JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

4. Einheit

Plan für heute

- **Univariate Datenanalyse: Streumaße (Dispersionsmaße)**

Lernziele

- Kenntnis von Streumaßen der univariaten Statistik
- Bestimmung und Berechnung von Streumaßen der univariaten Statistik: Spannweite, Interquartilsabstand, Varianz, Standardabweichung
- Kenntnis von Normalverteilung und Formmaßen

Statistische Kennwerte

- Statistische Kennwerte beschreiben spezifische Eigenschaften einer empirischen Merkmalsverteilung
- Unterscheidung Lage- und Streu(ungs)maße
- Lagemaße allein nicht ausreichend, um Daten angemessen zu beschreiben

Statistische Kennwerte: Streumaße

- **Streumaße** (auch Dispersionsmaße) **beantworten Fragen wie**
  - Über welchen Bereich erstrecken sich die Beobachtungen?
  - Wie stark unterscheiden sich die Einzelwerte voneinander?
  - Wie stark „streuen“ die Werte? Wie groß sind die Unterschiede in der Merkmalsausprägung zwischen den Beobachtungseinheiten?
  - Wie groß ist die durchschnittliche Abweichung vom Mittelwert?

Streumaße

- Verteilung der Einzelwerte kann sehr stark „auseinandergezogen“ sein (Werte sind sehr ungleich, stärkere Streuung)
- Verteilung der Einzelwerte kann „schmal“ sein (Werte sind eher gleich, geringere Streuung)

Beispiel mit Übung

Beispiel: Lebenszufriedenheit auf einer Skala von 0 („ganz unzufrieden“) bis 10 („ganz zufrieden“)

Übung: Bitte berechnen Sie für beide Gruppen („Stadt“, „Land“) Modus, Median und arithmetisches Mittel.

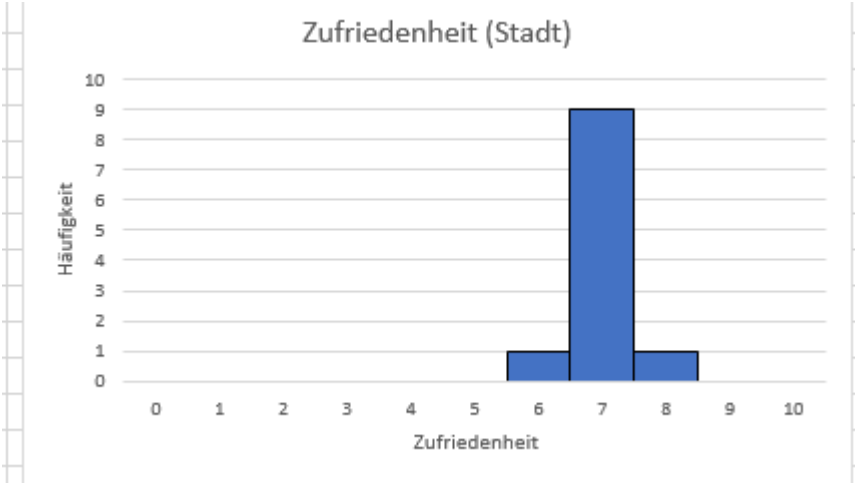
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Stadt	6	7	7	7	7	7	7	7	7	7	8
Land	1	2	7	7	7	7	8	8	10	10	10

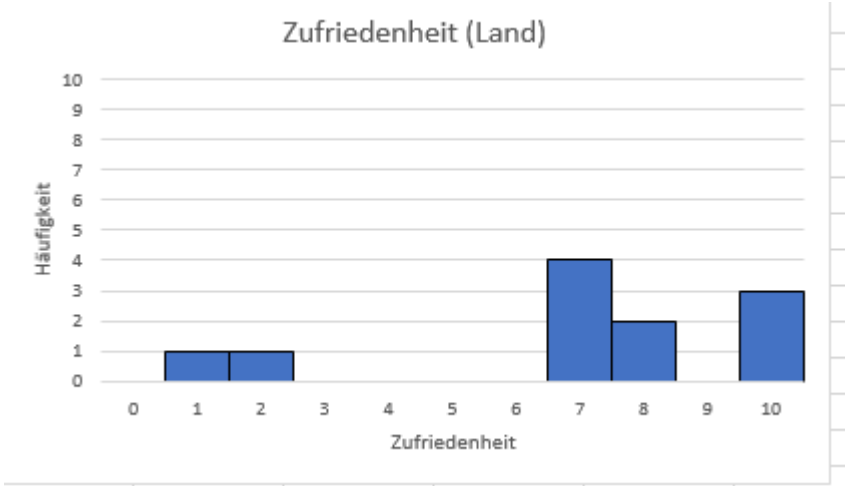
Übung

Modus, Median und arithmetisches Mittel liegt jeweils bei 7

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Stadt	6	7	7	7	7	7	7	7	7	7	8
Land	1	2	7	7	7	7	8	8	10	10	10

