Statistik I - Sitzung 4

Bernd Schlipphak

Institut für Politikwissenschaft

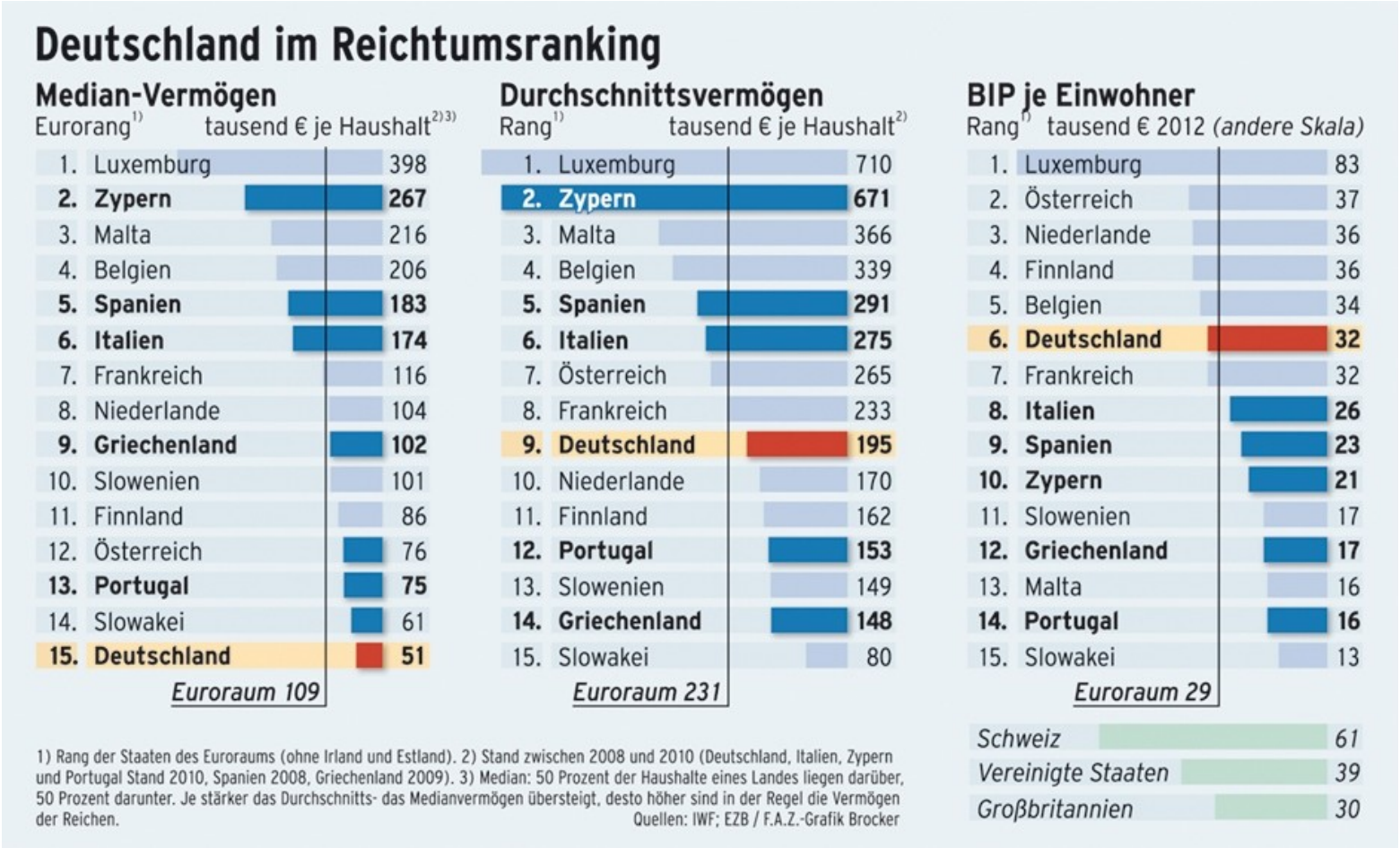
Woche 4

Statistik I - Sitzung 4

1. Maße zentraler Tendenz
   - Der Modus
   - Der Median
   - Der Arithmetische Mittelwert
   - Die Quantile

2. Streuungsmaße
   - Spannweite und Interquartilsabstand
   - Boxplot
   - Varianz und Standardabweichung
   - Schiefe

Was ist das Problem?



