Universität der Bundeswehr München

Fakultät für Humanwissenschaften
Sozialwissenschaftliche Methodenlehre
Prof. Dr. Daniel Lois

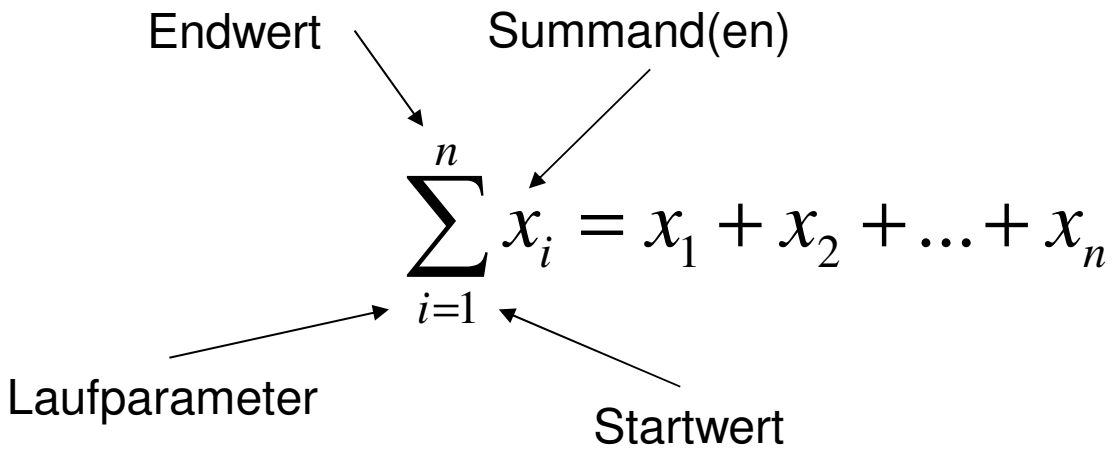
# Deskriptive Statistik

Stand: April 2015 (V2)

# Inhaltsverzeichnis

# Notation

- n = Anzahl der Untersuchungseinheiten
- X = eine Variable
- i = eine einzelne, aber keine bestimmte Untersuchungseinheit
- $x_i$ = der Wert der Variable X für Untersuchungseinheit i



# Messniveau

- Das Messniveau einer Variablen ist von zentraler Bedeutung dafür, welche statistischen Auswertungsverfahren für diese Variable sinnvoll bzw. zulässig sind
- Die Messgenauigkeit bzw. der Informationsgehalt der Daten steigt mit dem Messniveau an
- Die folgende Folie zeigt die Messung des Merkmals „Zufriedenheit“ mit ansteigendem Messniveau
- Wir unterscheiden im Folgenden hauptsächlich zwischen kategorialen Variablen (Nominalskala, Ordinalskala) und metrischen Variablen (Intervallskala, Verhältnisskala)

# Messniveau

**Beispiel: Drei Möglichkeiten der Erfassung von Zufriedenheit**

**Nominal:**

0 = unzufrieden
1 = zufrieden

**Ordinal:**

1 = sehr unzufrieden
2 = unzufrieden
3 = zufrieden
4 = sehr zufrieden

**Metrisch:**

0 = sehr unzufrieden bis
10 = sehr zufrieden

# Messniveau

- Bei einer **Nominalskala** handelt es sich um einen Satz rangmäßig nicht geordneter Kategorien (Beispiele: Geschlecht, Religionszugehörigkeit, Beruf, Familienstand)
- Es sind nur Aussagen über Gleichheit oder Verschiedenheit eines Merkmals möglich (Mann ungleich Frau, SPD-Wähler ungleich CDU-Wähler)
- Die Kategorien sind vollständig (d.h. schließen alle Fälle ein) und disjunkt (überschneidungsfrei)
- Nominale Variablen können beliebig viele Ausprägungen haben, bei zwei Ausprägungen spricht man dichotomen (zweistufigen) Variablen

# Messniveau

- Die **Ordinalskala** hat eine zusätzliche Eigenschaft: Die Ausprägungen der Variablen lassen sich sinnvoll in eine Ordnungsrelation bringen (z.B. Schulnoten, subjektive Schichteinstufung)
- Beispiel: Oberschicht > obere Mittelschicht > Mittelschicht > Arbeiterschicht
- Ordinales Messen informiert jedoch nicht über die Größe der Differenzen zwischen den Ausprägungen einer Variablen
- Zum Beispiel ist nicht bekannt, ob der Abstand zwischen „Oberschicht“ und „oberer Mittelschicht“ größer oder kleiner ist als der Abstand zwischen „oberer Mittelschicht“ und „Mittelschicht“

# Messniveau

- Bei der **Intervallskala** müssen zusätzlich die Abstände zwischen den einzelnen Ausprägungen einer Variablen gleich sein (Äquidistanz der Intervalle)
- Zum Beispiel beträgt die Differenz zwischen zwei Lebensjahren genau 12 Monate oder die Differenz zwischen zwei Tagen genau 24 Stunden – unabhängig davon, um welches Lebensjahr oder um welche Tage es sich handelt
- Existiert zusätzlich ein absoluter (invarianter) Nullpunkt (z.B. bei Variablen wie Alter oder Einkommen), spricht man von einer **Verhältnisskala**

# Messniveau

| | **Messniveau** | **(Zusätzliche) Eigenschaften** | **Aussagen möglich über** | **Beispiele** |
|---|---|---|---|---|
| Kategoriale Variablen | Nominalskala | Vollständige und disjunkte Kategorien | Gleichheit / Verschiedenheit | Geschlecht, Nationalität |
| | Ordinalskala | Ordnungsrelation | Größer / Kleiner | Schulnoten |
| Metrische Variablen | Intervallskala | Abstände definiert | Gleichheit von Differenzen | Geburtsjahr |
| | Verhältnisskala | Nullpunkt definiert | Gleichheit von Verhältnissen | Alter, Einkommen |

# Häufigkeitsverteilungen

- Im Folgenden wird dargestellt, wie kategoriale oder metrische Variablen tabellarisch bzw. grafisch dargestellt werden können
- Im Falle von nominal und ordinal skalierten Variablen besteht die Aufgabe darin, absolute und relative Häufigkeiten für die einzelnen Merkmalsausprägungen darzustellen, bei metrischen Variablen werden Häufigkeitsdichten betrachtet
- Diese Arbeitsschritte stehen im Rahmen einer explorativen Datenanalyse oft am Beginn der Datenauswertung

# Häufigkeitsverteilung

In den Daten vorhandene Ausprägungen

Absolute Häufigkeit

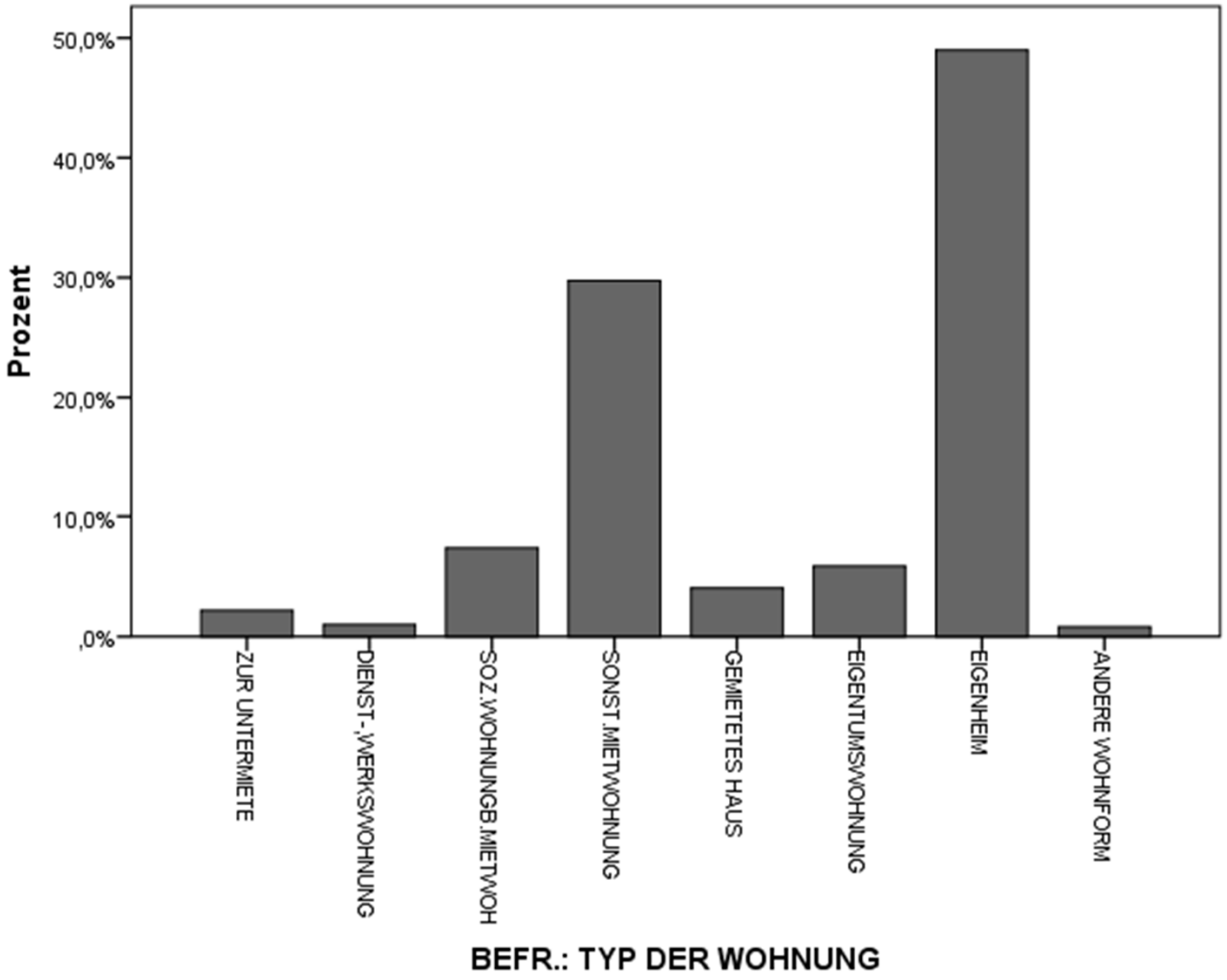
Relative Häufigkeit

Kumulierte relative Häufigkeit

**BEFR.: TYP DER WOHNUNG**

| | | Häufigkeit | Prozent | Gültige Prozent | Kumulative Prozente |
|---|---|---|---|---|---|
| Gültig | ZUR UNTERMIETE | 61 | 2,2 | 2,2 | 2,2 |
| | DIENST-, WERKSWOHNUNG | 28 | 1,0 | 1,0 | 3,2 |
| | SOZ.WOHNUNGB. MIETWOH | 208 | 7,4 | 7,4 | 10,5 |
| | SONST.MIETWOHNUNG | 838 | 29,7 | 29,7 | 40,3 |
| | GEMIETETES HAUS | 114 | 4,0 | 4,0 | 44,3 |
| | EIGENTUMSWOHNUNG | 165 | 5,9 | 5,9 | 50,2 |
| | EIGENHEIM | 1380 | 48,9 | 49,0 | 99,2 |
| | ANDERE WOHNFORM | 23 | ,8 | ,8 | 100,0 |
| | Gesamtsumme | 2817 | 99,9 | 100,0 | |
| Fehlend | KEINE ANGABE | 3 | ,1 | | |
| Gesamtsumme | | 2820 | 100,0 | | |