Aber: Gruppen unterscheiden sich trotzdem → Wichtig, über Streuung in den Daten informiert zu sein





Spannweite (Variationsbreite)

Auch: Range

- Definition: Differenz zwischen dem größten x_{max} und dem kleinsten Wert x_{min} der Daten (zwischen Maximum und Minimum)
- beschreibt die Größe des Bereichs, innerhalb derer sich die Werte befinden

- Berechnung: $V = x_{max} - xmi_n$
- Aber:
  - V ist anfällig gegenüber Extremwerten, „Ausreißern“
  - Berücksichtigt nur zwei Werte, alle anderen werden vernachlässigt

Spannweite: Beispiel mit Übung

Einkommensverteilung:

x_{max} = 1000€; x_{min} = 900€

- Berechnung: $V = x_{max} - x_{min}$=

Spannweite: Beispiel mit Übung

Einkommensverteilung, x_{max} = 1000€; x_{min} = 900€

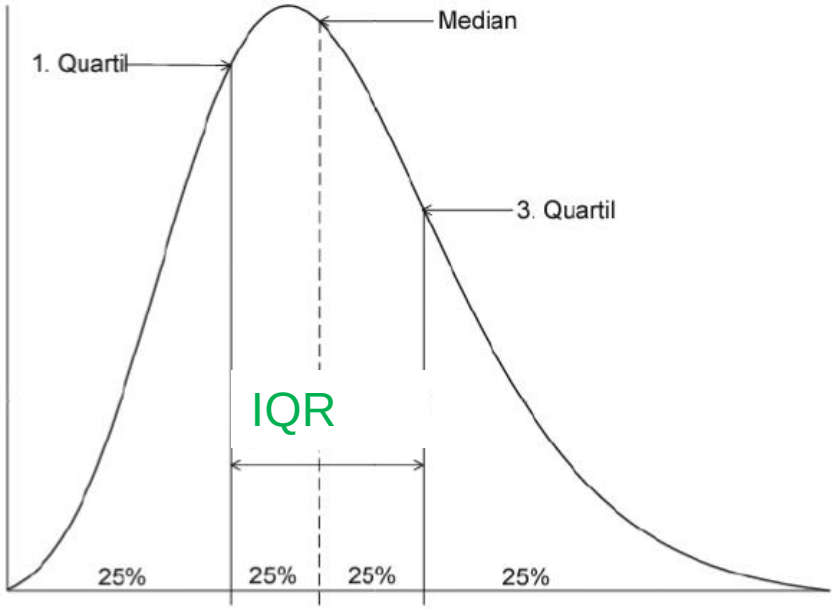
- Berechnung: $V = x_{max} - x_{min}$ = 1000 – 900 = 100

Interquartilsabstand

IQR

- Definition: Differenz zwischen dem oberen Quartil und dem unteren Quartil
- Berechnung: $\mathrm{IQR} = Q_{0.75} - Q_{0.25}$
- Berücksichtigt mittlere 50% der Verteilung, weniger anfällig ggü. Extremwerten
- vor allem sinnvoll, wenn der Kernbereich einer Häufigkeitsverteilung - d.h. wenn die zentral gelegenen 50% der Merkmalsausprägungen – interessieren
- Je größer der IQR desto stärker streuen die Beobachtungen

Wdh. Quartile



- 25% der Werte sind kleiner oder gleich und 75% der Werte sind größer oder gleich dem 1. Quartil ($Q_{0.25}$)
- Das zweite Quartil ist der Median. Der Median ($Q_{0.5}$) zerteilt eine Verteilung in zwei gleich große Hälften.
- 75% der Werte sind kleiner oder gleich und 25% der Werte größer oder gleich dem 3. Quartil ($Q_{0.75}$).

Beispiel IQR

- Wie lautet im Beispiel der IQR?

Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
$\sum$	*11*	*100*	*100*

Beispiel IQR

- Wie lautet im Beispiel der IQR → 14-11=3 Semester

Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
∑	*11*	*100*	*100*

Beispiel IQR

- Wie lautet im Beispiel der IQR → 14-11=3 Semester

→ Interpretation: 50% der Befragten haben zwischen 11 und 14 Semester für ihr Studium benötigt

Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
∑	*11*	*100*	*100*

Varianz

- Zählt zusammen mit der Standardabweichung zu den am Häufigsten verwendeten Streuungsmaßen
- Ermittelt die **durchschnittliche „Abweichung“** der Ausprägungen eines Merkmals vom arithmetischen Mittelwert aller Merkmalsausprägungen
- Definition: Summe der quadrierten Abweichungen der Merkmalswerte vom arithmetischen Mittelwert, dividiert durch die Anzahl der Beobachtungen
- Notation: σ^2 (für Grundgesamtheiten) oder s^2 (für Stichprobe)

- Warum quadriert? (Summe der Abweichungen von $\bar{x}$ ist immer 0)
- Voraussetzung ist mindestens (pseudo-)metrisches Skalenniveau

Beispiel mit Übung

- In einer Klausur haben *n*= 5 Prüflinge jeweils die folgenden Anzahl richtig gelöster Aufgaben erzielt:

$$x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 5 \; und \; x_5 = 5$$

- Wie streuen die Werte um das arithmetische Mittel $\bar{x}$?
- Berechnen Sie das arithmetische Mittel :

Beispiel mit Übung

- In einer Klausur haben *n*= 5 Prüflinge jeweils die folgenden Anzahl richtig gelöster Aufgaben erzielt:

$$x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 5 \; und \; x_5 = 5$$

- Berechnen Sie das arithmetische Mittel

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(2+3+4+5+5) = \frac{19}{5} = 3.8$$

Beispiel mit Übung

- Formel Varianz „normal“ bzw. bei Grundgesamtheit
- Summe der quadrierten Abweichungen vom arithmetischen Mittel, geteilt durch n

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Beispiel mit Übung

- In einer Klausur haben *n*= 5 Prüflinge jeweils die folgenden Anzahl richtig gelöster Aufgaben erzielt:

$$x_1 = 2, x_2 = 3, x_3 = \ 4, x_4 = 5 \ und \ x_5 = 5$$

- Berechnen Sie das arithmetische Mittel

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(2+3+4+5+5) = \frac{19}{5} = 3.8$$

- Die Abweichungen vom arithmetischen Mittel werden als Differenzen berechnet:
- 2 -3.8 = -1.8; 3 -3.8 = -0.8; 4 -3.8 = 0.2, 5 -3.8 = 1.2 und 5 -3.8 = 1.2

Beispiel mit Übung

- Formel Varianz „normal“ bzw. bei Grundgesamtheit

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$= \frac{1}{5}(-1{,}8^2 - 0{,}8^2 + 0{,}2^2 + 1{,}2^2 + 1{,}2^2) = \frac{6.8}{5} = 1.36$$

Beispiel mit Übung

- Formel Varianz „normal“ bzw. bei Grundgesamtheit

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$= \frac{1}{5}(-1{,}8^2 - 0{,}8^2 + 0{,}2^2 + 1{,}2^2 + 1{,}2^2) = \frac{6.8}{5} = 1.36$$

Ergebnis: Quadrierte Anzahl der Aufgabenlösungen?
Vereinfachung: Ziehen der Quadratwurzel aus der Varianz → Standardabweichung

$$\sqrt{1.36} = 1.17$$

Achtung: wichtige Unterscheidung

- Daten der Grundgesamtheit oder Daten einer Stichprobe?
- Unterscheidung zwischen empirischer Varianz σ^2
  und korrigierter Varianz s^2
- Anpassung durch Korrekturfaktor $\frac{1}{n-1}$

Achtung: wichtige Unterscheidung

- Daten der Grundgesamtheit oder Daten einer Stichprobe?
- Unterscheidung zwischen empirischer Varianz σ^2 und korrigierter Varianz s^2
- Anpassung durch Korrekturfaktor $\frac{1}{n-1}$

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$s^2 = \frac{1}{n-1} * \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standardabweichung

- Varianz als „durchschnittliche quadrierte Abweichung vom arithmetischen Mittel“ nicht besonders „intuitiv“ verständlich bzw. interpretierbar

→ Standardabweichung: Wurzel der Varianz, gleiche Maßeinheit wie Ausgangsvariable

- Varianzberechnung wird aber als Zwischenschritt benötigt

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

„normal“, bei Grundgesamtheit/Vollerhebung

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Korrigierte Standardabweichung, bei Stichprobe

Standardabweichung, Beispiel Interpretation

1.000 Personen wurden nach ihren wöchentlichen Ausgaben befragt. Der Mittelwert liegt bei 200 Euro, die Standardabweichung liegt bei s = 70 Euro.