Deutschland im Reichtumsranking

Median-Vermögen

Eurorang[1]		tausend € je Haushalt[2)3)]
1.	Luxemburg	398
2.	**Zypern**	**267**
3.	Malta	216
4.	Belgien	206
5.	**Spanien**	**183**
6.	**Italien**	**174**
7.	Frankreich	116
8.	Niederlande	104
9.	**Griechenland**	**102**
10.	Slowenien	101
11.	Finnland	86
12.	Österreich	76
13.	**Portugal**	**75**
14.	Slowakei	61
15.	**Deutschland**	**51**

Euroraum 109

Durchschnittsvermögen

Rang[1]		tausend € je Haushalt[2)]
1.	Luxemburg	710
2.	**Zypern**	**671**
3.	Malta	366
4.	Belgien	339
5.	**Spanien**	**291**
6.	**Italien**	**275**
7.	Österreich	265
8.	Frankreich	233
9.	**Deutschland**	**195**
10.	Niederlande	170
11.	Finnland	162
12.	**Portugal**	**153**
13.	Slowenien	149
14.	**Griechenland**	**148**
15.	Slowakei	80

Euroraum 231

BIP je Einwohner

Rang[1]		tausend € 2012 *(andere Skala)*
1.	Luxemburg	83
2.	Österreich	37
3.	Niederlande	36
4.	Finnland	36
5.	Belgien	34
6.	**Deutschland**	**32**
7.	Frankreich	32
8.	**Italien**	**26**
9.	**Spanien**	**23**
10.	**Zypern**	**21**
11.	Slowenien	17
12.	**Griechenland**	**17**
13.	Malta	16
14.	**Portugal**	**16**
15.	Slowakei	13

Euroraum 29

Schweiz	*61*
Vereinigte Staaten	*39*
Großbritannien	*30*

1) Rang der Staaten des Euroraums (ohne Irland und Estland). 2) Stand zwischen 2008 und 2010 (Deutschland, Italien, Zypern und Portugal Stand 2010, Spanien 2008, Griechenland 2009). 3) Median: 50 Prozent der Haushalte eines Landes liegen darüber, 50 Prozent darunter. Je stärker das Durchschnitts- das Medianvermögen übersteigt, desto höher sind in der Regel die Vermögen der Reichen.

Quellen: IWF; EZB / F.A.Z.-Grafik Brocker

Abbildung: Aus der FAZ (http://www.faz.net/-gqe-78jxy)

Maße zentraler Tendenz

- Grundlegend bestehen die Maße zentraler Tendenz aus den **Mittelwerten** und den sogenannten **Lagemaßen** (oder **Quantilen**)
- Die Mittelwerte lassen sich in drei Maße aufteilen
  - **Modus** (Mode, Modalwert)
  - **Median**
  - **Arithmetisches Mittel** (Durchschnitt, Mean)

Der Modus

- Der Modus ist der am häufigsten vorkommende Wert in einer Datenmenge
- Oder: 'Der Modus h ist die Ausprägung einer diskreten Variablen, die die größte Häufigkeit hat.' (Diaz-Bone 2006: 45)
- Der Modus kann für alle Skalenniveaus berechnet / bestimmt werden

Hypothetisches Beispiel für den Modus

- Haarfarbe (nominal skaliert): Welcher Wert (welche Haarfarbe) kommt am häufigsten in der Datenmenge vor?
- Vorkommen der Ausprägungen: Blond = 23 Fälle, Schwarz = 21 Fälle, Rot = 3 Fälle, Grün = 1 Fall
- Blond kommt am häufigsten vor $\Rightarrow$ **Modus = Blond**

Der Median $\tilde{x}$

- Der Median ist der Wert in der Mitte einer geordneten Datenmenge
- ODER: 'Der Wert einer geordneten Datenmenge, der diese so unterteilt, dass sich links und rechts von diesem Wert jeweils höchstens 50 Prozent der Datenwerte befinden.' (Behnke/Behnke 2006: 124)
- Der Median kann für Variablen, die mindestens Ordinalskalenniveau besitzen, berechnet werden

Hypothetisches Beispiel für den Median I

- Ein Kurs am IfPol mit 9 Studierenden (A - I) diskutiert die potentiellen Wähleranteile der AfD in Münster