# Häufigkeitsverteilung



# Häufigkeitsverteilung

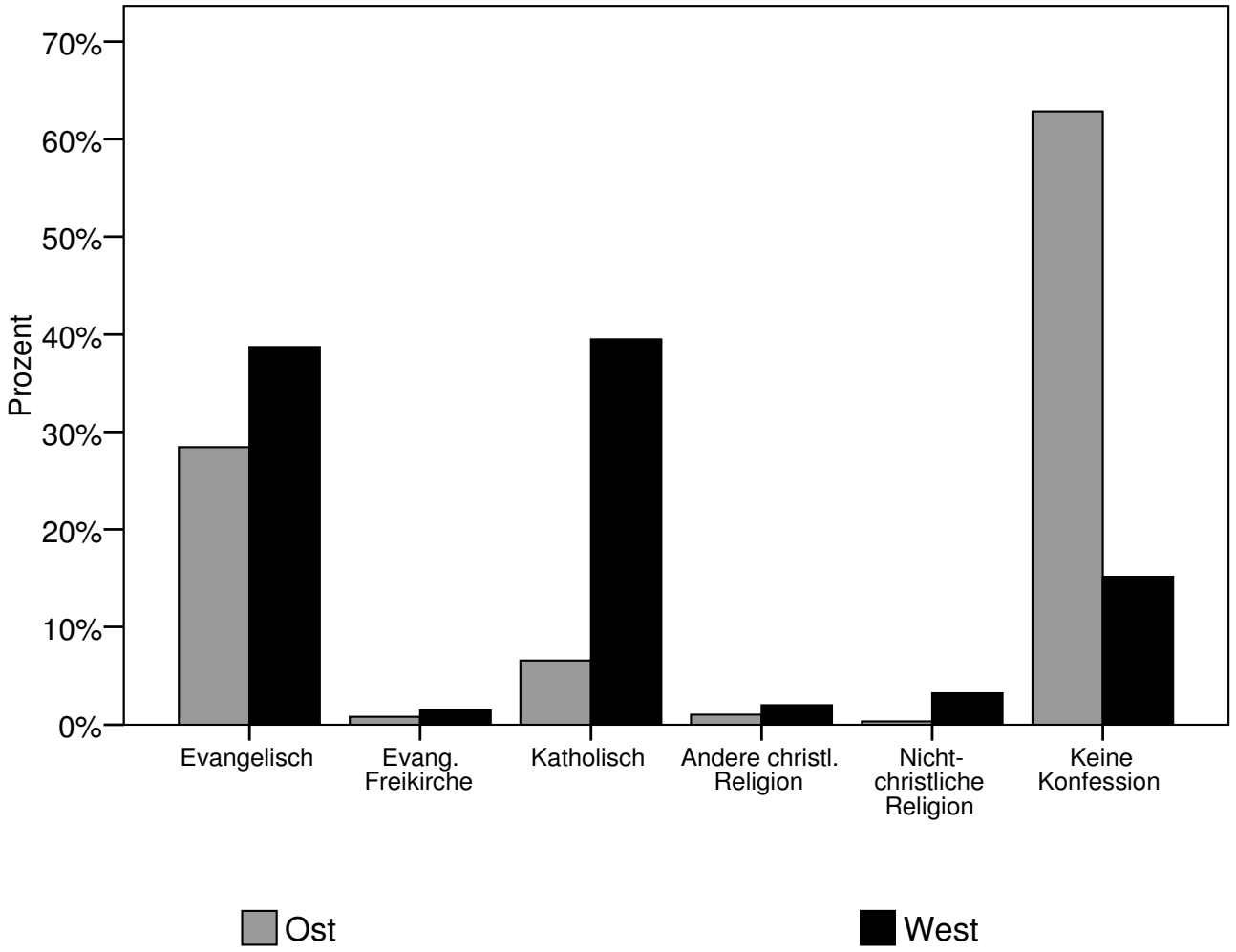
Klar strukturierte, übersichtliche Tabellen erstellen (Regel: Tinte sparen):

Tabelle: Häufigkeitsverteilung von Konfessionen in West- und Ostdeutschland

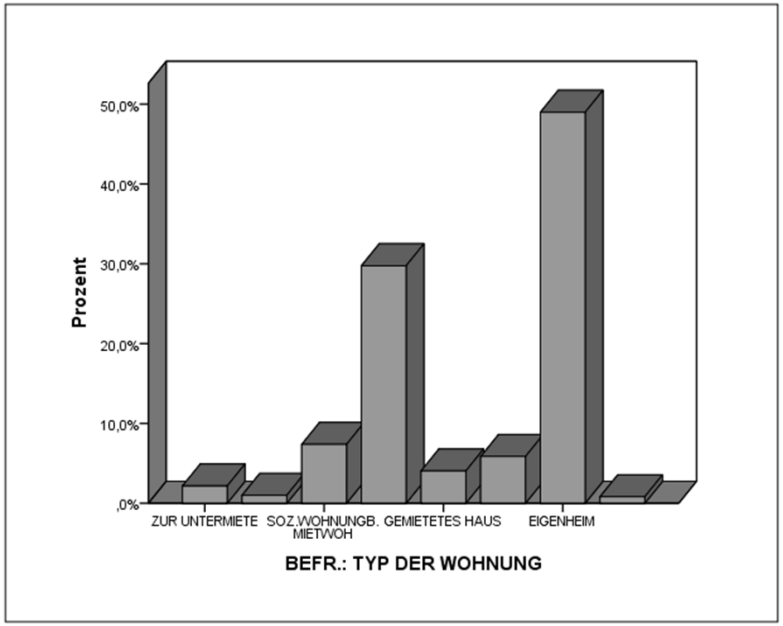
| | Westdeutschland | | Ostdeutschland | |
|---|---|---|---|---|
| | Häufigkeit | Prozent | Häufigkeit | Prozent |
| Katholisch | 759 | 39.5 | 58 | 6.6 |
| Evangelisch ohne Freikirchen | 744 | 38.7 | 251 | 28.4 |
| Evangelische Freikirchen | 28 | 1.5 | 7 | 0.8 |
| Andere christliche Religion | 38 | 2.0 | 9 | 1.0 |
| Nicht-christliche Religion | 62 | 3.2 | 3 | 0.3 |
| Keine Religionsgemeinschaft | 291 | 15.1 | 555 | 62.9 |
| Insgesamt | 1922 | 100.0 | 883 | 100.0 |

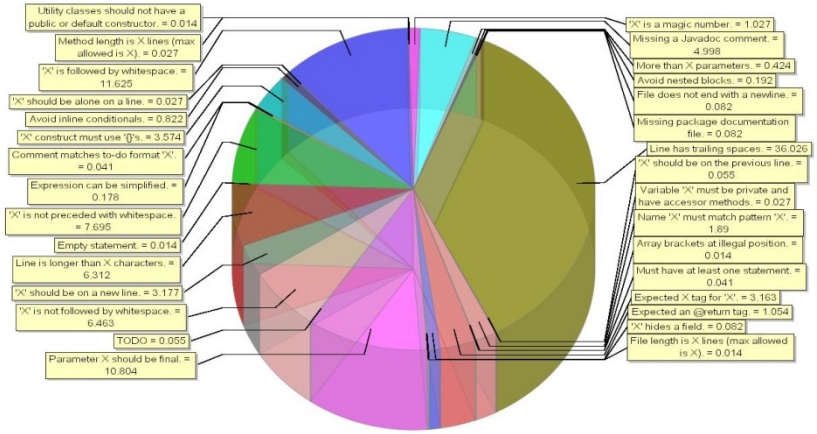
Quelle: ALLBUS 2002

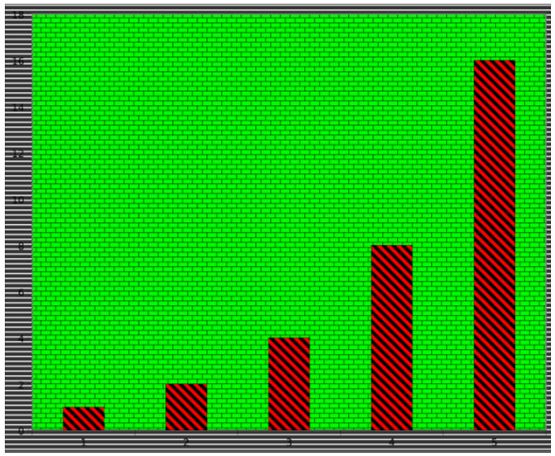
# Häufigkeitsverteilung



# Häufigkeitsverteilung



Negativbeispiele (Chart-Junk):
Bitte kein 3D, mit Designelementen sparsam umgehen, Kreisdiagramme grundsätzlich vermeiden

# Häufigkeitsverteilung

- Um die Verteilung metrisch skalierter Merkmale mit sehr vielen Merkmalsausprägungen zu beschreiben, sind Häufigkeitstabellen oder Balkendiagramme wenig geeignet:

**ALTER: BEFRAGTE<R>**

| | | Häufigkeit | Prozent | Gültige Prozent | Kumulative Prozente |
|---|---|---|---|---|---|
| Gültig | 18 | 32 | ,9 | ,9 | ,9 |
| | 19 | 62 | 1,8 | 1,8 | 2,7 |
| | 20 | 60 | 1,7 | 1,7 | 4,5 |
| | 21 | 42 | 1,2 | 1,2 | 5,7 |
| | 22 | 41 | 1,2 | 1,2 | 6,9 |
| | 23 | 52 | 1,5 | 1,5 | 8,4 |
| | 24 | 45 | 1,3 | 1,3 | 9,7 |
| | 25 | 37 | 1,1 | 1,1 | 10,7 |
| | 26 | 44 | 1,3 | 1,3 | 12,0 |
| | 27 | 43 | 1,2 | 1,2 | 13,2 |

usw.

