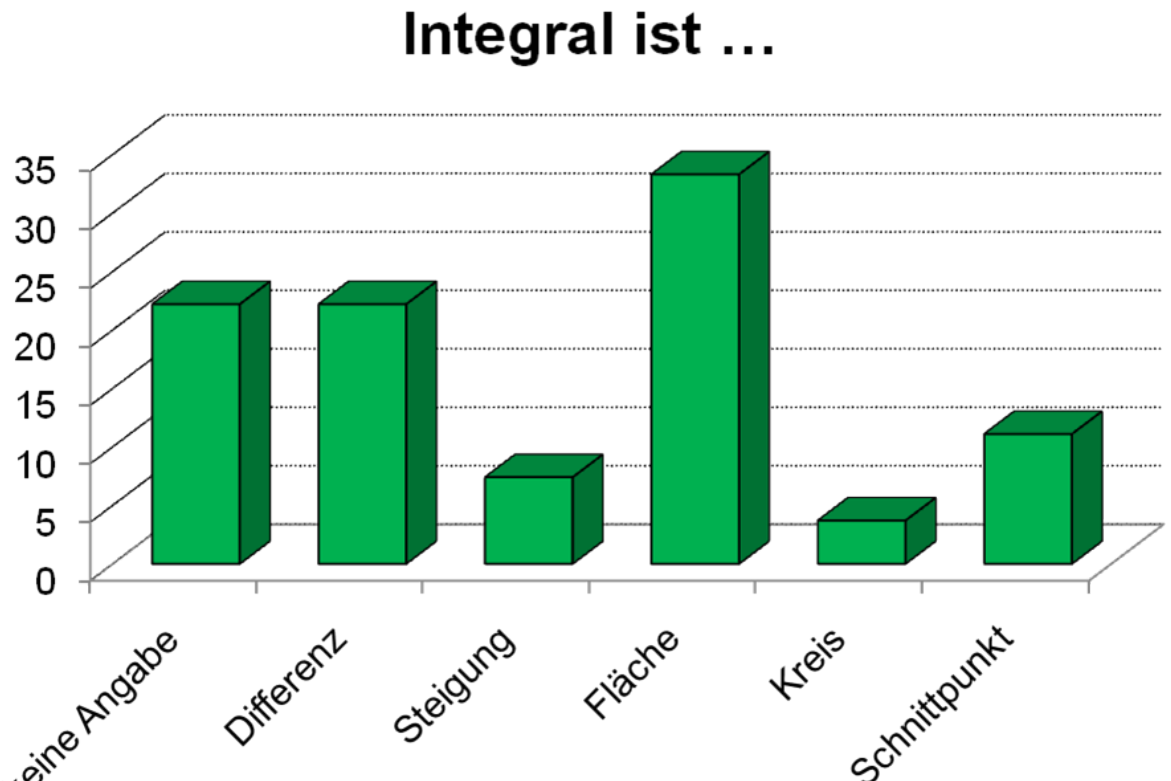
Die kumulierte Verteilungsfunktion eignet sich auch zum Vergleich der Verteilung von zwei oder mehr Gruppen (Beispiel: Alter):



Ist die Verteilung am Anfang steil, heißt das, dass mehr Fälle im unteren Wertebereich liegen, also niedrige Werte aufweisen.

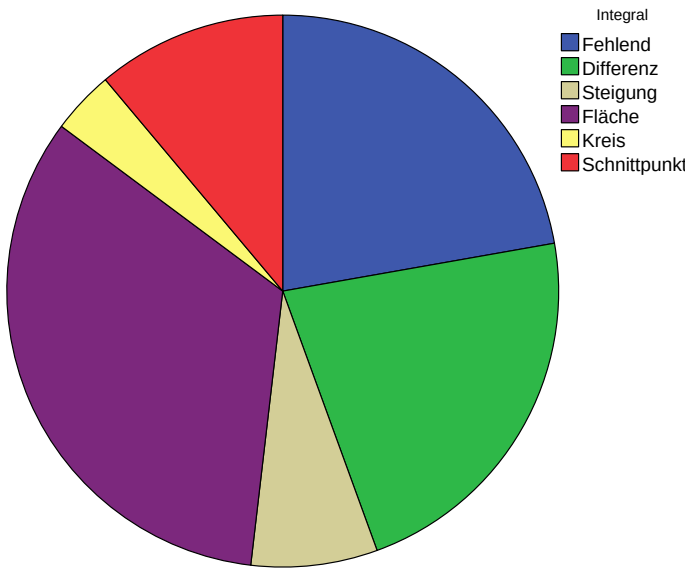
Nicht empfohlene Diagramme I

Der beliebte Pseudo-3-D-Effekt verzerrt die Daten und wird von Experten einhellig abgelehnt. Er hat sich nur aufgrund der Unfähigkeit bei gleichzeitiger Marktdominanz eines großen US-amerikanischen Softwareunternehmes verbreitet. In schriftlichen Arbeiten führt er bei mir zur einer Abwertung um mindestens 0,3 Notenpunkte.



