3. Deskriptive Statistik

„Die deskriptive (auch: beschreibende) Statistik hat zum Ziel, […] *Daten durch Tabellen, Kennzahlen* […] *und Grafiken übersichtlich darzustellen und zu ordnen. Dies ist vor allem bei umfangreichem Datenmaterial sinnvoll, da dieses nicht leicht überblickt werden kann.*“ – Wikipedia.

3.1 Wichtige Größen

3.1.1 Lagemaße

Häufig möchte man einen Parameter bei zwei Gruppen vergleichen, **beispielsweise die Körpergröße von Männern und Frauen**. Eine beliebte Methode hierfür ist die Bildung des (sicher jedem bekannten) **arithmetischen Mittelwerts:** Man addiert die Werte aller Messungen und teilt sie durch die Anzahl der Messungen (s. **Abb. 2**). Der arithmetische Mittelwert darf nur bei **intervallskalierten Daten** angewendet werden. Einfach ausgedrückt sind das Daten, bei denen ein z.B. doppelt so hoher Wert auch tatsächlich bedeutet, dass die Ausprägung des gemessenen Parameters doppelt so groß ist (vgl. mit **ordinal skalierten** Daten).

	Körpergröße in [m]				
Person 1	1,78		1,78	+	
Person 2	1,56		1,56	+	
Person 3	1,80		1,80	+	
Person 4	1,82		1,82	+	
Person 5	1,69		1,69	+	
Person 6	1,78		1,78	+	
Person 7	1,90		1,90	+	
Person 8	1,85		1,85		
Mittelwert:	**1,77**		14,18	: 8	= 1,77

Abb. 2: Der Mittelwert (das **arithmetische Mittel**) der Körpergröße von acht Personen errechnet sich durch die Addition der gemessenen Werte geteilt durch deren Anzahl.

Weniger bekannt dürfte der **Median** sein: Beim Median werden alle vorhandenen Messwerte aufsteigend geordnet. Der Wert, der genau in der Mitte steht, ist der Median (s. **Abb. 3**). Falls eine gerade Anzahl an Werten vorliegt, berechnet sich der Median aus dem Mittelwert der beiden mittleren Werte.

	Anzahl WLAN-fähiger Endgeräte
Haushalt 1	3
Haushalt 2	4
Haushalt 3	3
Haushalt 4	1
Haushalt 5	4
Haushalt 6	3
Haushalt 7	5
Median:	**3**

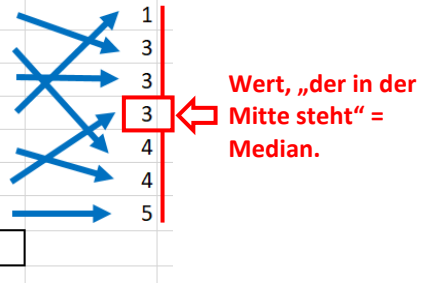

Abb. 3: Der **Median** ist derjenige Wert, der nach dem aufsteigenden Sortieren aller Werte in der Mitte steht.

Der Median kann auch bei **ordinal skalierten Daten** angewendet werden. Bei diesen Daten geht es nur darum, eine Reihenfolge bzw. Rangfolge festzustellen: Bei einem Autorennen kann ein **erster, zweiter und dritter Platz** vergeben werden. Dabei spielt es keine Rolle, ob der Erstplatzierte eine Stunde oder eine Minute schneller war als der Zweitplatzierte. Häufig kommen ordinal skalierte Daten auch bei Umfragen zum Einsatz:

Frage: Wie häufig nutzen Sie Ihr Handy, um Ihren Kontostand abzurufen?
Antwortmöglichkeiten: **sehr oft (1), oft (2), gelegentlich (3), selten (4), nie (5)**.

Über die Abstände zwischen den vorgegebenen Antwortmöglichkeiten lässt sich kaum eine Aussage machen: Ob jemand der „gelegentlich“ seinen Kontostand per Handy checkt, das doppelt so oft oder viermal so oft tut, wie jemand der „selten“ seinen Kontostand am Handy checkt, bleibt unklar. Auch **Schulnoten** sind ordinal skalierte Daten. Man kann nämlich z.B. nicht sagen, dass jemand der die Note 3 erhalten hat, dreimal so viele Fehler gemacht hat, wie jemand der die Note 1 erhalten hat. Insofern ist die Berechnung von arithmetischen Mittelwerten bei Schulnoten aus statistischer Sicht eigentlich unzulässig (s. **Abb. 4**).