Problem: zu viele Werte

# Häufigkeitsverteilung

- Lösung: Daten gruppieren (hier werden 10-Jahres-Altersklassen gebildet):

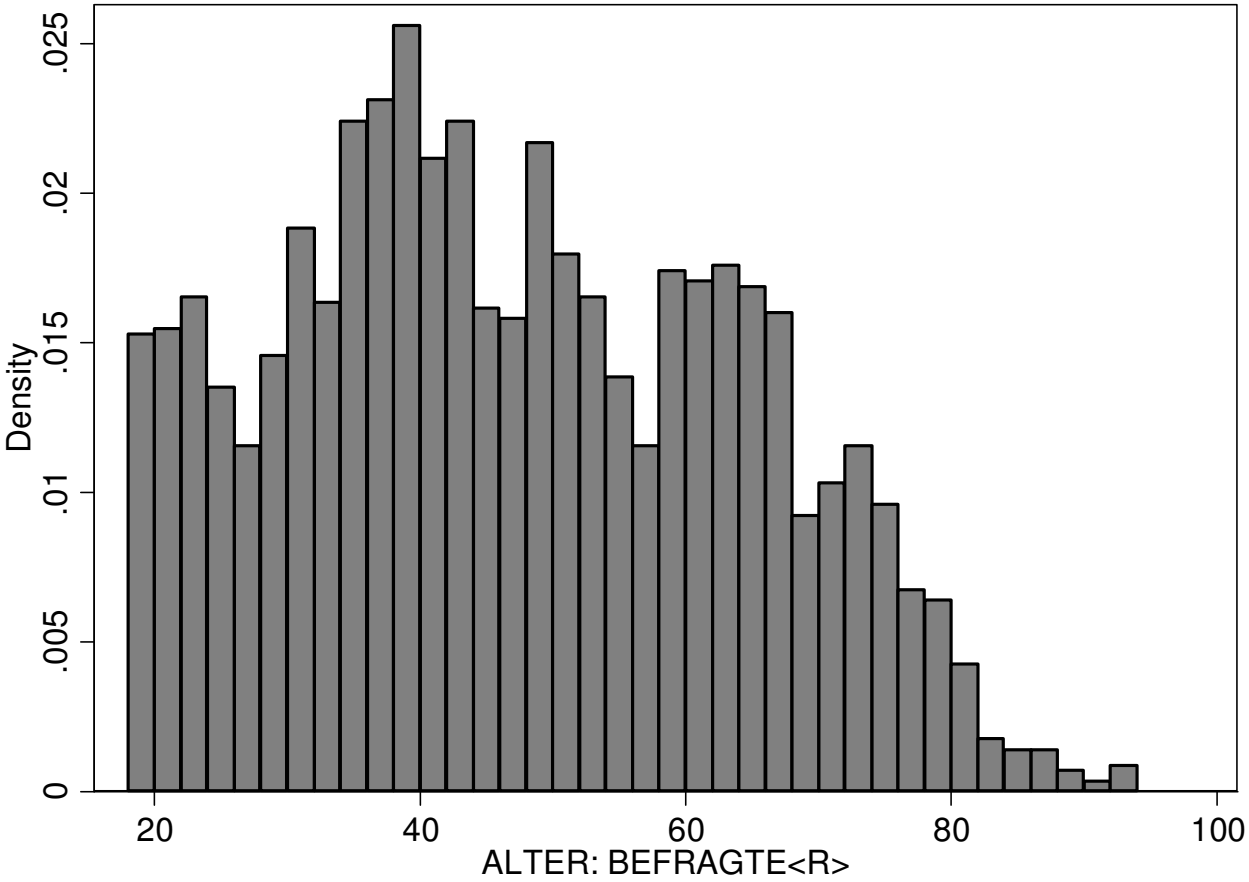
**ALTER: BEFRAGTE<R>, KATEGORISIERT**

| | | Häufigkeit | Prozent | Gültige Prozent | Kumulative Prozente |
|---|---|---|---|---|---|
| Gültig | 18-29 JAHRE | 489 | 17,3 | 17,4 | 17,4 |
| | 30-44 JAHRE | 889 | 31,5 | 31,6 | 49,0 |
| | 45-59 JAHRE | 691 | 24,5 | 24,6 | 73,6 |
| | 60-74 JAHRE | 584 | 20,7 | 20,8 | 94,3 |
| | 75-89 JAHRE | 153 | 5,4 | 5,4 | 99,8 |
| | UEBER 89 JAHRE | 7 | ,2 | ,2 | 100,0 |
| | Gesamtsumme | 2813 | 99,8 | 100,0 | |
| Fehlend | KEINE ANGABE | 7 | ,2 | | |
| Gesamtsumme | | 2820 | 100,0 | | |

# Häufigkeitsverteilung

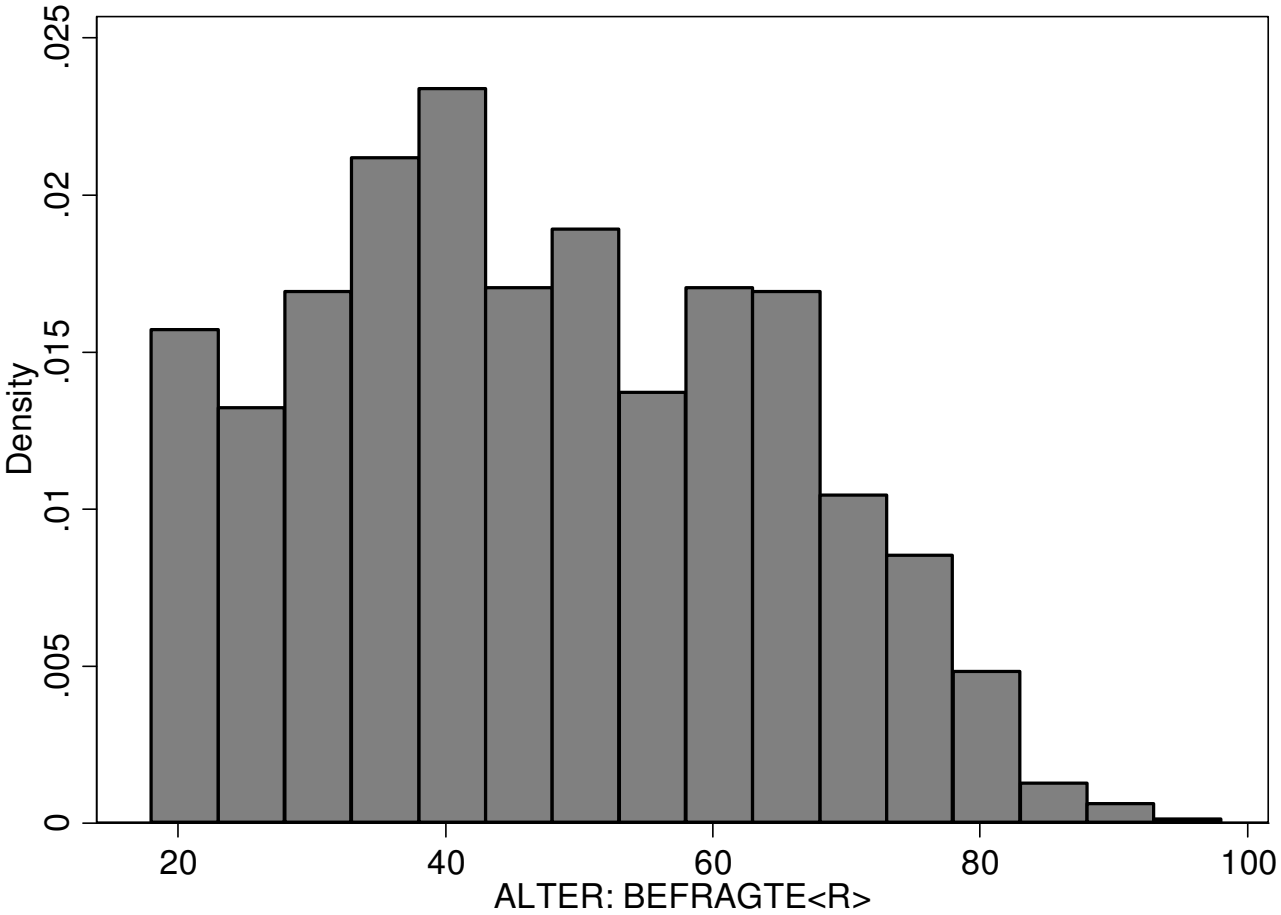
- Eine grafische Lösung ist das **Histogramm** mit folgenden Eigenschaften
  - Histogramm = Form der graphischen Darstellung der Häufigkeitsverteilung eines metrischen Merkmals, dessen Merkmalswerte in Klassen zusammengefasst wurden
  - Klassenhäufigkeiten werden durch die Flächeninhalte aneinander angrenzender Rechtecke repräsentiert
  - Die nächsten Folien zeigen sog. normierte Histogramme. Hier ist die Rechteckhöhe gleich der relativen Häufigkeitsdichte („density“), die als Quotient aus relativer Klassenhäufigkeit und Klassenbreite definiert ist
  - Die Flächeninhalte der aneinander angrenzenden Rechtecke addieren sich bei einem normierten Histogramm stets zu 1

# Häufigkeitsverteilung




Histogramm,
Klassenbreite 2

# Häufigkeitsverteilung




Histogramm,
Klassenbreite 5

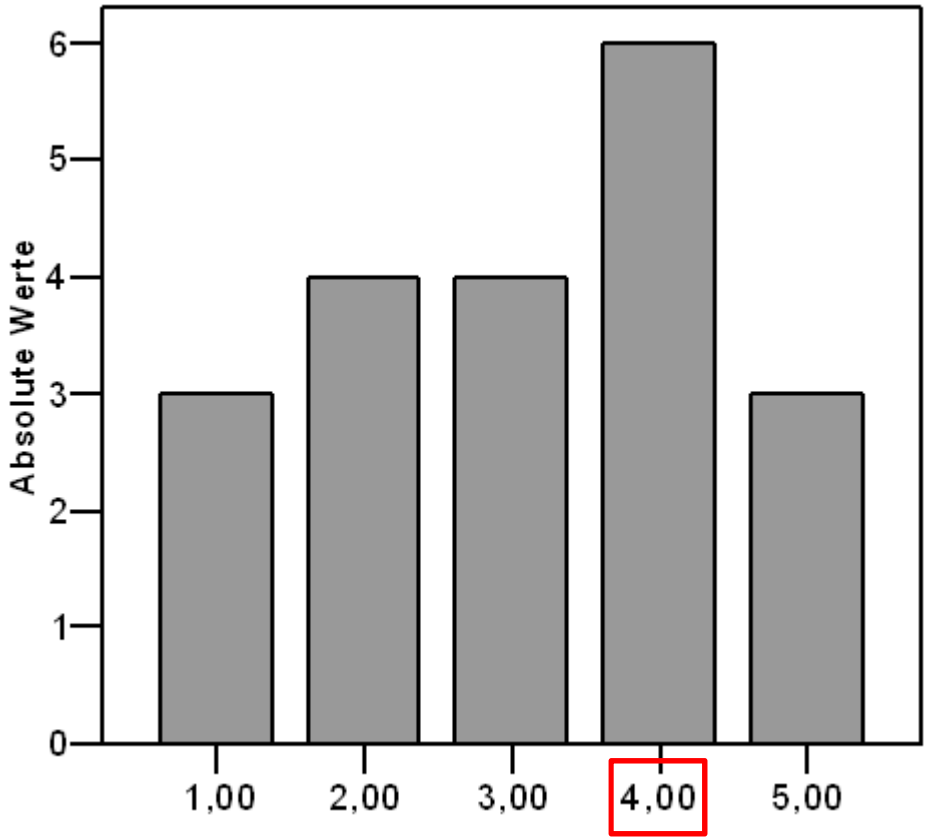
# Zentrale Tendenz

- Insbesondere bei metrischen Merkmalen werden tabellarische Darstellungen schnell unübersichtlich
- Neben einer grafischen Betrachtung (Histogramm, Box-Plot) verwenden wir daher zusätzlich statistische Kennziffern, um zentrale Eigenschaften von Verteilungen zu beschreiben
- Zunächst werden drei Maßzahlen der zentralen Tendenz vorgestellt: Modus, Median und arithmetischer Mittelwert