Nicht empfohlene Diagramme II

Die Anteilswerte werden in Kreiswinkel übersetzt (Anteil von 360 Grad). Grundsätzlich hat dieser Diagrammtyp in Präsentationen nichts zu suchen, da das menschliche Auge die (relative) Größe von Winkeln nur grob beurteilen kann.



Graphische Darstellung von Verteilungen, Teil II: Ordinale und metrische Merkmale mit vielen Ausprägungen

Möglichkeit I: Stamm-Blatt-Diagramm

Oft auch Englisch: Stem-and-leaf-display (nach seinem Erfinder, John W. Tukey). Hier in der Darstellung der Software SPSS (Daten: Alter).

```
  4.00        2 .  1244
 11.00        2 .  56777788889
 12.00        3 .  111122233344
 23.00        3 .  55555666667777777889999
 13.00        4 .  0011111222334
 14.00        4 .  55567777888999
  8.00        5 .  22233344
  6.00        5 .  555678
  4.00        6 .  0014
  5.00        6 .  55555
Stem width:   10.00
Each leaf:     1 case(s)
```

Quelle: Simulierte Daten.

Konstruktion des Stamm-Blatt-Diagramms

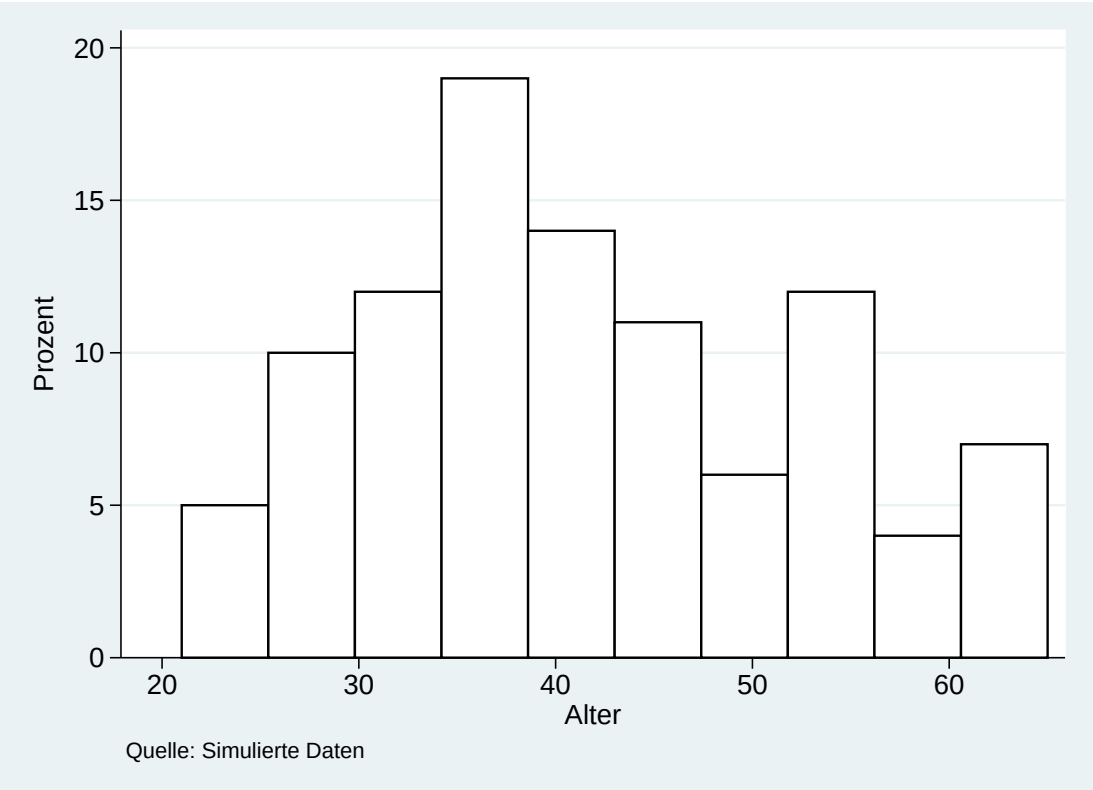
Eine mögliche Regel zur Konstruktion lautet:

1. Der Datenbereich wird in Intervalle gleicher Breite eingeteilt, die das 0,5 oder 1-fache einer Potenz von 10 betragen. Die erste Ziffer der Werte wird links als „Stamm“ abgetragen.
   In unserem Beispiel: Die Intervalle haben die Breite 5 (das 0,5-fache von 10^1).
2. Die Werte werden, wenn erforderlich, auf die Stelle gerundet, die nach den Ziffern des Stammes kommt (in unserem Beispiel: nicht nötig). Diese gerundeten Werte werden dann der Größe nach geordnet als „Blätter“ rechts vom Stamm abgetragen.