	Schüler 1	Schüler 2	Schüler 3
Note 1	2	2	2
Note 2	3	2	3
Note 3	3	2	2
Note 4	4	4	4
Note 5	4	4	4
Note 6	2	2	2
arith. Mittelwert	3,0	2,7	2,8
Median	3,0	2,0	2,5

Abb. 4: Vergleich von **Median** und **arithmetischem Mittelwert** bei Schulnoten verschiedener Schüler.

Der Median ist sehr **robust** gegenüber **Ausreißern**. Das sind einzelne Werte, die sich deutlich von den anderen Messwerten einer Gruppe unterscheiden. Daher ist der Median in der Biologie oft beliebt, da gerade beim Umgang mit Lebewesen immer wieder Besonderheiten auftreten (s. **Abb.5**).

	Häufigkeit des Gefiederputzens bei Vögeln pro Stunde
Vogel 1	3
Vogel 2	4
Vogel 3	3
Vogel 4	1
Vogel 5	4
Vogel 6	3
Vogel 7	4
arith. Mittelwert	3,14
Median	3,00

		Häufigkeit des Gefiederputzens bei Vögeln pro Stunde
	Vogel 1	3
	Vogel 2	4
	Vogel 3	3
	Vogel 4	1
	Vogel 5	4
	Vogel 6	3
Ausreißer:	Vogel 7	75
arith. Mittelwert		13,29
Median		3,00

Abb. 5: Einzelne **Ausreißer** haben oft große Auswirkungen auf den Mittelwert, nicht jedoch auf den Median.

3.1.2 Streuungsmaße

Der Mittelwert oder Median alleine sagt noch wenig aus. Manchmal liegen die gemessenen Werte nämlich eng am tatsächlichen Mittelwert, manchmal **streuen** sie aber auch stark (s. **Abb. 6**).

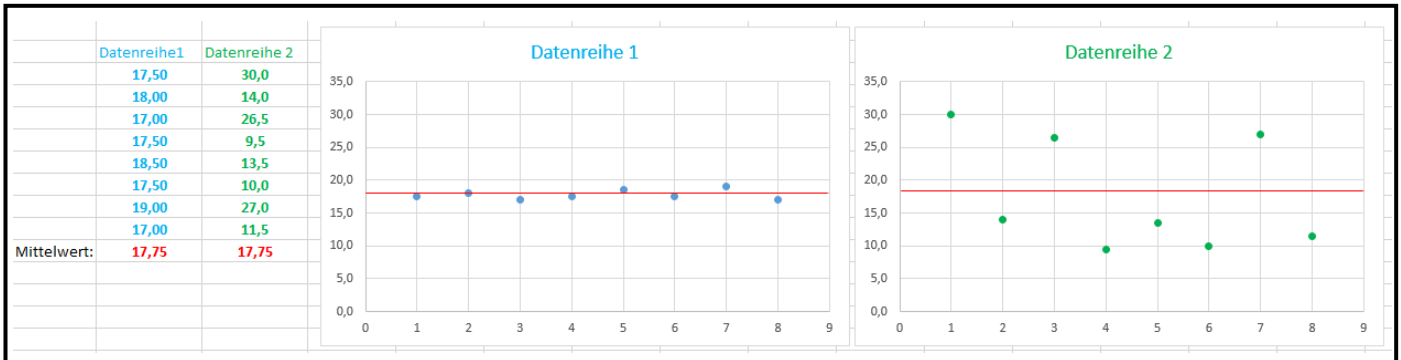

Abb. 6: Bei beiden hier dargestellten Datenreihen erhält man als Mittelwert 17,75. Allerdings **streuen** die Einzelwerte bei der Datenreihe 2 sehr viel stärker als bei Datenreihe 1.

Es gibt verschiedene Maße, die in der Lage sind diese Streuung in Zahlen auszudrücken: Zum Beispiel die **Varianz,** die vereinfacht ausgedrückt ungefähr den Mittelwert der Abweichungsquadrate darstellt. Aber auch die (empirische) **Standardabweichung**, die als Wurzel aus der Varianz definiert ist. Beide Parameter kann man in Tabellenkalkulationsprogrammen in der Regel einfach berechnen lassen (s. **Abb. 7**).