Ungeordnete Verteilung	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Erwarteter Anteil an AfD-WählerInnen in %	5	0	5	4	3	1	2	6	30

Hypothetisches Beispiel für den Median I

- Ein Kurs am IfPol mit 9 Studierenden (A - I) diskutiert die potentiellen Wähleranteile der AfD in Münster

Ungeordnete Verteilung	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Erwarteter Anteil an AfD-WählerInnen in %	5	0	5	4	3	1	2	6	30

Geordnete Verteilung	**B**	**F**	**G**	**E**	**D**	**A**	**C**	**H**	**I**
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- **Median = 4 (D)**. Dieser Wert liegt genau in der Mitte der Verteilung (5. Wert von 9 Werten, rechts und links neben ihm genau 4 Werte)

Hypothetisches Beispiel für den Median - II

- Ein Kurs am IfPol mit 10 Studierenden (A - J) diskutiert die potentiellen Wähleranteile der AfD in Münster

Geordnete Verteilung	**B**	**F**	**G**	**E**	**D**	**A**	**C**	**H**	**I**	**J**
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30	99

- **Median = 4.5**. Median bei geraden Wertevorkommen (= geraden Fallzahlen) jener Wert, der zwischen den beiden mittleren Werten (d.h., (D + A) / 2) liegt

Der Median

Merksatz Median

Für eine solche Auszählung des Medians müssen die Ausprägungen oder Datenwerte der Größe der Werte nach geordnet sein! Daher ist der Median nur ab Ordinalniveau einsetzbar!

Der Arithmetische Mittelwert $\bar{x}$

- Den Arithmetischen Mittelwert stellt man immer durch den folgenden Ausdruck dar: $\bar{x}$
- Formal: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Die Berechnung des arithmetischen Mittelwerts ist nur für Variablen mit metrischem Skalenniveau möglich und sinnvoll!

Hypothetisches Beispiel für den Mittelwert

- Ein Kurs am IfPol mit 9 Studierenden (A - I) diskutiert die potentiellen Wähleranteile der AfD in Münster

Geordnete Verteilung	**B**	**F**	**G**	**E**	**D**	**A**	**C**	**H**	**I**
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- Mittelwert: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i$
- **Mittelwert =** 1/9 * (0+1+2+3+4+5+5+6+30) = 1/9*(56) = **6.2**

Besonderheit / Problem des Mittelwerts

- Der arithmetische Mittelwert ist **ausreißersensitiv**, d.h. er reagiert auf Ausreißer sehr stark und verzerrend!
- Dies steht im Gegensatz zu Modus und Median, die **ausreißerresistent** sind

Besonderheit / Problem des Mittelwerts

- Modus für unser Beispiel = 5 / Median für unser Beispiel = 4
- Arithmetischer Mittelwert = 6.2 weicht stark von anderen Werten ab!
- Schließt man den Ausreißerwert in der Verteilung (I = 30) aus, so erhält man einen Mittelwert von 3.3

Besonderheit / Problem des Mittelwerts

- Für unser Beispiel mit 10 Studierenden: Modus = 5 / Median = 4.5
- Arithmetischer Mittelwert = 15.5 weicht dann noch stärker ab!
- Ausreißerwert von J = 99 weist auf extremen Ausreißer oder auf fehlenden Wert hin