Abwandlungen dieser Regel sind möglich (siehe Beispiel vorherige Seite: die Häufigkeiten in der jeweiligen Zeile werden angegeben), die Grundidee ist immer die gleiche.

Möglichkeit II: Das Histogramm

Vor allem metrische Merkmale werden häufig anhand von Histogrammen visualisiert.

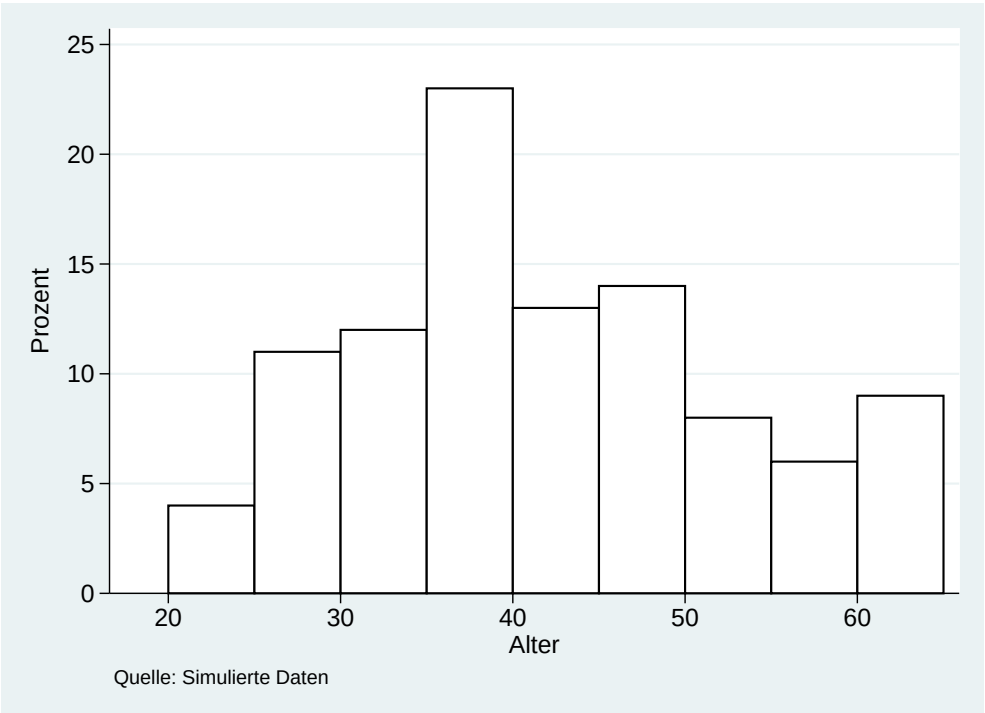


Konstruktion von Histogrammen

1. Die Daten werden in Klassen eingeteilt, zweckmäßigerweise von gleicher Klassenbreite. In der Regel werden rechtsoffene Intervalle gewählt, im Beispiel:
   21,0 bis $<$ 25,4
   25,4 bis $<$ 29,8
   29,8 bis $<$ 33,2
   usw. Zwingend ist die Wahl rechtsoffener Intervalle nicht!
2. In der Graphik werden über den Klassen Säulen abgetragen, deren Fläche proportional zu den (absoluten oder relativen) Häufigkeiten ist. (Bei gleicher Klassenbreite heißt dies auch: Die Höhe der Säule ist proportional zu den absoluten oder relativen Häufigkeiten.)
3. Um zu kennzeichnen, dass die Klassen aneinander grenzen (und dass man im Grunde von einem stetigen Merkmal ausgeht), berühren die Balken einander.