	Datenreihe1	Datenreihe 2
	17,5	13,5
	18,0	11,5
	17,0	30,0
	17,5	26,5
	18,5	9,5
	17,5	27,0
	19,0	14,0
	17,0	10,0
Mittelwert:	17,8	17,75
Standardabweichung:	0,71	8,55
Varianz:	0,50	73,07

	Datenreihe 2	Abweichung vom Mittelwert (=17,75)	Quadrat der Abweichung	
	13,5	-4,3	18,0625	
	11,5	-6,3	39,0625	
	30,0	12,3	150,0625	
	26,5	8,8	76,5625	
	9,5	-8,3	68,0625	
	27,0	9,3	85,5625	
	14,0	-3,8	14,0625	
	10,0	-7,8	60,0625	
arith. Mittelwert	17,75		73,07	Mittelwert* der Quadrate

Abb. 7: *links*: Die Varianz und die Standardabweichung kann man von Tabellenkalkulationsprogrammen einfach über eine Formel in der entsprechenden Zelle berechnen lassen.

rechts: Veranschaulichung des mathematischen Verfahrens zur Berechnung der Varianz: Von jedem Messwert wird die Abweichung (Differenz) zum Mittelwert bestimmt. Dieser Wert wird quadriert. Der Mittelwert der Quadrate entspricht der Varianz.
* Wer genau nachrechnet, erhält für den Mittelwert hier eigentlich 63,94. Dies liegt daran, dass man bei der Berechnung dieses speziellen Mittelwerts die Summe nicht durch die Anzahl der Messungen teilt, sondern durch die um 1 verringerte Zahl (hier also: Summe geteilt durch 7, obwohl 8 Messungen vorliegen).

Auch bei ordinal skalierten Daten, kann man neben dem Median ein Streuungsmaß angeben: Den Interquartilsabstand. Die Quartile sind vereinfacht ausgedrückt die Werte unterhalb derer 25% (1. Quartil) bzw. 75% (3. Quartil) aller anderen Messwerte liegen.

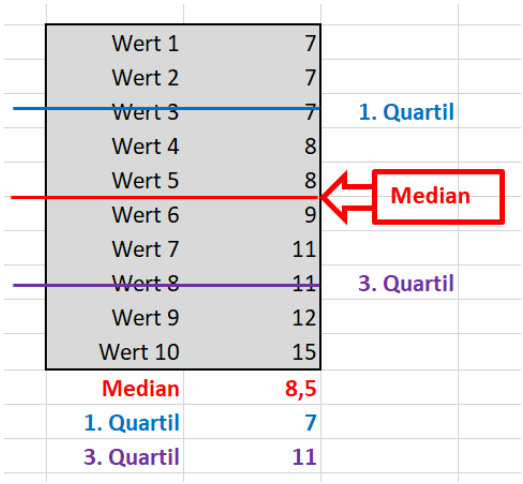

Abb. 8: Median und **Quartile** einer Datenreihe. Üblicherweise können Quartile sinnvoll erst ab ca. 20 Werten verwendet werden.

3.2 Diagrammtypen

Es gibt Standard-Diagrammtypen, die mit Tabellenkalkulationsprogrammen in der Regel leicht erstellt werden können. Jeder Diagrammtyp hat Vor- und Nachteile, die gegeneinander abgewogen werden müssen, eine detaillierte Betrachtung erfolgt hier jedoch nicht. Lediglich auf zwei typische Fehlerquellen soll hier hingewiesen werden.

Unterbrochene Skalierung der y-Achse

Gelegentlich beginnt in Diagrammen die y-Achse im Ursprung nicht mit dem Wert 0. Diese Darstellung wird dann gewählt, wenn man eigentlich kleine Unterschiede deutlich machen möchte. Es ist zwar nicht wirklich ein Fehler, sollte aber klar gekennzeichnet werden, z.B. durch eine Unterbrechung in der y-Achse (s. **Abb. 9**).