Was heißt das?

Standardabweichung, Beispiel Interpretation

1.000 Personen wurden nach ihren wöchentlichen Ausgaben befragt. Der Mittelwert liegt bei 200 Euro, die Standardabweichung liegt bei s = 70 Euro.

Was heißt das? → **die durchschnittliche Entfernung aller beobachteten Werte zum Mittelwert liegt bei 70 Euro**

Ergänzung: Formmaße

Neben Streu- und Lagemaßen gibt es auch noch (die weniger häufig berechneten) **Formmaße**

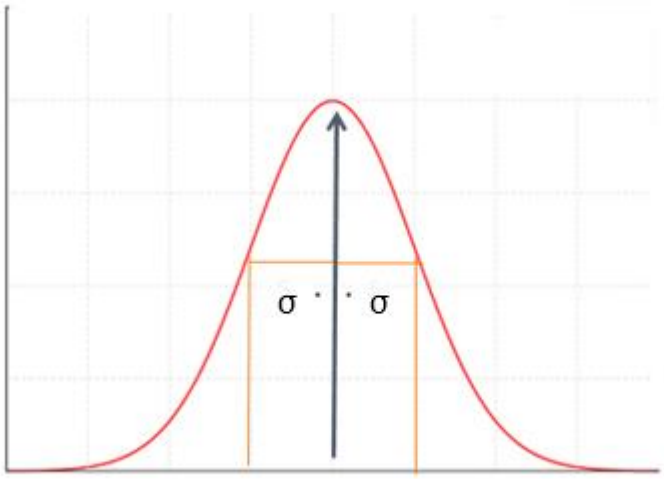
- **Schiefe** (engl.: skewness)
- **Wölbung** (Kurtosis)

Beide Maße beschreiben die Abweichung einer Verteilung von der sog. „Normalverteilung“

Normalverteilung

Deskriptiv:

- Beschreibt eine symmetrische Verteilungsform in Form einer Glocke („Gaußsche Glockenkurve“)
- Werte konzentrieren sich in der Mitte, treten mit größerem Abstand zur Mitte immer seltener auf



Normalverteilung

- Im „wirklichen Leben“, in der Natur: einige Merkmale treten normalverteilt in der Bevölkerung auf (IQ, Körpergröße)
- Inferenzstatistik: zentrales Modell für Wahrscheinlichkeitsverteilungen für kontinuierliche Zufallsvariablen, sog. „stetige Verteilungen“ (→ v.a. Statistik 2)

Ergänzung: Formmaße

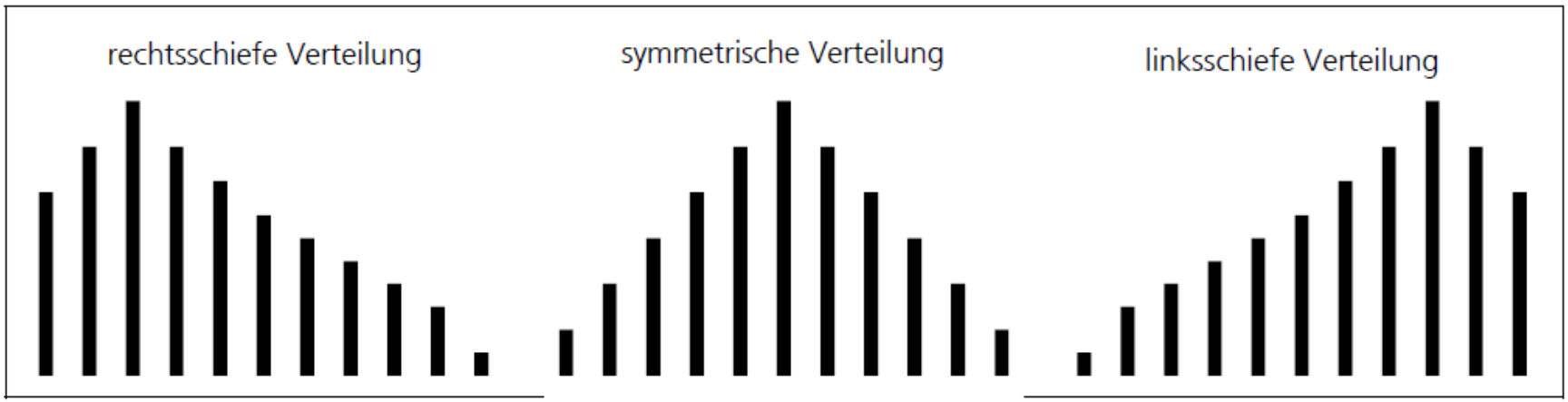
Formmaße Schiefe und Wölbung: **beschreiben die Abweichung einer Verteilung von der sog. „Normalverteilung“**