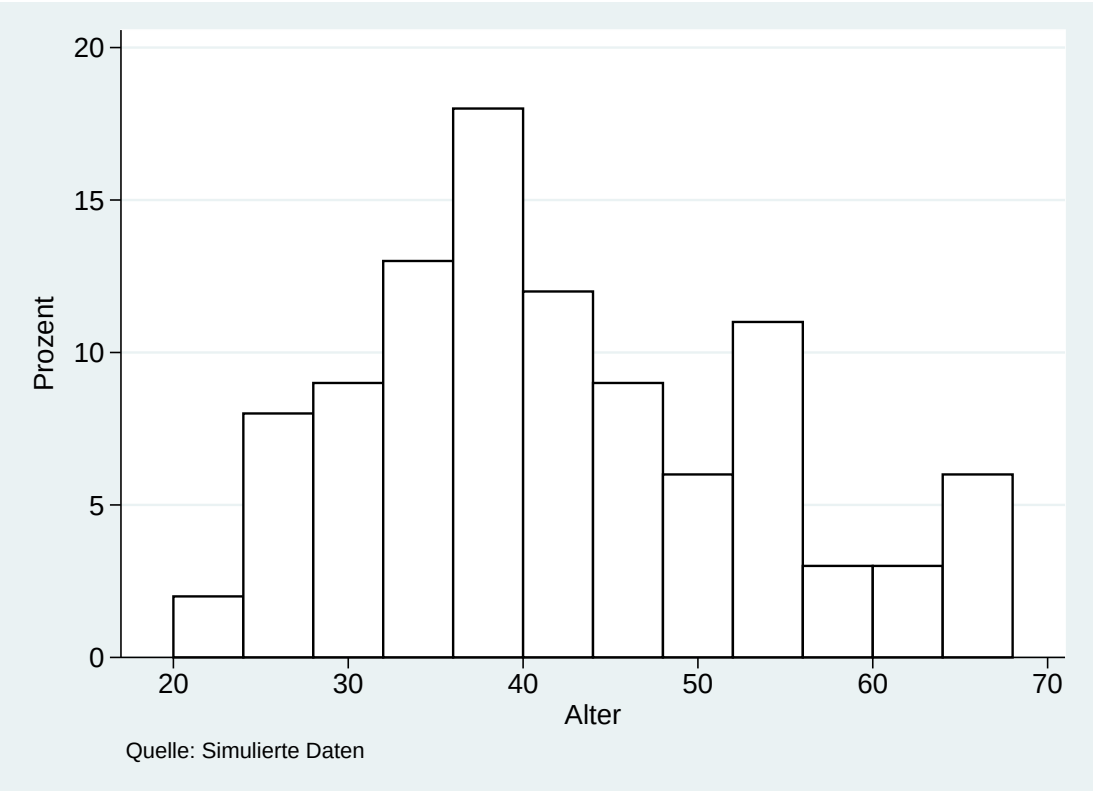
Probleme von Histogrammen

Die Wahl des Startpunktes der Klassenbildung (also: der untersten Klassengrenze) ebenso wie die Breite der Klassen wird in der Regel von Software getroffen. Da beides das Aussehen des Histogramms beeinflusst, sollte man die Vorgaben nicht unbesehen übernehmen.
Die gleichen Daten mit Startpunkt = 20 und Klassenbreite = 5 führen zu einem Histogramm, dessen Form dem Stamm-Blatt-Diagramm entspricht.