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30	99

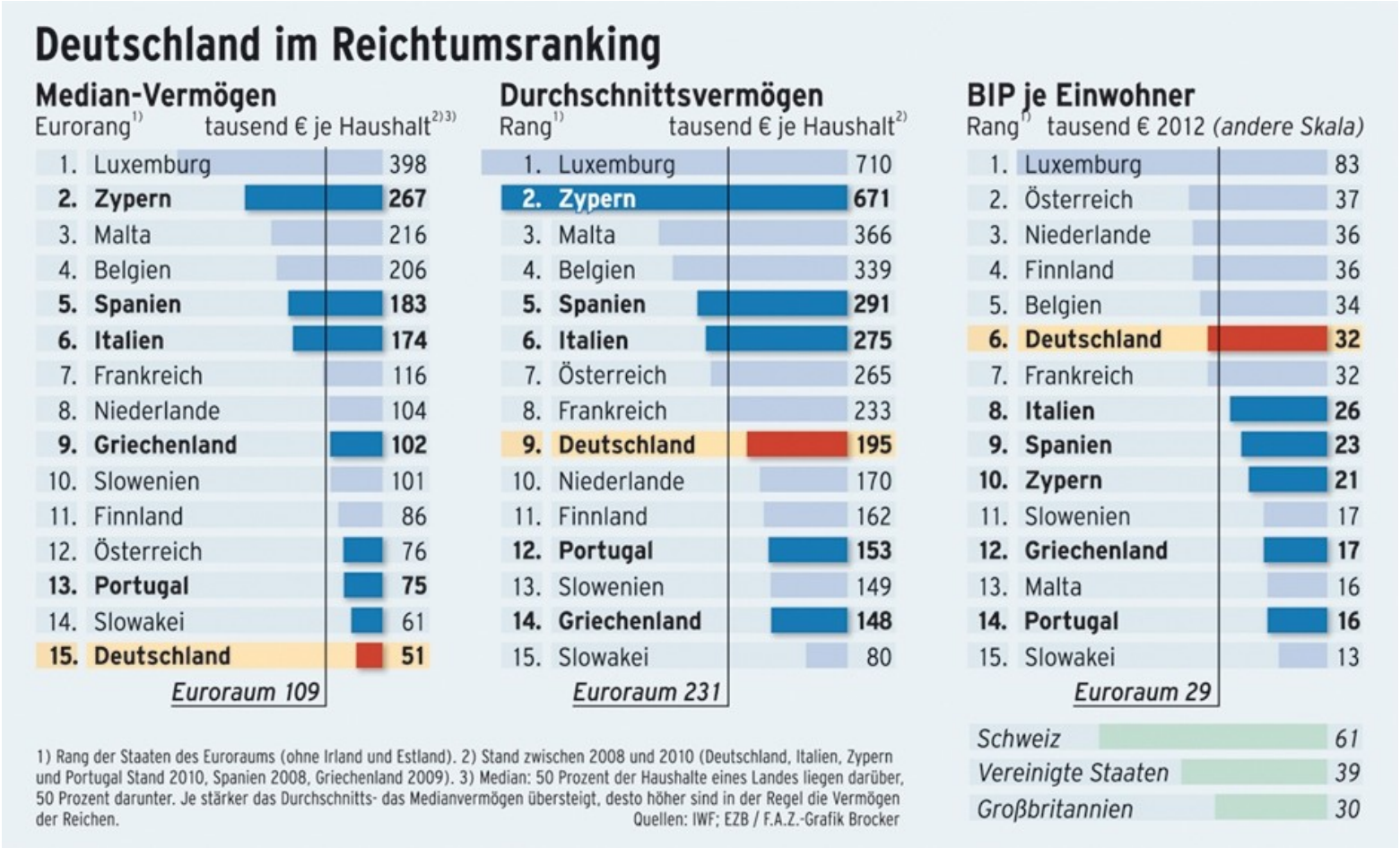
Besonderheit / Problem des Mittelwerts

- Fälle mit fehlenden Werten müssen für die Berechnung des Mittelwerts immer ausgeschlossen werden, da die Werte keine substantiell interpretierbare Größe darstellen und daher den Mittelwert unsinnig verzerren!
- Fälle mit extremen Werten müssen näher angeschaut werden – liegt eventuell ein Codier-Fehler oder eine bewusst unsinnige Antwort vor? Lässt sich der Wert noch sinnvoll erklären?

Merksatz Arithmetischer Mittelwert

Der arithmetische Mittelwert sollte immer zusammen mit dem Median dargestellt werden – so lassen sich Verzerrungen erkennen und auf extreme/fehlende Werte überprüfen!

Das ist das Problem!