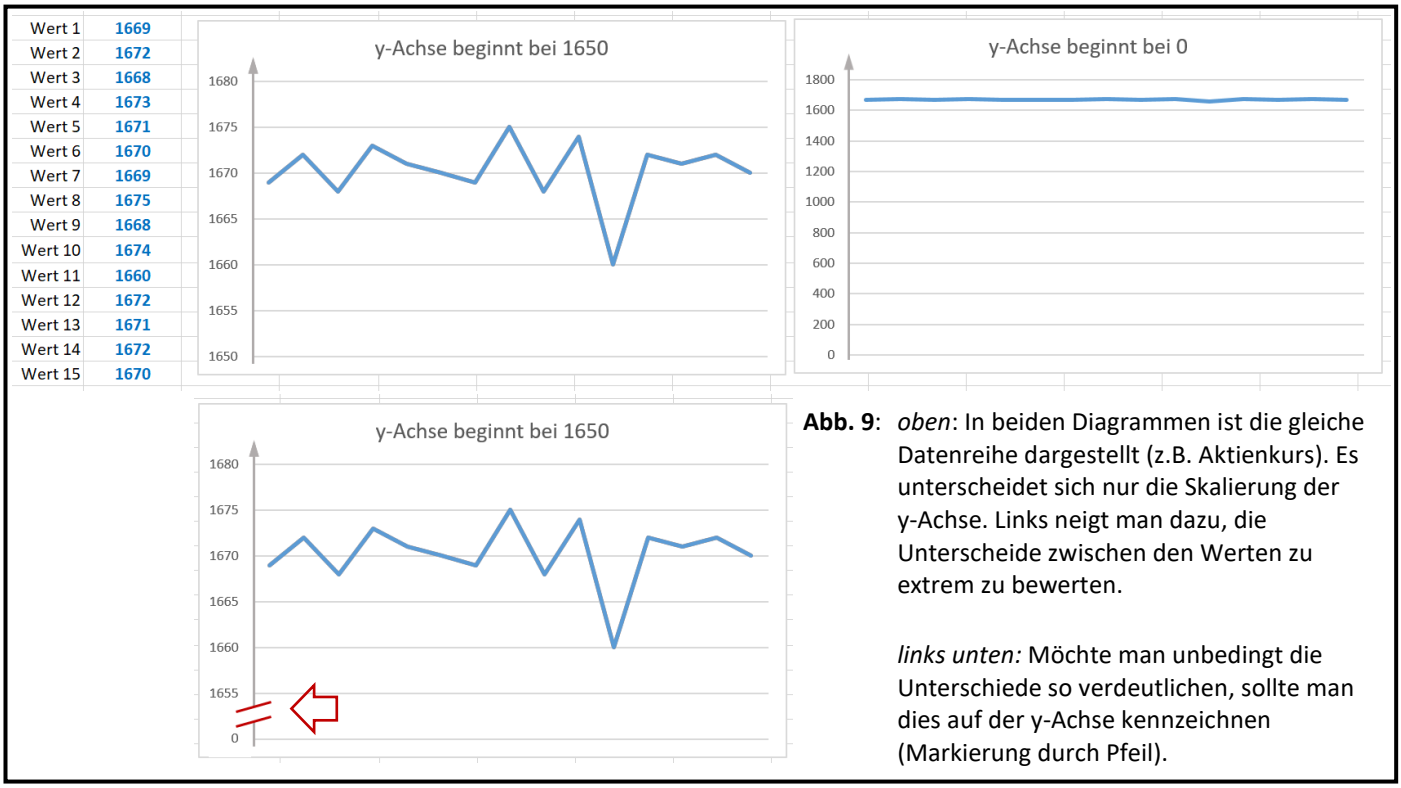

Abb. 9: *oben*: In beiden Diagrammen ist die gleiche Datenreihe dargestellt (z.B. Aktienkurs). Es unterscheidet sich nur die Skalierung der y-Achse. Links neigt man dazu, die Unterscheide zwischen den Werten zu extrem zu bewerten.

links unten: Möchte man unbedingt die Unterschiede so verdeutlichen, sollte man dies auf der y-Achse kennzeichnen (Markierung durch Pfeil).

Falscher Diagrammtyp

Liniendiagramme bei denen Datenpunkte **verbunden** sind, dürfen nur verwendet werden, wenn zwischen den Datenpunkten auch tatsächlich eine **Verbindung** besteht, z.B. wenn sie zeitlich aufeinander folgen. Werden unterschiedliche Gruppen betrachtet, sind Säulendiagramme o.ä. zu verwenden.

Kreis- bzw. Tortendiagramme können nur verwendet werden, wenn die dargestellten Werte als Summe tatsächlich eine 100%-Gesamtheit ergeben.