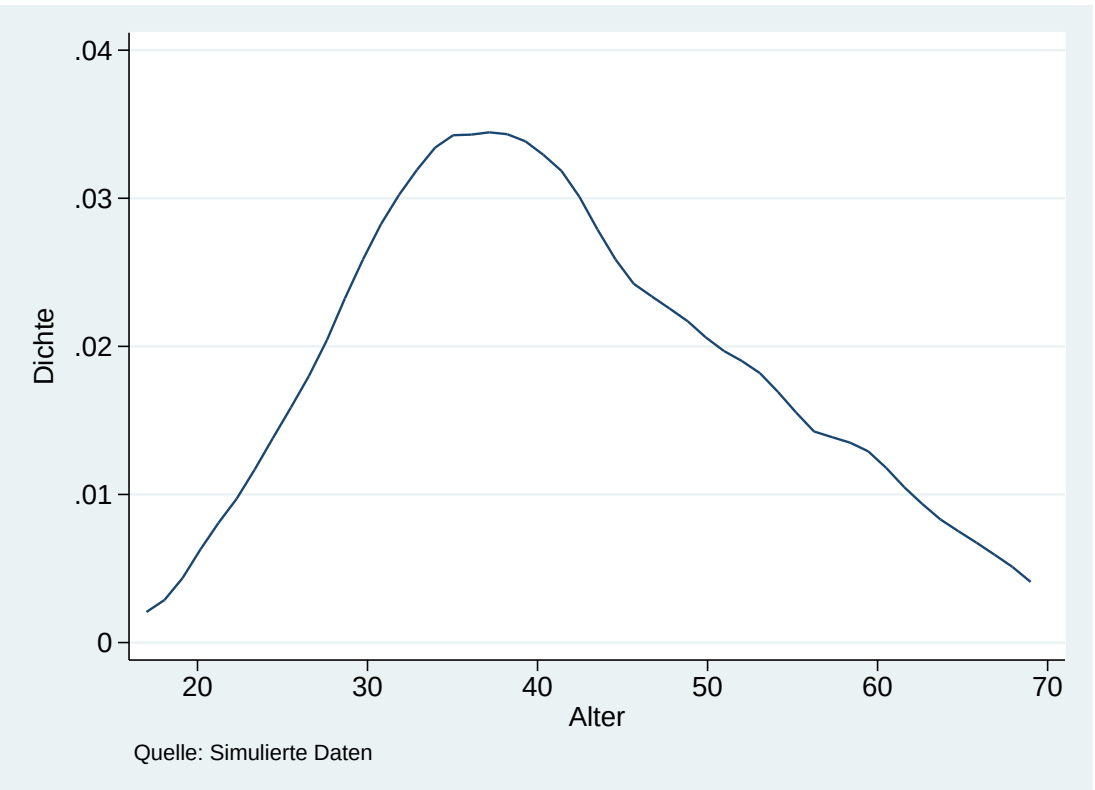
Die gleichen Daten, nochmals anders

Startpunkt = 20, Klassenbreite = 4
→ deutlicher ausgeprägte Gipfel rechts (siehe auch erstes Beispiel)



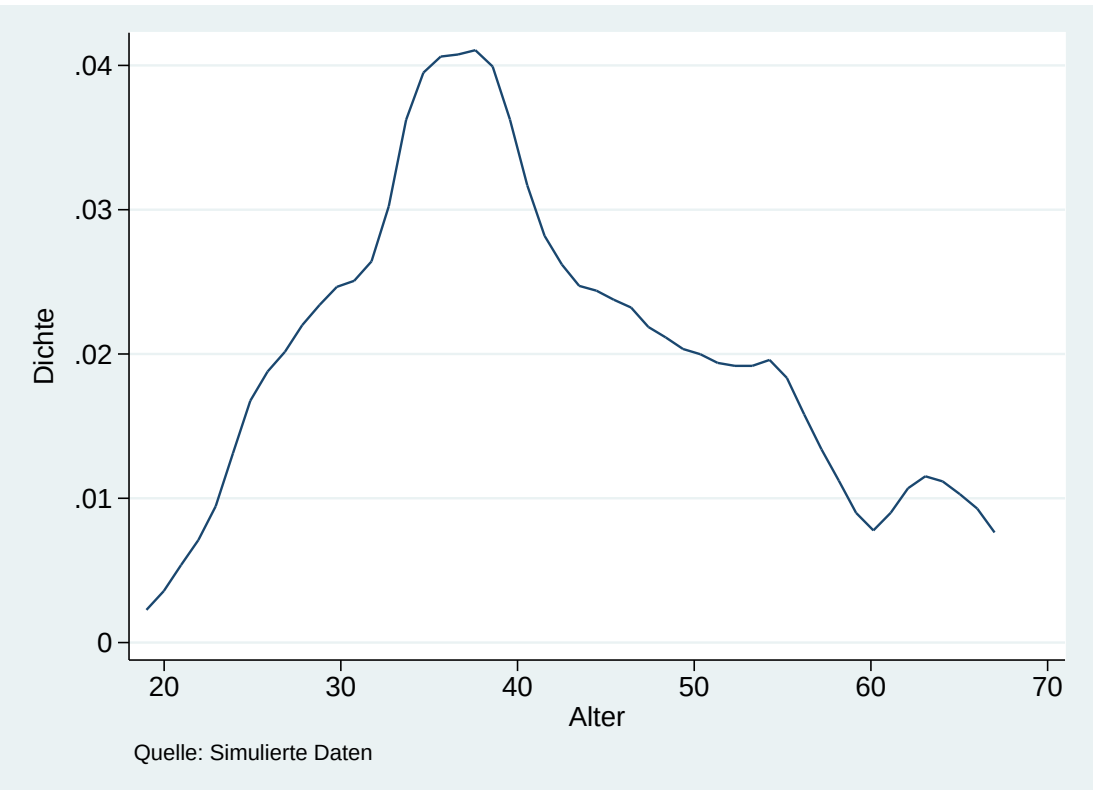
Möglichkeit III: Geglättete Histogramme

Soweit es sich um (prinzipiell) stetige Merkmale handelt, widerspricht das (diskret aussehende) Histogramm dem Charakter der Daten. Geglättete Histogramme (üblicherweise auf der Grundlage sog. Kern-Dichte-Schätzung) entsprechen diesem Charakter eher.



Möglichkeit III: Geglättete Histogramme

Auf die Konstruktion der Kern-Dichte-Schätzer kann aus Zeitgründen nicht eingegangen werden. Grundsätzlich ist zu beachten, dass auch hier (durch die sog. Bandbreite; nachfolgend: halbe Bandbreite der vorherigen Graphik) das Aussehen der Verteilung beeinflusst werden kann.