Deutschland im Reichtumsranking

Median-Vermögen

Eurorang[1]		tausend € je Haushalt[2)3)]
1.	Luxemburg	398
2.	**Zypern**	**267**
3.	Malta	216
4.	Belgien	206
5.	**Spanien**	**183**
6.	**Italien**	**174**
7.	Frankreich	116
8.	Niederlande	104
9.	**Griechenland**	**102**
10.	Slowenien	101
11.	Finnland	86
12.	Österreich	76
13.	**Portugal**	**75**
14.	Slowakei	61
15.	**Deutschland**	**51**

Euroraum 109

Durchschnittsvermögen

Rang[1]		tausend € je Haushalt[2)]
1.	Luxemburg	710
2.	**Zypern**	**671**
3.	Malta	366
4.	Belgien	339
5.	**Spanien**	**291**
6.	**Italien**	**275**
7.	Österreich	265
8.	Frankreich	233
9.	**Deutschland**	**195**
10.	Niederlande	170
11.	Finnland	162
12.	**Portugal**	**153**
13.	Slowenien	149
14.	**Griechenland**	**148**
15.	Slowakei	80

Euroraum 231

BIP je Einwohner

Rang[1]		tausend € 2012 *(andere Skala)*
1.	Luxemburg	83
2.	Österreich	37
3.	Niederlande	36
4.	Finnland	36
5.	Belgien	34
6.	**Deutschland**	**32**
7.	Frankreich	32
8.	**Italien**	**26**
9.	**Spanien**	**23**
10.	**Zypern**	**21**
11.	Slowenien	17
12.	**Griechenland**	**17**
13.	Malta	16
14.	**Portugal**	**16**
15.	Slowakei	13

Euroraum 29

Schweiz	*61*
Vereinigte Staaten	*39*
Großbritannien	*30*

1) Rang der Staaten des Euroraums (ohne Irland und Estland). 2) Stand zwischen 2008 und 2010 (Deutschland, Italien, Zypern und Portugal Stand 2010, Spanien 2008, Griechenland 2009). 3) Median: 50 Prozent der Haushalte eines Landes liegen darüber, 50 Prozent darunter. Je stärker das Durchschnitts- das Medianvermögen übersteigt, desto höher sind in der Regel die Vermögen der Reichen.

Quellen: IWF; EZB / F.A.Z.-Grafik Brocker

Abbildung: Aus der FAZ (http://www.faz.net/-gqe-78jxy)

Die p-Quantile (Lagemaße)

- Innerhalb einer geordneten Datenmenge unterteilen Quantile eine Datenmenge so, dass unter und über dem Quantil bestimmte Anteile an Fällen stehen
- ODER: Der Wert ($= x_p$), der eine geordnete Datenmenge so unterteilt, dass sich links davon höchstens p*100 % der Werte und rechts davon höchstens (1-p)*100 % der Werte befinden. (Behnke/Behnke 2006: 131)

Die p-Quantile (Lagemaße)

- Das 50%-Quantil ($= x_{.50}$) ist daher der **Median**, da es sich genau in der Mitte der Verteilung befindet und rechts und links davon jeweils 50% der Datenwerte vorkommen
- Weitere häufig verwendete Quantile sind
  - $x_{.10}$ = Links von diesem Wert befinden sich in einer geordneten Datenmenge 10% aller Werte der Verteilung
  - $x_{.90}$ = Links von diesem Wert befinden sich in einer geordneten Datenmenge 90% aller Werte der Verteilung ODER Rechts von diesem Wert befinden sich in einer geordneten Datenmenge 10% aller Werte

Die p-Quantile (Lagemaße)

- Weitere häufig verwendete Quantile sind
  - $x_{.25}$ = Links von diesem Wert befinden sich in einer geordneten Datenmenge 25% aller Werte der Verteilung
  - $x_{.75}$ = Links von diesem Wert befinden sich in einer geordneten Datenmenge 75% aller Werte der Verteilung
- Diese beiden Maße nennen sich auch die **Quartile** (da sie die Verteilung in Viertel unterteilen)!
- Die Quartile werden für die Berechnung des **Interquartilsabstands** benötigt!

Die Streuungsmaße

- Spannweite und Interquartilsabstand (oder: Quartilsabstand, vgl. Diaz-Bone 2006)
  - **Boxplot** als Darstellung der Streuung
- Varianz und Standardabweichung
- Schiefe einer Verteilung

Die Spannweite

- Die **Spannweite** (oder: **Range**) bezeichnet den Abstand zwischen größter und kleinster Ausprägung
- Spannweite = größte Ausprägung – kleinste Ausprägung ODER
- Spannweite = $x_1 - x_0$
- Achtung: Spannweite ist ausreißersensitiv!