# Zentrale Tendenz

- Maße der zentralen Tendenz (auch: Lagemaße) werden insofern „repräsentative“ Werte genannt, weil sie den typischen, zentralen oder durchschnittlichen Wert einer Verteilung beschreiben
- Welche Kennziffer jeweils verwendet wird, richtet sich nach dem Messniveau der Daten
- Eine einfache Maßzahl der zentralen Tendenz, die bei jedem Messniveau berechnet werden kann, ist der **Modus** (abgekürzt mit h)
- Er ist definiert als der am häufigsten vorkommende Wert einer Verteilung

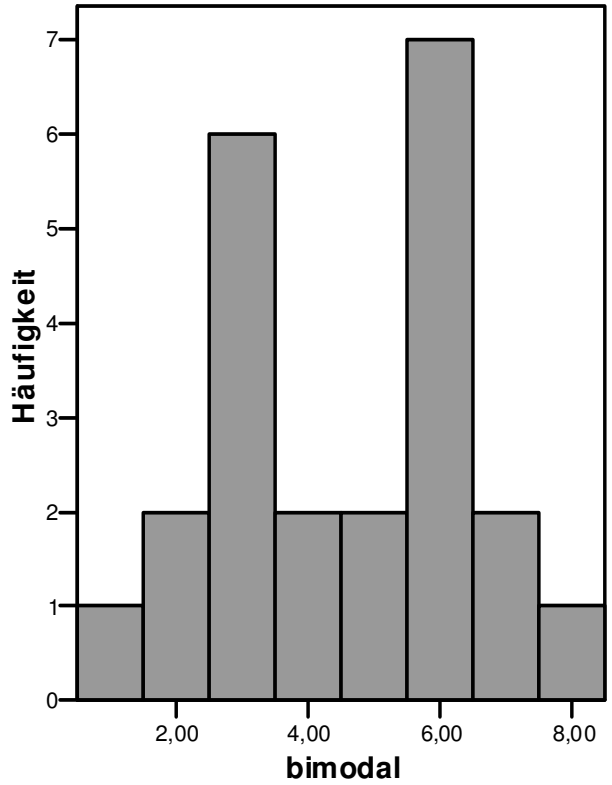
# Zentrale Tendenz




Modus (h) = 4

# Zentrale Tendenz

- Eigenschaften des Modus
  - Ist eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt
  - Kann bereits ab Nominalskalenniveau berechnet werden
  - Ist problematisch bei bi- und multimodalen Verteilungen (siehe Grafik rechts); allgemein bei sehr vielen, ähnlich besetzten Kategorien



# Zentrale Tendenz

- Der **Median** ( $\tilde{x}$ ) kennzeichnet die exakte Mitte einer Verteilung, deren Ausprägungen (aufsteigend) geordnet sind, d.h.:

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{für n ungerade} \\ \frac{1}{2}\left( x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) & \text{für n gerade} \end{cases}$$

# Zentrale Tendenz

- Beispiel 1 (ungerade,  n = 11): 3, 4, 4, 5, 6, **7**, 8, 8, 8, 9, 10
- Hier ist der Median = 7, weil ebenso viele Fälle unterhalb wie oberhalb des sechsten Falles liegen (11 + 1 / 2 = 6, d.h. die sechste Beobachtung ist der Median)
- Beispiel 2 (gerade, n = 10): 3, 4, 4, 5, **6, 7**, 7, 8, 8, 9
- Hier ist der Median der halbierte Wert des 10/2-ten und des (10/2+1)-ten Falls, also = 6 + 7 / 2 = 6,5

# Zentrale Tendenz

- Eigenschaften des Median:
  - Sinnvoll ab dem ordinalen Messniveau
  - Unempfindlich gegenüber Ausreißern
  - Mindestens 50% der Fälle sind kleiner oder gleich dem Median
  - Mindestens 50% der Fälle sind größer oder gleich dem Median

# Zentrale Tendenz

- Die bekannteste Maßzahl der zentralen Tendenz einer Verteilung ist der **arithmetische Mittelwert** ( $\overline{x}$), dessen Berechnung metrische Daten voraussetzt
- Das arithmetische Mittel ist definiert als die Summe der Messwerte, definiert durch ihre Anzahl:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Wenn Werte mehr als einmal vorkommen, kann man sie mit der Häufigkeit ($f_i$) multiplizieren, mit der sie vorkommen:

$$\overline{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_k x_n}{n} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$

# Zentrale Tendenz

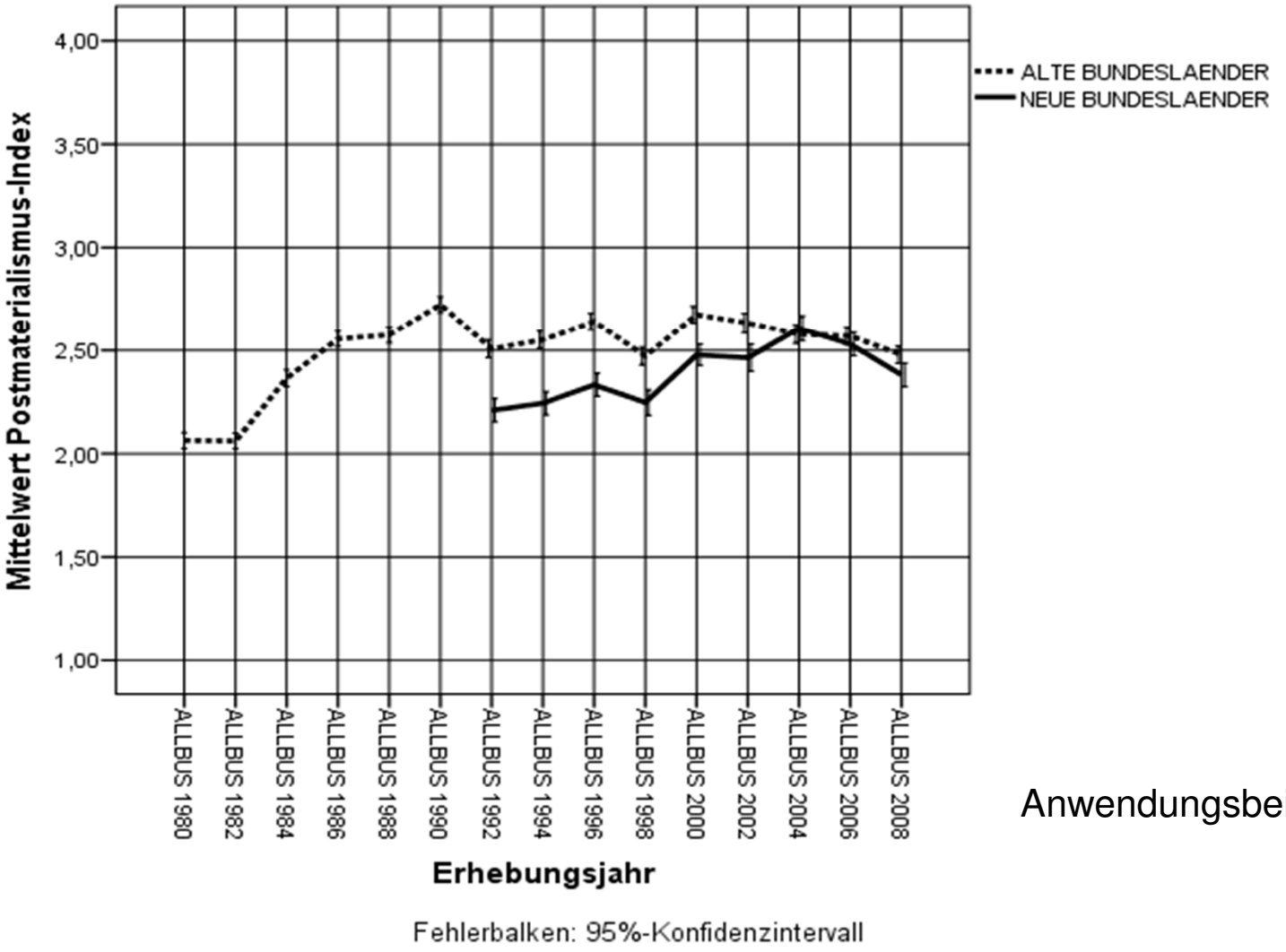
- Als Beispiel sei folgende Verteilung gegeben (n = 8): 6, 7, 7, 8, 8, 8, 9, 9

$$\bar{x} = \frac{1(6) + 2(7) + 3(8) + 2(9)}{8} = \frac{62}{8} = 7{,}75$$

- Eigenschaften des arithmetischen Mittelwertes:
  - Berechnung setzt metrisches Messniveau voraus
  - Empfindlich gegen Ausreißer
  - „Schwerpunkteigenschaft“ (Summe der positiven und negativen Abweichungen vom arithmetischen Mittel = 0, → Kovarianz):