→ „horizontale“ Abweichung: Schiefe (engl.: skewness)

→ „vertikale“ Abweichung: Wölbung (Kurtosis)

Formmaße: Schiefe

Händische Berechnung sehr aufwändig, daher Statistikprogramm!



Quelle: Eigene Darstellung

$$\text{Schiefe} = \frac{\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)^3}{n}$$

Formmaße: Schiefe

Interpretation der Werte:

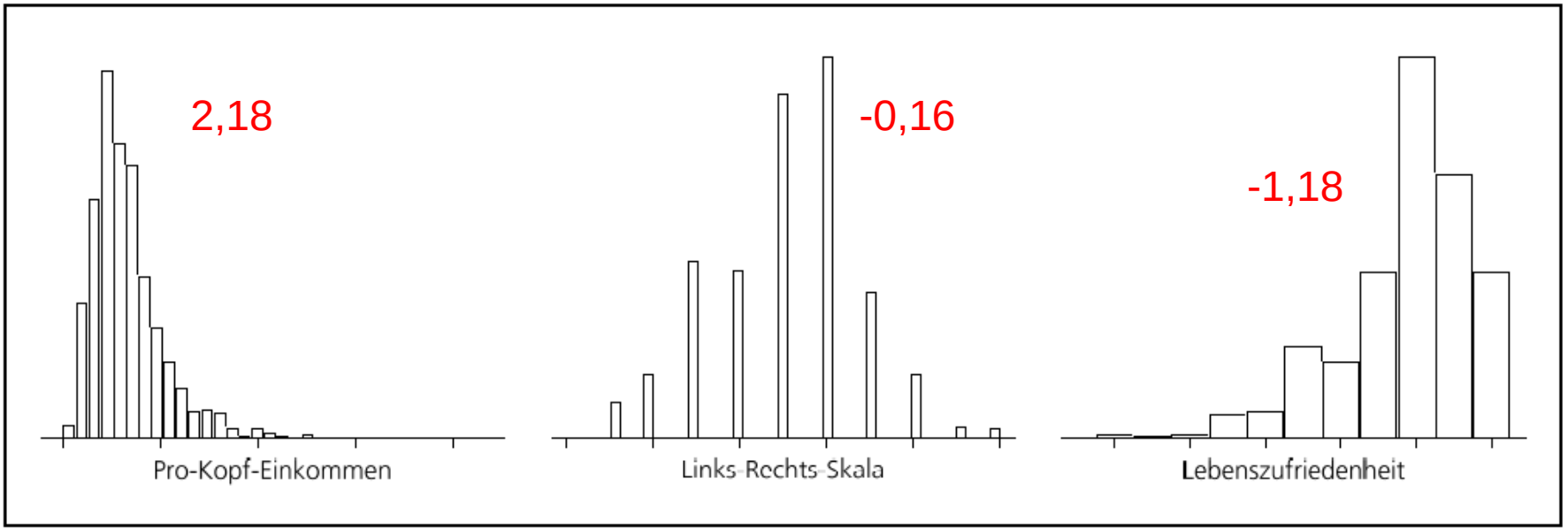
Schiefe < 0: linksschiefe Verteilung (rechtssteil/-gipflig)

Schiefe = 0: symmetrische Verteilung

Schiefe > 0: rechtsschiefe Verteilung (linkssteil)