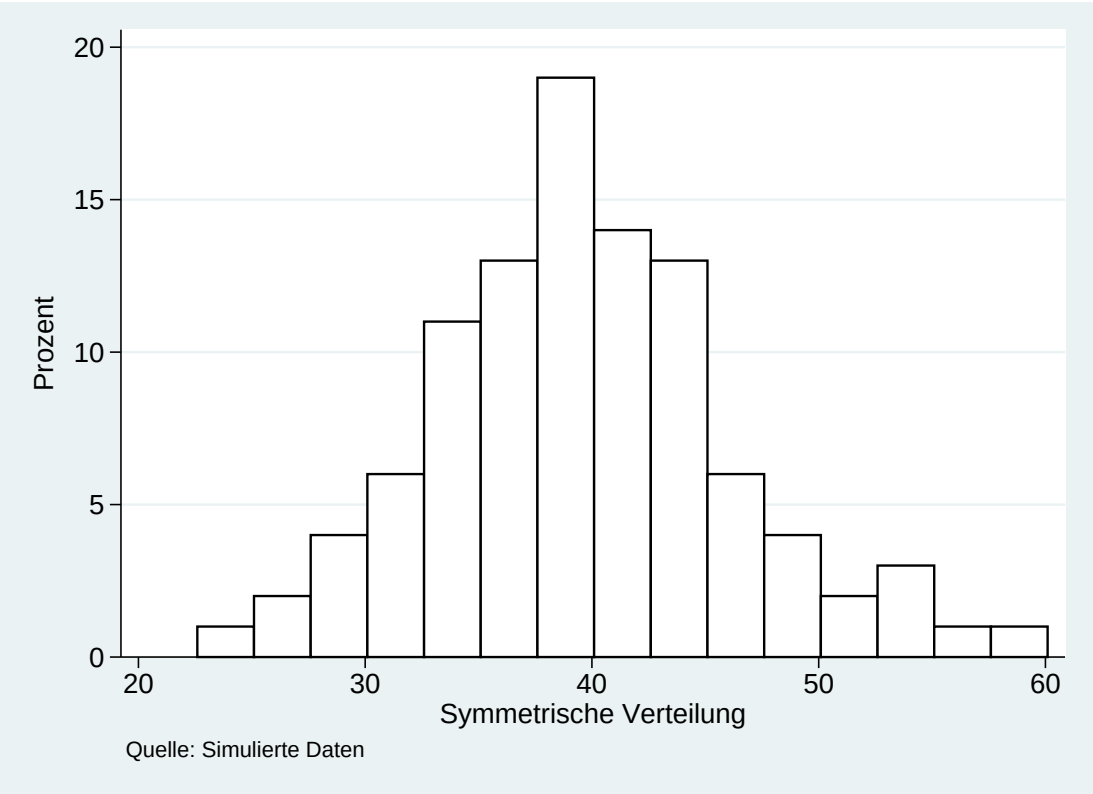
Erste abschließende Bemerkungen zu Graphiken

Aufgepasst: Die vorgestellten Visualisierungen von Verteilungen lassen die einzelnen Datenwerte immer mehr zurücktreten.

1. Im Stamm-Blatt-Diagramm lassen sich (jedenfalls bei kleineren Datensätzen) noch die einzelnen Datenwerte erkennen (bei großen Untersuchungen mit mehr als einigen Hundert Fällen kann es aber Probleme geben).
2. Beim Histogramm kann man nicht erkennen, wie sich die Datenwerte innerhalb der einzelnen Klassen verteilen. Auch sind die Grenzen zwischen den einzelnen Klassen oft nur annäherungsweise erkennbar.
3. Bei geglätteten Verteilungen lässt sich nur mehr grob abschätzen, in welchen Wertebereichen mehr oder weniger Fälle liegen.