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

# Zentrale Tendenz




Anwendungsbeispiel

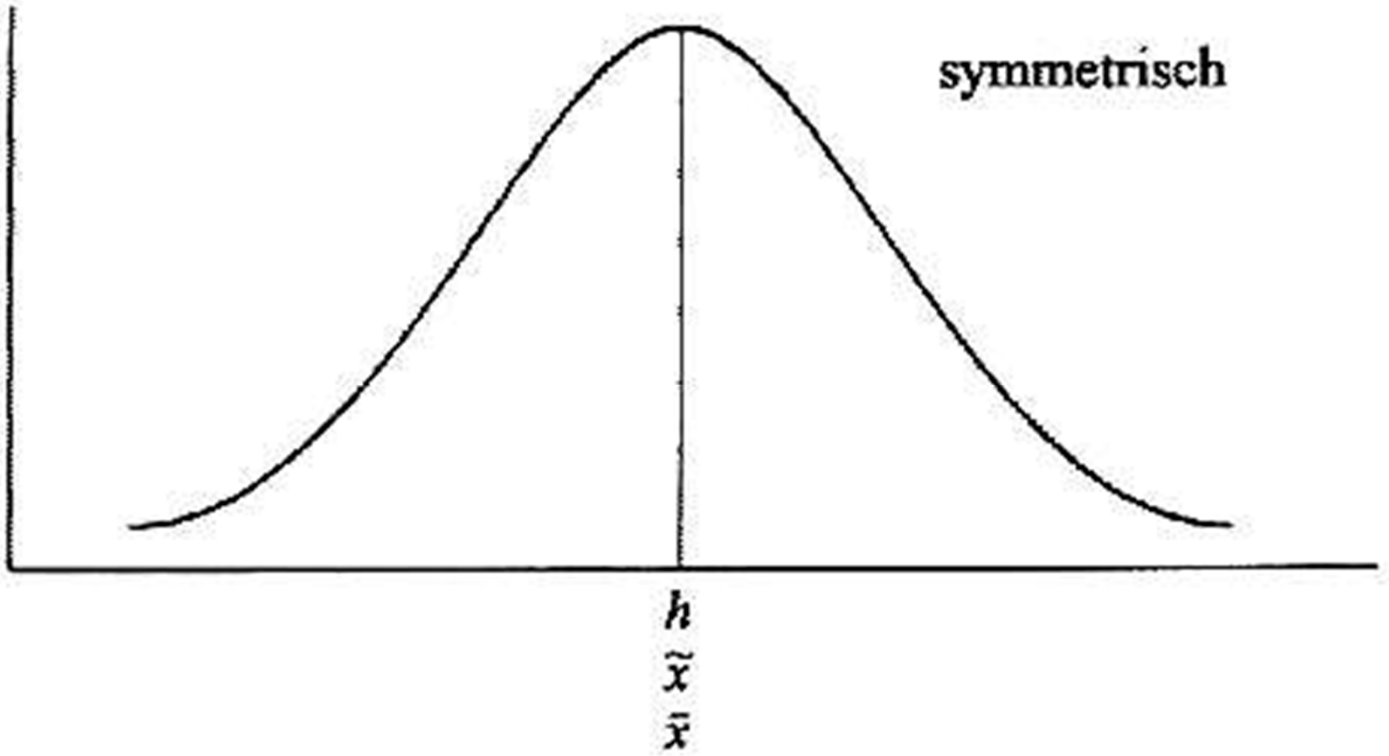
# Zentrale Tendenz

- Keiner der Mittelwerte ist einem anderen in jeder Hinsicht überlegen und sollte universell verwendet werden
- Jeder reflektiert viel mehr einen anderen Aspekt einer Verteilung
- Der Modus gibt Aufschluss über die größte Häufigkeit, der Median über die mittlere Position und das arithmetische Mittel über die Zentralität der Werte
- Die Maße der zentralen Tendenz haben jedoch spezielle Eigenschaften, die mit der Form einer Verteilung zusammenhängen
- Ist eine Verteilung unimodal und exakt symmetrisch, sind Modus, Median und arithmetischer Mittelwert identisch

# Zentrale Tendenz

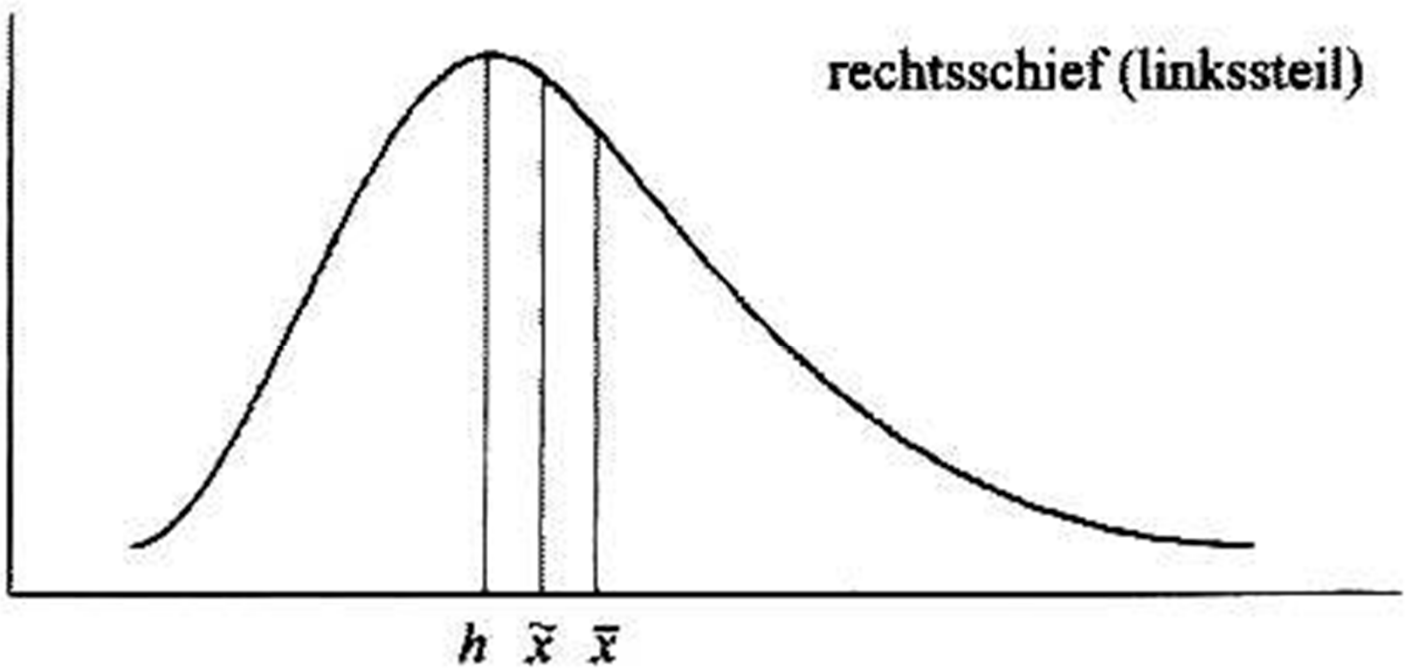
- Bei unimodalen rechtsschiefen (linkssteilen) Verteilungen besteht folgende Beziehung zwischen den Mittelwerten: Modus < Median < Arithmetisches Mittel
- Bei linksschiefen (rechtssteilen) Verteilungen lautet die Beziehung: Arithmetisches Mittel < Median < Modus
- Die wichtigste Schlussfolgerung daraus lautet: Der arithmetische Mittelwert reagiert stärker auf Ausreißer / Extremwerte als Median und Modus!
- Auf den nächsten Folien sind die genannten Eigenschaften der Maße der zentralen Tendenz dargestellt, wobei die vertikale y-Achse für die Häufigkeit und die horizontale x-Achse für eine mindestens ordinal skalierte Variable steht