Der Interquartilsabstand (oder Quartilsabstand)

- Der **Interquartilsabstand** ist der Abstand zwischen dem 25% und dem 75%-Quantil (oder: zwischen den beiden Quartilen)
- Interquartilsabstand (IQA/IQR) $= x_{.75} - x_{.25}$

Der Boxplot (= Fünf-Punkte-Zusammenfassung)

- Spannweite (d.h., höchste und niedrigste Ausprägung), IQR und Median bilden zusammen eine gute Charakterisierung der Streuung einer Variablen (= Fünf-Punkte-Zusammenfassung)
- Die Zusammenfassung umfasst also $x_0, x_{.25}, x_{.50}, x_{.75}, x_1$!
- Diese 5 Punkte können in einem sogenannten **Boxplot** visualisiert werden

Der Boxplot

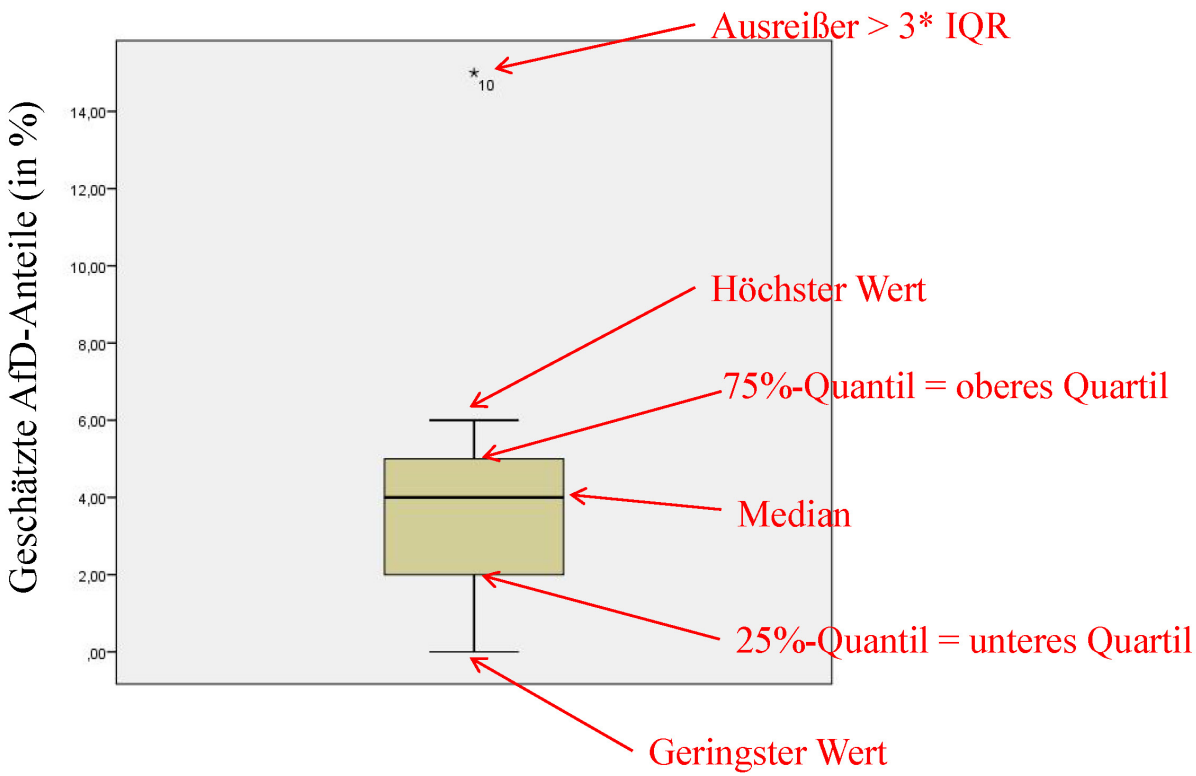


Abbildung: Eigene Darstellung des hypothetischen Beispiels

Varianz und Standardabweichung

- Die **Varianz** und die **Standardabweichung** sind zwei sehr wichtige Konzepte, die eng miteinander zusammenhängen
- Die **Varianz** bezeichnet den 'Mittelwert der quadrierten Abweichungen der individuellen Werte zum Mittelwert der Verteilung' (Behnke/Behnke 2006: 132)
- Formal: $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Varianz und Standardabweichung

- Die **Standardabweichung** stellt die Wurzel aus der Varianz dar
- Formal: $s(x) = \sqrt{var(x)}$

Merksatz Varianz und Standardabweichung

Generell verwendet man eher die Standardabweichung als die Varianz, um die Streuung einer Verteilung auszudrücken!