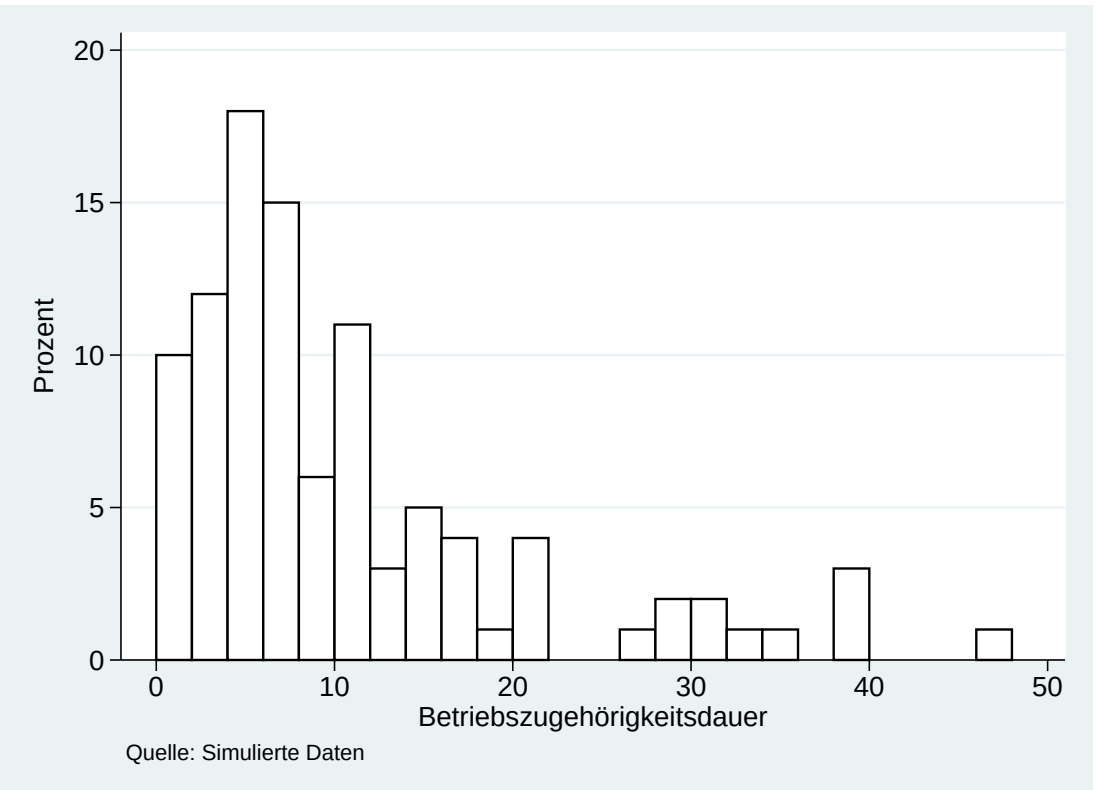
Verteilungsformen I: Symmetrische Verteilung

Verteilungen lassen sich häufig anhand ihrer Form charakterisieren. Eine symmetrische Verteilung hat einen Gipfel in der Mitte und links und rechts davon etwa gleich viele Datenwerte.



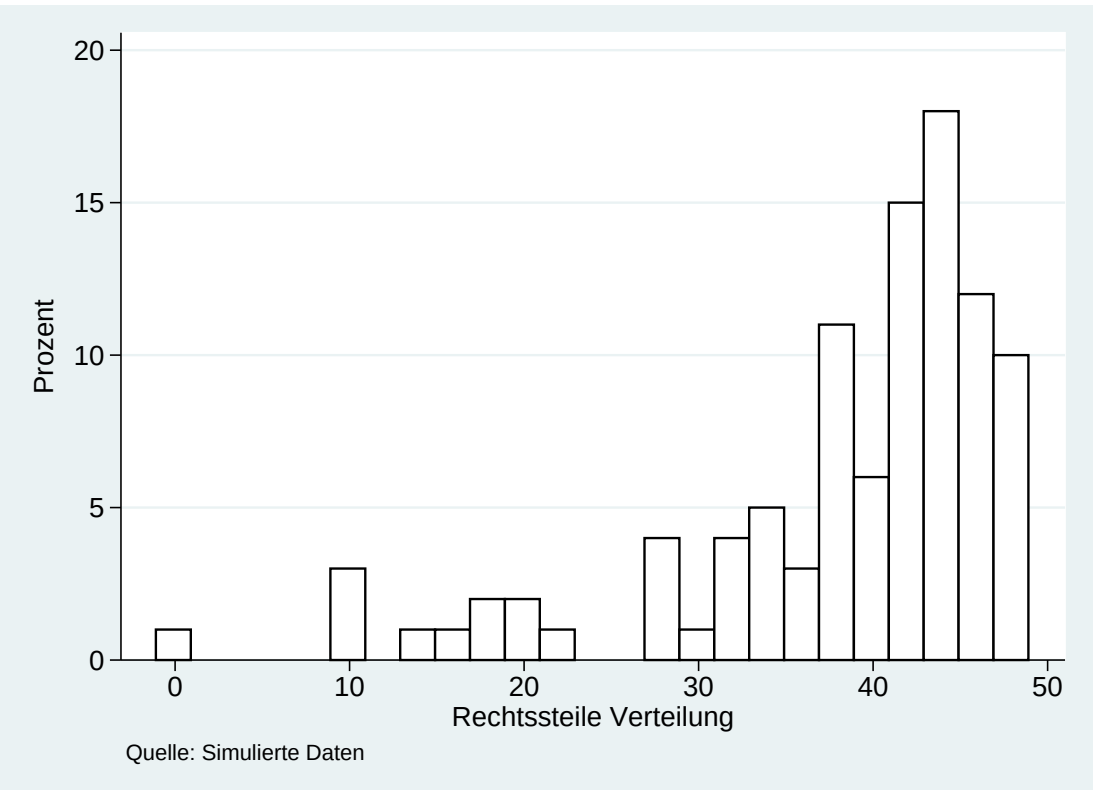
Verteilungsformen II: Linkssteile (rechtsschiefe) Verteilung