# Zentrale Tendenz



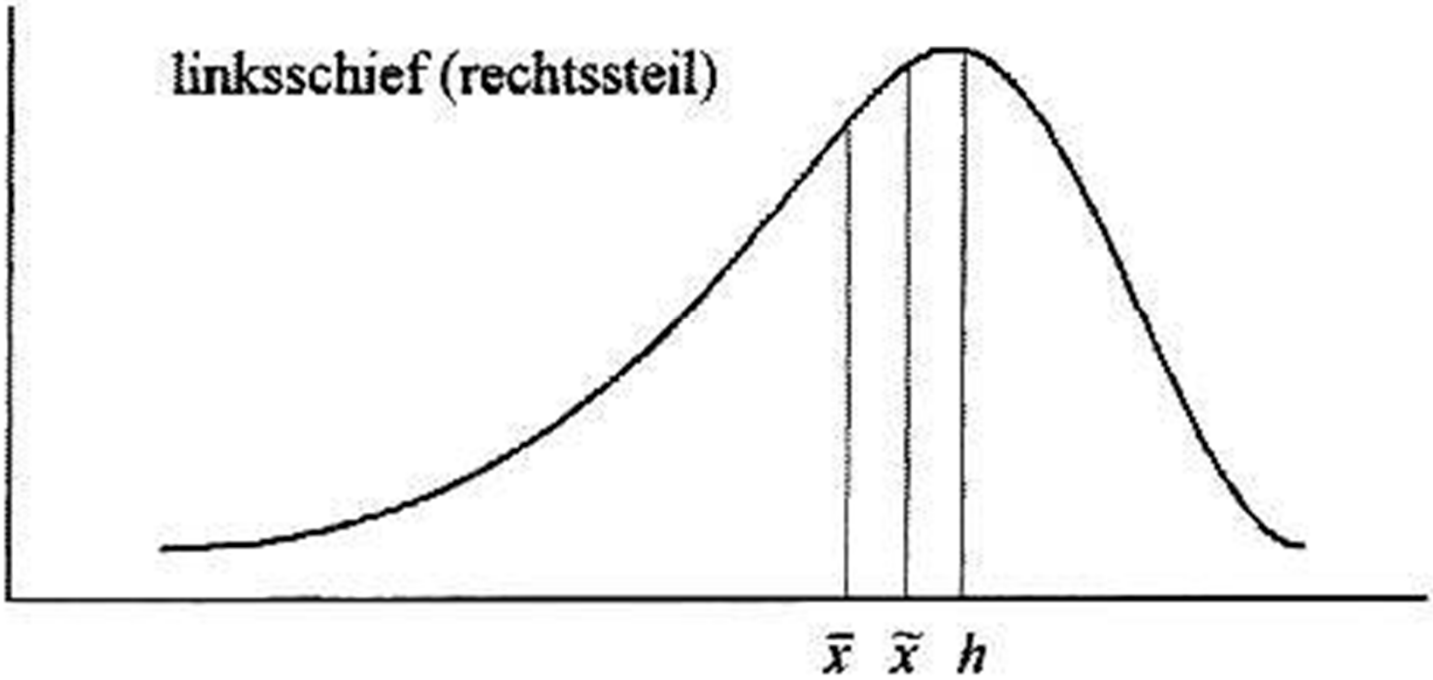

Aus: Benninghaus (1998): Deskriptive Statistik, S. 50

# Zentrale Tendenz




Aus: Benninghaus (1998): Deskriptive Statistik, S. 50

# Zentrale Tendenz



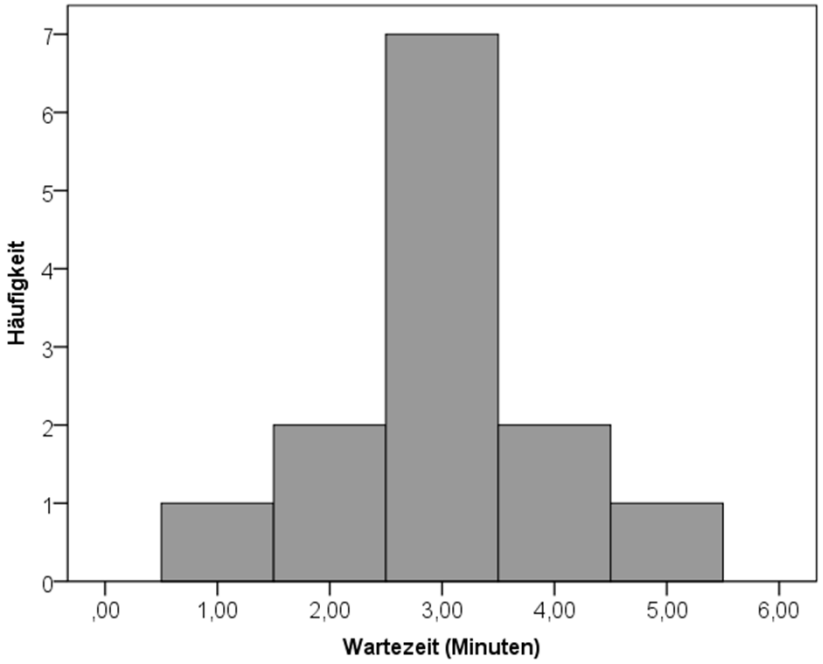

Aus: Benninghaus (1998): Deskriptive Statistik, S. 50

# Streuung

- Die zentrale Tendenz gibt keinen Aufschluss über die Homogenität bzw. Heterogenität der Variablenwerte; hierzu benötigt man Streuungsmaße
- Die nächste Folie zeigt zwei Verteilungen der Wartezeit bis zum Eintreffen eines Taxis nach dem Anruf
- Bei beiden Taxiunternehmen betragen alle Maßzahlen der zentralen Tendenz 3 Minuten – nach Maßgabe dieser Kennziffern gibt es also keinen Unterschied

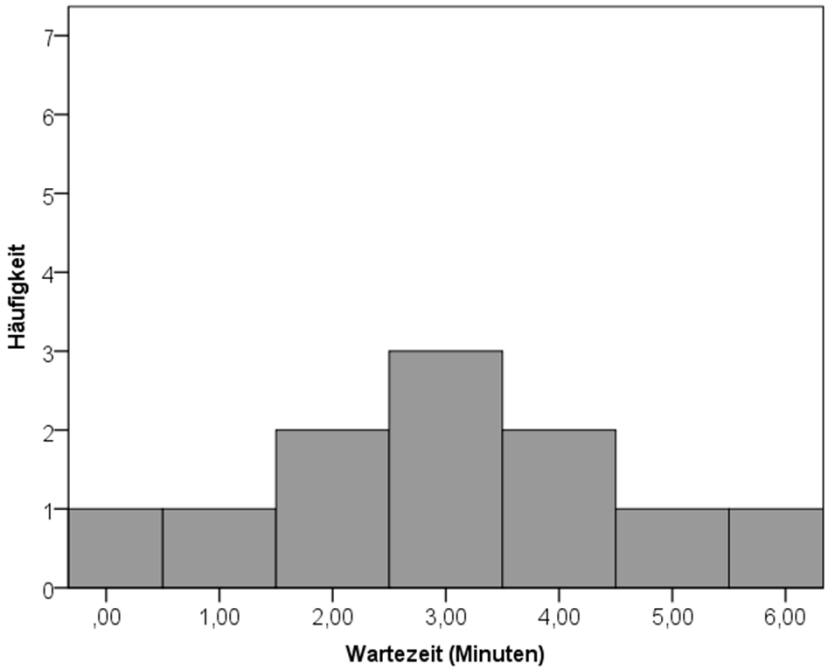
# Streuung

Unternehmen A




Arithmetischer Mittelwert: 3
Median: 3
Modus: 3

Unternehmen B




Arithmetischer Mittelwert: 3
Median: 3
Modus: 3

# Streuung

- Die Streuung der Verteilungen ist aber sehr unterschiedlich
- Die **Spannweite** (auch: Range) der Verteilung, definiert als die Differenz zwischen dem größten und kleinsten Wert, beträgt bei Unternehmen A 4 Minuten (5 - 1 = 4)
- Bei Unternehmen B beträgt sie dagegen 6 Minuten (6 – 0 = 6)
- Zudem kommt der zentrale Wert 3 bei Unternehmen A deutlich häufiger vor als bei Unternehmen B
- Bei Unternehmen A handelt es sich deswegen um eine steilgipflige Verteilung und bei Unternehmen B um eine flachgipfelige (→ Kurtosis)
- Wir suchen nun Koeffizienten, welche die Unterschiede in der Streuung erfassen

# Streuung

- Die gebräuchlichsten Streuungsmaße sind die **Standardabweichung (s)**, definiert als die Quadratwurzel aus der **Varianz ($s^2$)**, die ihrerseits definiert ist als die durch die Anzahl der Messwerte (für Stichprobendaten: n-1) geteilte Summe der quadrierten Abweichungen aller Messwerte von ihrem arithmetischen Mittel:

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

- Varianz und Standardabweichungen setzen ein metrisches Messniveau voraus
- Zur Berechnung der beiden Kennwerte sind die folgenden Arbeitstabellen hilfreich:

# Streuung

Unternehmen A

| $x_i$ | $f_i$ | $(x_i - \overline{x})$ | $(x_i - \overline{x})^2$ | $f_i(x_i - \overline{x})^2$ |
|---|---|---|---|---|
| 1 | 1 | -2 | 4 | 4 |
| 2 | 2 | -1 | 1 | 2 |
| 3 | 7 | 0 | 0 | 0 |
| 4 | 2 | 1 | 1 | 2 |
| 5 | 1 | 2 | 4 | 4 |
| $\sum$ | 13 | | | 12 |

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{12}{12} = 1 \qquad s = \sqrt{1} = 1$$

# Streuung

Unternehmen B

| $x_i$ | $f_i$ | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ | $f_i(x_i - \bar{x})^2$ |
|---|---|---|---|---|
| 0 | 1 | -3 | 9 | 9 |
| 1 | 1 | -2 | 4 | 4 |
| 2 | 2 | -1 | 1 | 2 |
| 3 | 3 | 0 | 0 | 0 |
| 4 | 2 | 1 | 1 | 2 |
| 5 | 1 | 2 | 4 | 4 |
| 6 | 1 | 3 | 9 | 9 |
| $\sum$ | 11 | | | 30 |

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{30}{10} = 3 \qquad s = \sqrt{3} = 1{,}73$$

# Streuung

- Die Varianz der Wartezeiten für Taxiunternehmen A liegt bei einer Quadratminute, wogegen sie für Unternehmen B drei Quadratminuten beträgt
- In der deskriptiven Statistik ist die Standardabweichung der Varianz vorzuziehen, weil sie ein Kennwert in der Messeinheit der zugrunde liegenden Variable ist (Minuten anstatt Minuten$^2$)
- Die Standardabweichung (auch: durchschnittliche Abweichung vom Mittelwert) beträgt für Unternehmen A 1 Minute und für Unternehmen B 1,73 Minuten

# Streuung

Anwendungsbeispiel

Tabelle: Religiöse Selbsteinstufung (1 = gar nicht religiös bis 10 = sehr religiös) in Abhängigkeit von konfessioneller Homogamie der Eltern

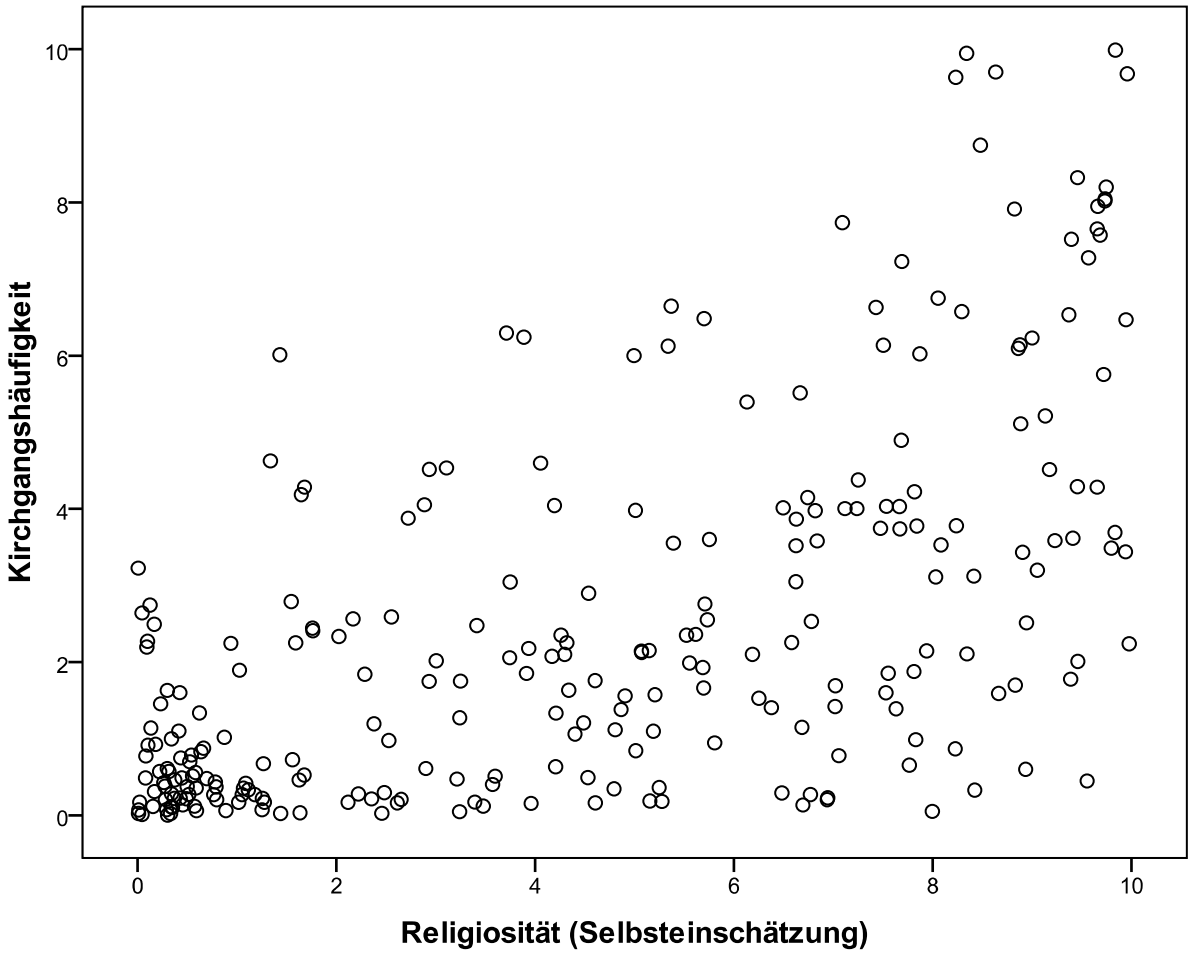
| | N | Mittelwert | s |
|---|---|---|---|
| Konfessionelle Konstellation Eltern | | | |
| Beide konfessionslos | 354 | 2.18 | 2.33 |
| Gleiche Konfession | 1807 | 5.78 | 2.96 |
| Unterschiedliche Konfessionen | 375 | 4.69 | 2.88 |
| Insgesamt | 2536 | 5.12 | 3.12 |