Werte, deren Beträge > als 1 sind werden als deutliche Abweichung interpretiert

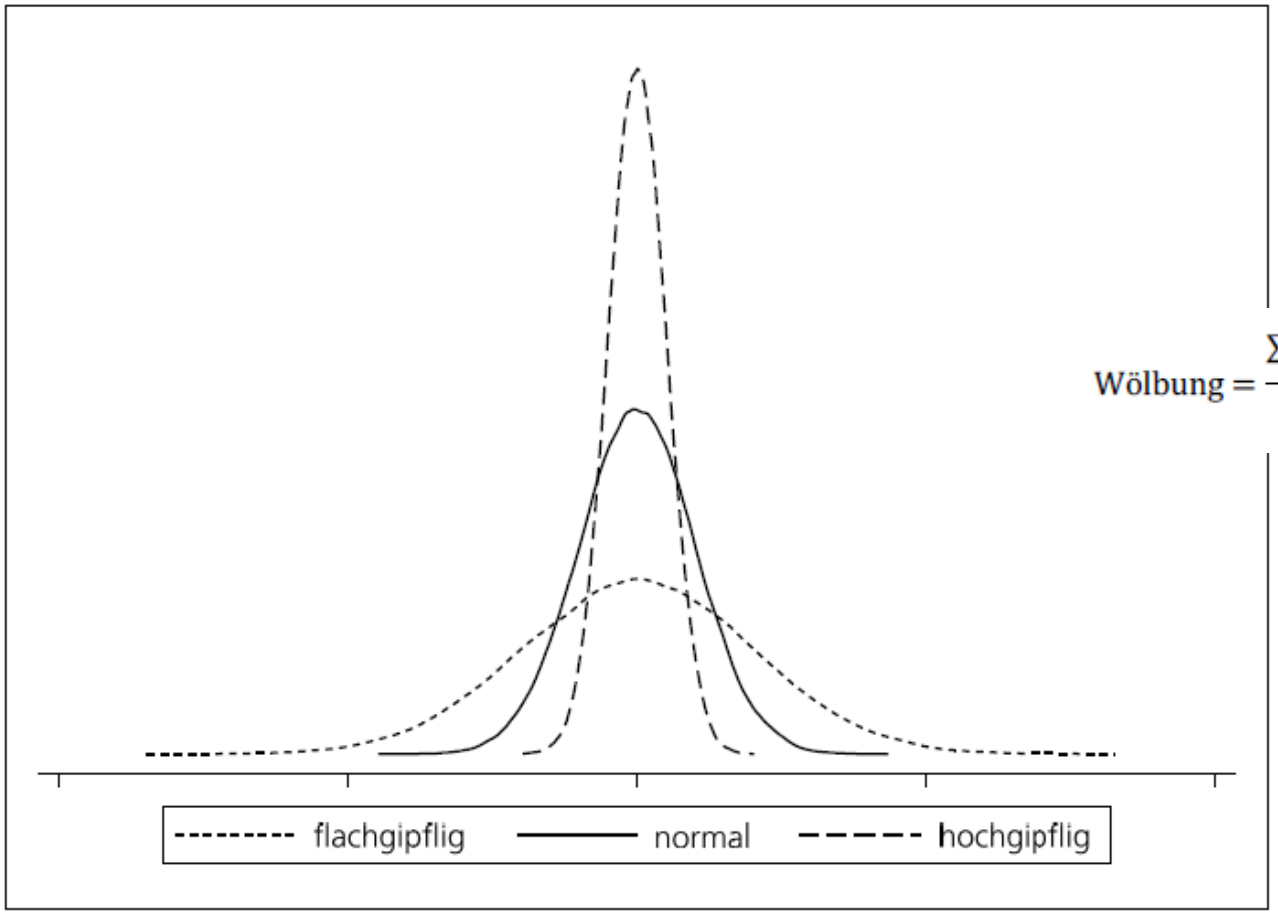
„Schiefe“ Verteilungen, Beispiele



Daten: ALLBUS 2016. Eigene Berechnungen

Ergänzung: Wölbung

„vertikale“ Abweichung von der Normalverteilung: Wölbung (Kurtosis)



$$\text{Wölbung} = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4}{n} - 3$$

Quelle: Eigene Darstellung

Formmaße: Wölbung

Interpretation der Werte:

– Kurtosis < 0: flachgipflige Verteilung

– Kurtosis = 0: Normalverteilung

– Kurtosis > 0: hochgipflige Verteilung

Lernziele

- Kenntnis von Streumaßen der univariaten Statistik
- Bestimmung und Berechnung von Streumaßen der univariaten Statistik: Spannweite, Interquartilsabstand, Varianz, Standardabweichung
- Kenntnis von Normalverteilung und Formmaßen

Übung: Zusammenfassung Kennwerte

Bitte ergänzen Sie folgende Tabelle

Kennwert	Ab Skalenniveau…
Modus	nominal
Median	
Arithmetisches Mittel	
Variationsweite	
Varianz	
Standardabweichung	
Schiefe	
Wölbung	