Varianz und Standardabweichung am Beispiel

	B	F	G	E	D	A	C	H	I
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 6.2$

Varianz und Standardabweichung am Beispiel

	B	F	G	E	D	A	C	H	I
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 6.2$
- $var(x) = \frac{1}{9}((B - \bar{x})^2 + (F - \bar{x})^2 + (...))$

Varianz und Standardabweichung am Beispiel

	B	F	G	E	D	A	C	H	I
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 6.2$
- $var(x) = \frac{1}{9}((B - \bar{x})^2 + (F - \bar{x})^2 + (...)9$
- $var(x) = \frac{1}{9}((0 - 6.2)^2 + (1 - 6.2)^2 + (2 - 6.2)^2 + (3 - 6.2)^2 + (4 - 6.2)^2 + (5 - 6.2)^2 + (5 - 6.2)^2 + (6 - 6.2)^2 + (30 - 6.2)^2)$

Varianz und Standardabweichung am Beispiel

	B	F	G	E	D	A	C	H	I
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 6.2$
- $var(x) = \frac{1}{9}((B - \bar{x})^2 + (F - \bar{x})^2 + (...))$
- $var(x) = \frac{1}{9}((0 - 6.2)^2 + (1 - 6.2)^2 + (2 - 6.2)^2 + (3 - 6.2)^2 + (4 - 6.2)^2 + (5 - 6.2)^2 + (5 - 6.2)^2 + (6 - 6.2)^2 + (30 - 6.2)^2)$
- $var(x) = \frac{1}{9}(38.44 + 27.04 + 17.64 + 10.24 + 4.84 + 1.44 + 1.44 + 0.04 + 566.44)$

Varianz und Standardabweichung am Beispiel

	B	F	G	E	D	A	C	H	I
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 6.2$
- $var(x) = \frac{1}{9}((B - \bar{x})^2 + (F - \bar{x})^2 + (...))$
- $var(x) = \frac{1}{9}((0 - 6.2)^2 + (1 - 6.2)^2 + (2 - 6.2)^2 + (3 - 6.2)^2 + (4 - 6.2)^2 + (5 - 6.2)^2 + (5 - 6.2)^2 + (6 - 6.2)^2 + (30 - 6.2)^2)$
- $var(x) = \frac{1}{9}(38.44 + 27.04 + 17.64 + 10.24 + 4.84 + 1.44 + 1.44 + 0.04 + 566.44)$

Varianz und Standardabweichung am Beispiel

	B	F	G	E	D	A	C	H	I
Erwarteter Anteil an AfD-WählerInnen in %	0	1	2	3	4	5	5	6	30

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 6.2$
- $var(x) = \frac{1}{9}(38.44+27.04+17.64+10.24+4.84+1.44+1.44+0.04+566.44)$
- Varianz: $var(x) \approx 74.17$
- Standardabweichung: $s(x) = \sqrt{var(x)} = \sqrt{74.2} = 8.61$

Varianz / Standardabweichung - Interpretation

	A	B	C	D	E	F	G	H	I
Erwarteter Anteil an AfD-WählerInnen in %	1	1	1	1	1	1	1	1	2

- Anderes Beispiel mit einer sehr homogenen Verteilung
  - Alle schätzen AfD-Wahlerfolg sehr ähnlich (gering) ein
  - Wie wirkt sich das auf Varianz und Standardabweichung aus?

Varianz / Standardabweichung - Interpretation

	A	B	C	D	E	F	G	H	I
Erwarteter Anteil an AfD-WählerInnen in %	1	1	1	1	1	1	1	1	2

- $var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ mit $\bar{x} = 1.1$
- $var(x) = \frac{1}{9}((1-1.1)^2 + (1-1.1)^2 + (1-1.1)^2 + (1-1.1)^2 + (1-1.1)^2 + (1-1.1)^2 + (1-1.1)^2 + (1-1.1)^2 + (2-1.1)^2)$
- $var(x) = \frac{1}{9}(0.89)$
- Varianz: $var(x) \approx 0.10$
- Standardabweichung: $s(x) = \sqrt{var(x)} \approx \sqrt{0.1} \approx 0.32$

Varianz / Standardabweichung - Interpretation

- Varianz und Standardabweichung (StA) sind also nicht nur Maße für den *durchschnittlichen Abstand der Ausprägung eines Falles zum Mittelwert*
- Sie sagen uns vor allem etwas über die **Homogenität / Heterogenität** einer Verteilung
  - Je größer die StA (im Verhältnis zum Mittelwert) ist, desto heterogener ist die Verteilung
  - Vereinfacht: Je größer die StA, desto unterschiedlicher sind die (Ausprägungen der) Fälle im Hinblick auf die untersuchte Variable / Dimension!