Quelle: ALLBUS 2002 (eigene Berechnungen)

# Streuung

- Wie zwei Variablen gemeinsam (in Abhängigkeit voneinander) streuen, zeigt das folgende **Streudiagramm**
  - Da die Variable Kirchgangshäufigkeit lediglich 6 diskrete Ausprägungen hat und die Variable religiöse Selbsteinstufung 10, wurden beiden Variablen bei der Erstellung des Diagramms kleine Zufalls-Fehler (Jitter) zugespielt
  - Die Anschaulichkeit und Aussagekraft von Streudiagrammen wird häufig durch die Einzeichnung von linearen oder nicht-linearen Anpassungslinien erhöht (→ Korrelation, Regression)

# Streuung


Kirchgangshäufigkeit
0
2
4
6
8
10
Religiosität (Selbsteinschätzung)
0
2
4
6
8
10

# Streuung

- Ein Heterogenitätsmaß für nominale oder ordinale Variablen ist der **Index qualitativer Variation**
- Grundidee: Der Index nimmt den Wert 0 an, wenn alle Fälle die gleiche Ausprägung aufweisen und den Wert 1, wenn die einzelnen Gruppen gleich stark besetzt sind
- Der Index kann nach folgender Formel berechnet werden ($p^2_k$ ist die relative Häufigkeit der k-ten von m möglichen Klassen):

$$IQV = \frac{1 - \sum_{k=1}^{m} p_k^2}{\frac{1}{m} * (m-1)}$$

# Streuung

**Konfession**

| ERHEBUNGSGEBIET: WEST - OST | | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|---|
| ALTE BUNDESLAENDER | Gültig | katholisch | 759 | 39,2 | 39,5 | 39,5 |
| | | evangelisch | 772 | 39,9 | 40,2 | 79,7 |
| | | andere Religionsgemeinschaft | 100 | 5,2 | 5,2 | 84,9 |
| | | keiner Konfession zugehrig | 291 | 15,0 | 15,1 | 100,0 |
| | | Gesamt | 1922 | 99,4 | 100,0 | |
| | Fehlend | fehlender Wert | 12 | ,6 | | |
| | Gesamt | | 1934 | 100,0 | | |
| NEUE BUNDESLAENDER | Gültig | katholisch | 58 | 6,5 | 6,6 | 6,6 |
| | | evangelisch | 258 | 29,1 | 29,2 | 35,8 |
| | | andere Religionsgemeinschaft | 12 | 1,4 | 1,4 | 37,1 |
| | | keiner Konfession zugehrig | 555 | 62,6 | 62,9 | 100,0 |
| | | Gesamt | 883 | 99,7 | 100,0 | |
| | Fehlend | fehlender Wert | 3 | ,3 | | |
| | Gesamt | | 886 | 100,0 | | |

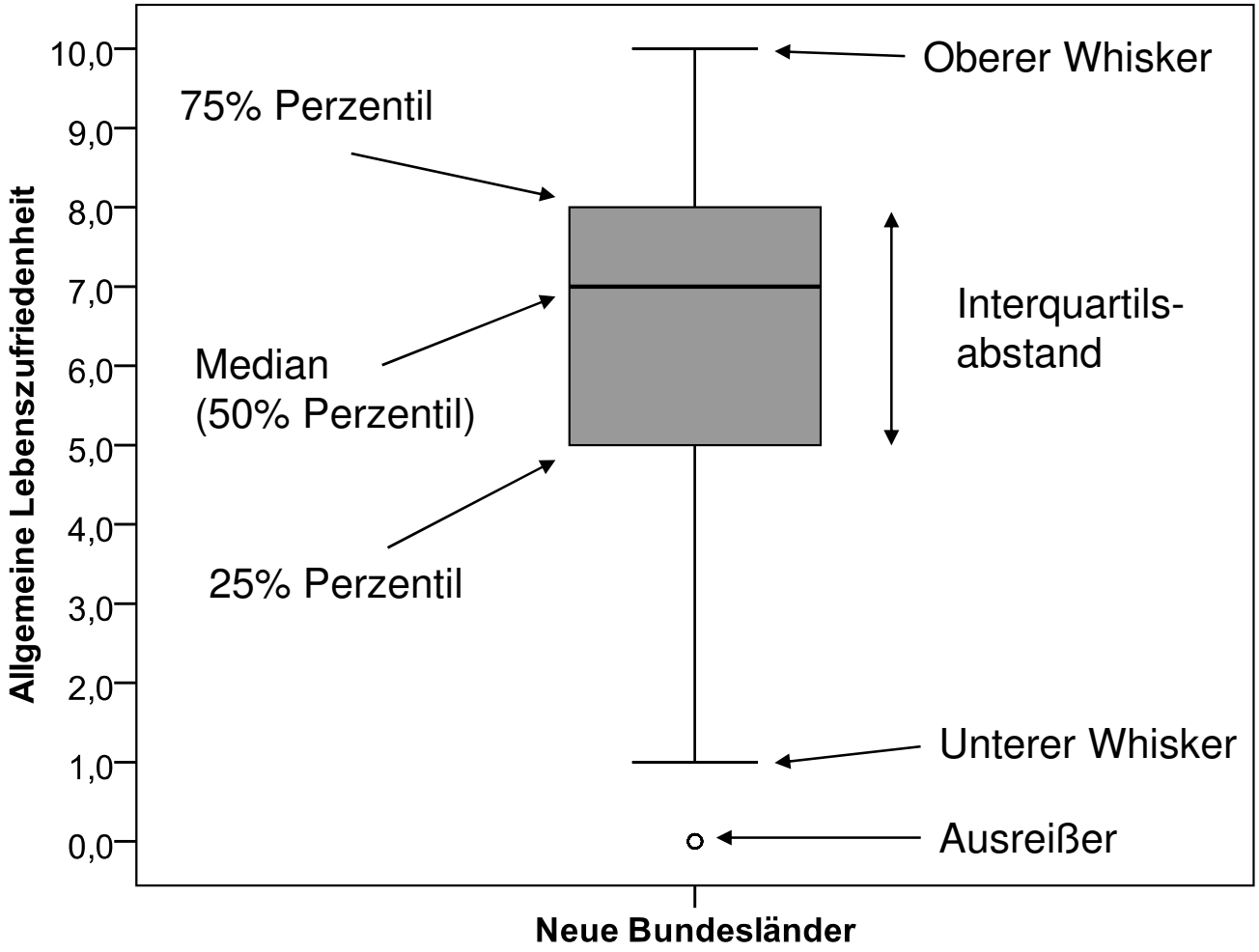
IQV-West: $1-(0{,}395^2+0{,}402^2+0{,}052^2+0{,}151^2) / (1/4)*(4-1) = 0{,}656 / 0{,}75 = 0{,}87$

IQV-Ost: $1-(0{,}066^2+0{,}292^2+0{,}014^2+0{,}629^2) / (1/4)*(4-1) = 0{,}51 / 0{,}75 = 0{,}68$