Bei einer linkssteilen (oder rechtsschiefen) Verteilung sind die Daten eher links konzentriert, mit der Folge, dass die größten Datenwerte eher weit weg vom Zentrum der Daten liegen. Dieser Verteilungstyp ist relativ häufig.



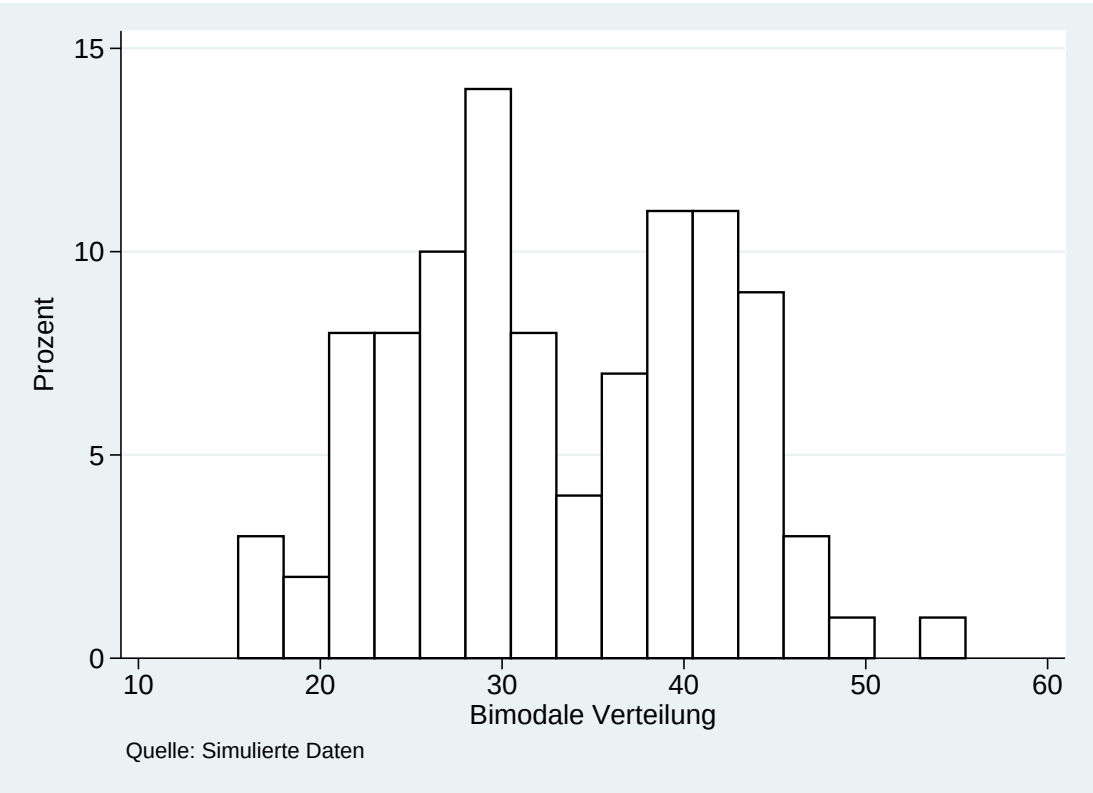
Verteilungsformen III: Rechtssteile (linksschiefe) Verteilung

Bei einer rechtssteilen (oder linksschiefen) Verteilung sind die Daten eher recht konzentriert, mit der Folge, dass die niedrigsten Datenwerte eher weit weg vom Zentrum der Daten liegen. Dieser Verteilungstyp ist (in den Sozialwissenschaften) eher selten.



Verteilungsformen IV: Zweigipflige (bimodale) Verteilung

Verteilungen mit zwei Gipfeln deuten darauf hin, dass es möglicherweise zwei Gruppen mit unterschiedlichen Eigenschaften in den Daten gibt.



Epilog

Wir haben nur einige einfache Möglichkeiten der Visualisierung von Verteilungen besprochen. Eine weitere Möglichkeit folgt in der nächste Vorlesung. Damit ist das Spektrum aber noch nicht erschöpft.

Auch die hier besprochenen Möglichkeiten sind, wie wir gesehen haben, nicht voraussetzungslos und sollten nicht unbedacht eingesetzt werden.