Die Schiefe einer Verteilung

- Anhand des Mittelwertes und der Standardabweichung kann zudem noch dargestellt werden, welche **Schiefe** eine Verteilung hat bzw. ob eine Variable symmetrisch, linkssteil (rechtsschief) oder rechtssteil (linksschief) verteilt ist

Abbildung 13.1: Symmetrische Verteilung

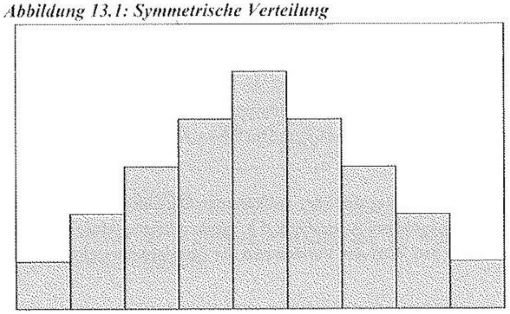


Abbildung 13.2: Linkssteile Verteilung

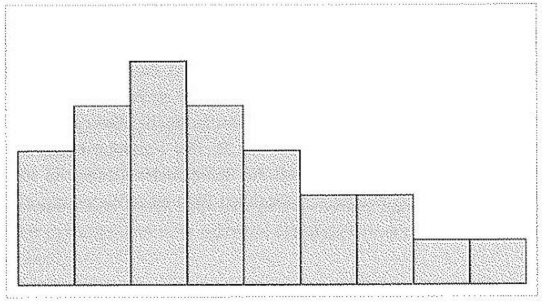
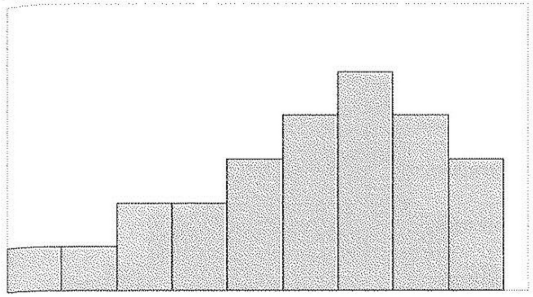


Abbildung 13.3: Rechtssteile Verteilung



Die Schiefe einer Verteilung

- Formal: Schiefe $= \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$
- Für die Werte der Schiefe und ihre Interpretation gilt
  - Schiefe ≈ 0 = Symmetrische Verteilung
  - Schiefe > 0 = Linkssteile Verteilung (rechtsschiefe Verteilung)
  - Schiefe < 0 = Rechtssteile Verteilung (linksschiefe Verteilung)

Besonderheiten / Probleme Streuungsmaße

- Standardabweichung, Varianz und Schiefe basieren stets auf dem arithmetischen Mittelwert einer Verteilung. Sie haben daher zwei besondere Eigenschaften.

Merksatz Streuungsmaße I

Standardabweichung, Varianz und Schiefe dürfen nur für Variablen berechnet werden, die metrisch skaliert sind!

Merksatz Streuungsmaße II

Standardabweichung, Varianz und Schiefe sind ausreißersensitiv – d.h., Ausreißer verzerren alle diese Maße!

Ausblick — Übungsaufgaben

- Auf Learnweb finden Sie inzwischen Übungsaufgaben zu dieser und der vorherigen Sitzung bereitgestellt
  - Lösen Sie zunächst zuhause als Vorbereitung für das Tutorium die Aufgabe 1
  - Im Tutorium werden wir dann die Lösung für Aufgabe 1 besprechen und anhand der Aufgabe 2 weiter in das Programm SPSS einsteigen
  - Nach dem Tutorium wird die Lösungsskizze für beide Aufgaben ebenfalls in Learnweb hochgeladen werden