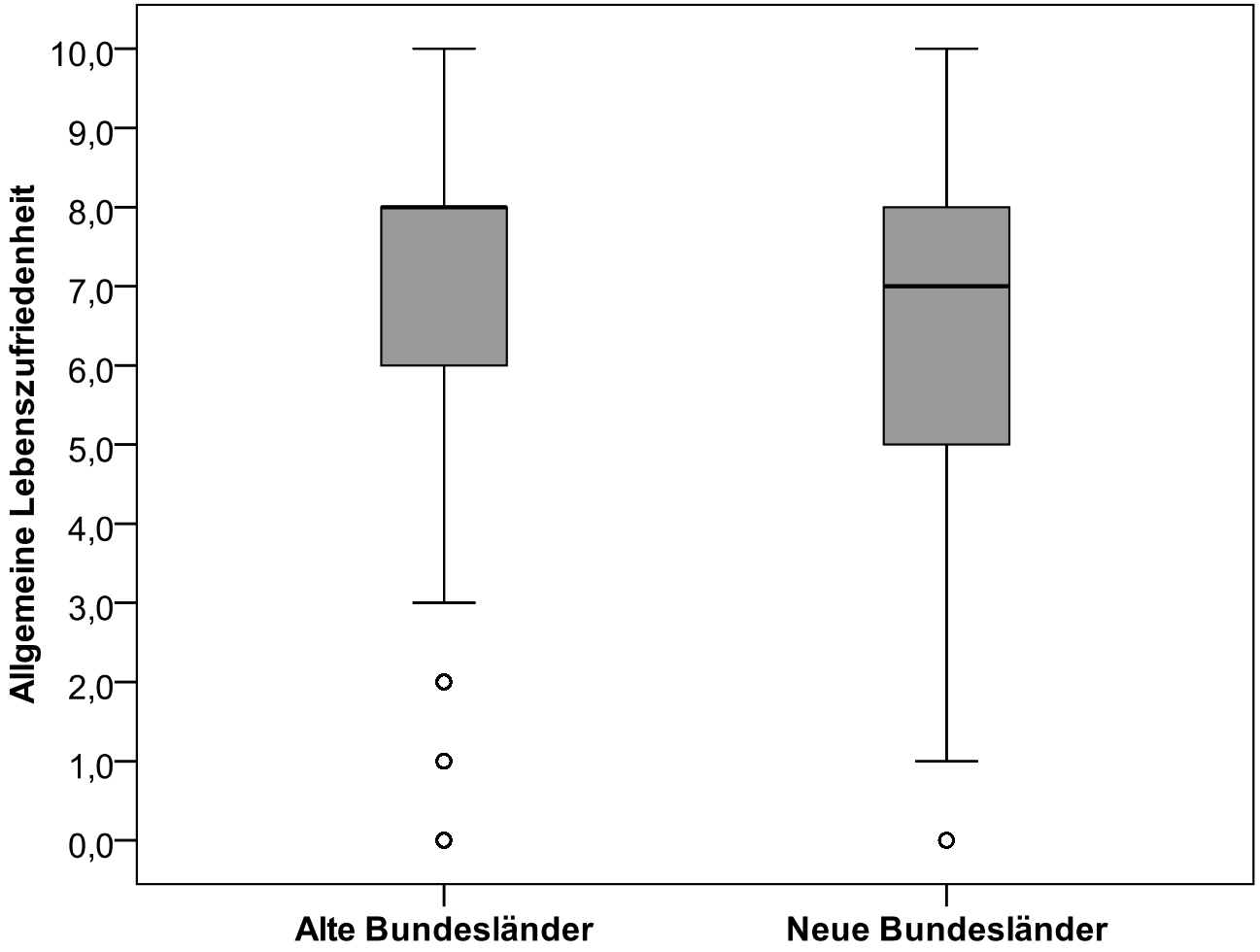
# Streuung

- Ein gebräuchliches Diagramm, das sowohl über die zentrale Tendenz als auch die Streuung einer Verteilung informiert, ist der **Box-Plot**
  - Die „Whisker“ (abgeleitet von Katzenschnurbart) kennzeichnen den größten bzw. kleinsten Wert, der noch keinen Ausreißer darstellt
  - Als Ausreißer gekennzeichnet sind Werte, die um mehr als das 1,5-fache der Boxhöhe (also um mehr als 1,5 Interquartilsabstände) über oder unterhalb der Box liegen
  - Werte, die um mehr als das 3-fache der Boxhöhe über oder unterhalb der Box liegen, gelten als Extremwerte und werden mit Sternen gekennzeichnet
  - Der Median muss nicht in der Mitte der Box liegen, seine Lage hängt von der Verteilung ab
  - Box-Plots werden meist eingesetzt, um die Verteilung einer mindestens ordinalen Variablen zwischen zwei oder mehr Gruppen zu vergleichen (siehe übernächste Folie)

# Streuung


10,0
9,0
8,0
7,0
6,0
5,0
4,0
3,0
2,0
1,0
0,0
Allgemeine Lebenszufriedenheit
Neue Bundesländer
Oberer Whisker
75% Perzentil
Interquartils-
abstand
Median
(50% Perzentil)
25% Perzentil
Unterer Whisker
Ausreißer

# Streuung


Allgemeine Lebenszufriedenheit
10,0
9,0
8,0
7,0
6,0
5,0
4,0
3,0
2,0
1,0
0,0
Alte Bundesländer
Neue Bundesländer

# Höhere Momente

- Neben zentraler Tendenz und Streuung werden nun Koeffizienten für die sog. höheren Momente der Verteilung betrachtet: **Schiefe** und **Kurtosis**
- Der Schiefe-Koeffizient ist 0 bei symmetrischer, negativ bei rechtssteiler und positiv bei linkssteiler Verteilung
- Der vierte Moment einer Verteilung ist die sog. Kurtosis oder Wölbung; hohe Werte des Kurtosis-Koeffizienten indizieren steilgipflige Verteilungen mit hoher Wölbung
- Die nächsten Folien zeigen die Berechnungsformeln
  - Achtung: Statistikprogramme wie Excel oder SPSS rechnen unter Umständen mit anderen Formeln, die für den Stichprobenumfang adjustiert sind
  - Gelegentlich wird die Kurtosis einer Verteilung als Differenz zur Kurtosis der Standardnormalverteilung angegeben, die 3 beträgt

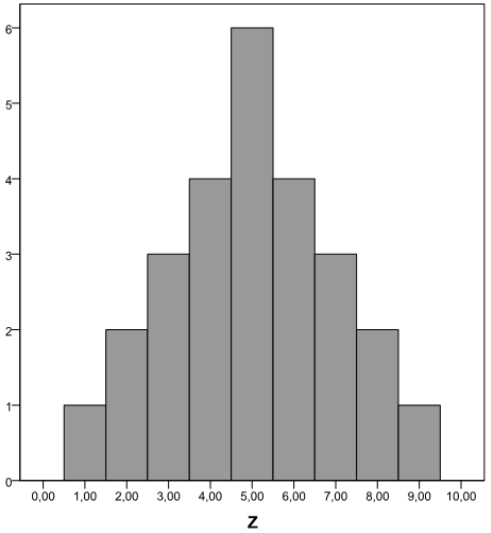
# Höhere Momente

- Die nachfolgenden Formeln zu Schiefe und Kurtosis basieren auf **z-standardisierten** Messwerten (→ Inferenzstatistik)
- Man erzeugt z-transformierte Werte $z_i$, indem man von jedem Messwert ($x_i$) das arithmetische Mittel subtrahiert und diese Differenz durch die Standardabweichung (s) dividiert:

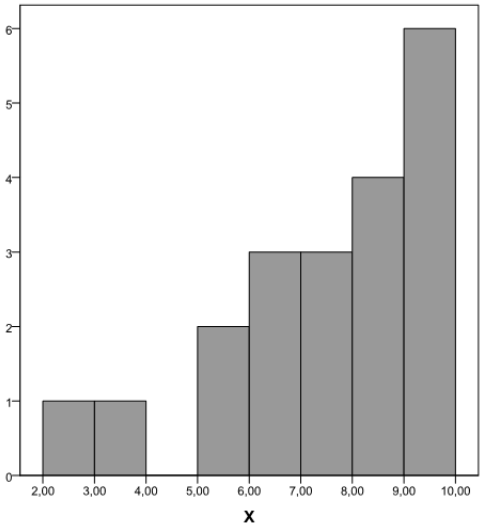
$$z_i = \frac{x_i - \bar{x}}{s}$$

- z-transformierte Variablen haben einen Mittelwert von 0 und eine Standardabweichung von 1; dadurch werden die Schiefe- und Kurtosis-Koeffizienten unabhängig von der Messeinheit der jeweiligen Variablen

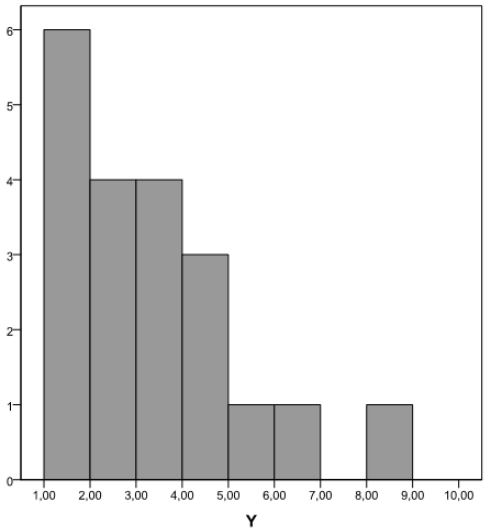
# Höhere Momente




Schiefe = 0
(symmetrisch)



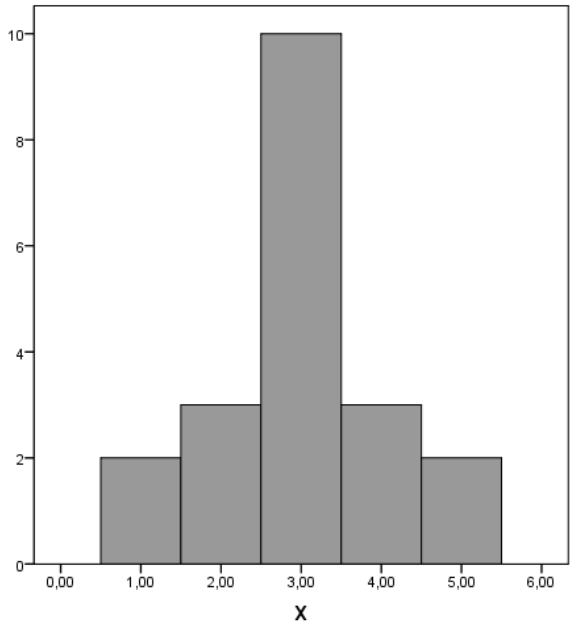

Schiefe = -0,75
(rechtssteil)

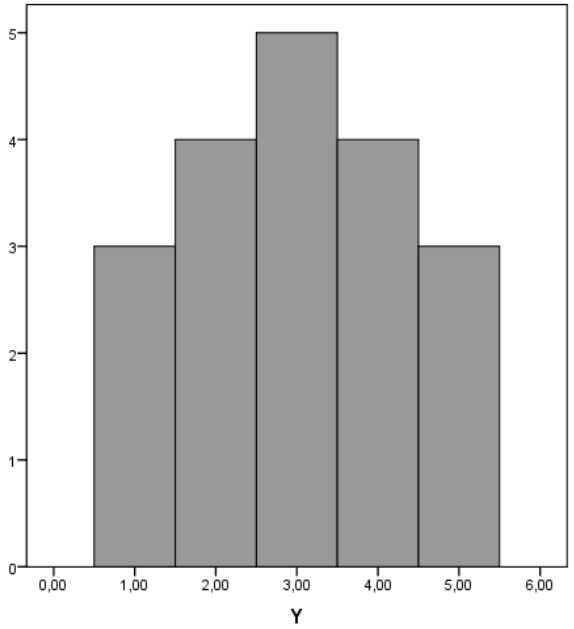



Schiefe = 1,22
(linkssteil)

$$Schiefe = \frac{\sum_{i=1}^{n} z_i^{\,3}}{n}$$

# Höhere Momente




Kurtosis = 2,61

Kurtosis = 1,73

$$Kurtosis = \frac{\sum_{i=1}^{n} z_i^4}{n}$$

# Übersicht: Zentrale Tendenz & Streuung

| | **Maße der zentralen Tendenz** | **Streuungsmaße** |
|---|---|---|
| **Nominal** | Modus | Index qualitativer Variation |
| **Ordinal** | Modus, Median | Index qualitativer Variation |
| **Metrisch** | Modus, Median, arithmetischer Mittelwert | Spannweite, Interquartilsabstand, Standardabweichung, Varianz |

Ohne Anspruch auf Vollständigkeit

# Übersicht: Einige Grafik-Typen

Betrachtung einer Variable (univariate Analyse):

| **Nominal** | **Ordinal** | **Metrisch** |
|---|---|---|
| Balkendiagramm | Balkendiagramm, Box-Plot | Histogramm, Box-Plot |

Betrachtung des Zusammenhangs zwischen zwei Variablen (bivariat):

| | **Nominal** | **Metrisch (ordinal)** |
|---|---|---|
| **Nominal** | Gruppiertes Balkendiagramm | Box-Plot |
| **Metrisch (ordinal)** | | Streudiagramm, Liniendiagramm |

# Ausgewählte Literatur


- Benninghaus, H. (2007): Deskriptive Statistik: Eine Einführung für Sozialwissenschaftler. 11. Auflage. Wiesbaden: Springer VS.
- Kopp, J. & Lois, D. (2014): Sozialwissenschaftliche Datenanalyse. Eine Einführung. 2. Auflage. Wiesbaden: Springer VS (Kapitel 3).
- Müller-Benedict, V. (2006): Grundkurs Statistik in den Sozialwissenschaften. Wiesbaden: VS.