Grundlagen und Anwendung der Wahrscheinlichkeitstheorie

WS 24/25

Datenprojekt
Gruppe 09

Paul Franzen, Ben Caspar, Johannes Christmann

R1.1:

Dieser Datensatz enthält eine Tabelle mit dem Thema „Household prizes“ mit zwei Spalten, welche Datum und „Percent change from a year ago“ vereinigt. Dabei läuft das Datum nicht tagesweise fort, sondern liegt in Quartal-Sprüngen vor, d.h. immer der erste Tag eines n+1-ten Quartals im Jahr folgt auf den ersten Tag des n-ten Quartals. Das Datum ist hierbei im Format DD-MM-YYYY gegeben und reicht vom Jahr 2000 bis 2023. Die zweite Spalte gibt eine prozentuale Veränderung einer unbekannten Größe, am wahrscheinlichsten aber eine Änderung der Haushaltspreise vom jeweils vorherigen Jahr, an, welche neben dem positiven Bereich auch in den negativen Bereich reicht. Der Prozentsatz liegt in der Größenordnung von etwa -15% bis +9%. Der Datensatz stammt von der Online-Database FRED, kurz Federal Reserve Economic Data. Diese wird bereitgestellt von der Federal Reserve Bank of St. Louis. (https://fred.stlouisfed.org/series/QDER628BIS#0)
In Quelltext-Ansicht besteht diese Datei mit dem Namen „data-1.csv“ aus 94 Zeilen Code (Datum mit jeweils zugehörigem Prozentsatz) und ist in UTF8-Format kodiert.

R1.2:
Der Variable des Prozentsatzes ist sinnvollerweise eine Verhältnisskala zuzuordnen. Diese Variable ist abhängig vom Datum, welches auf einer Intervallskala geführt wird.

R1.3:
Es wurde Python in VS-Code verwendet mit folgenden Bibliotheken:

- Pandas
- Scipy
- Numpy
- Matplotlib
- Statistics

Zudem wurde auch Excel, seine gängigen Funktionen sowie Statistik Add-ins verwendet.

R1.7:

Modus:	Es gibt keinen Modus, da kein Wert doppelt vorkommt. Alternativ ist jeder Wert ein Modus, da alle Werte mit der Häufigkeit eins vorkommen.
Arithmetischer Mittelwert:	0,67639
Median:	0,25056

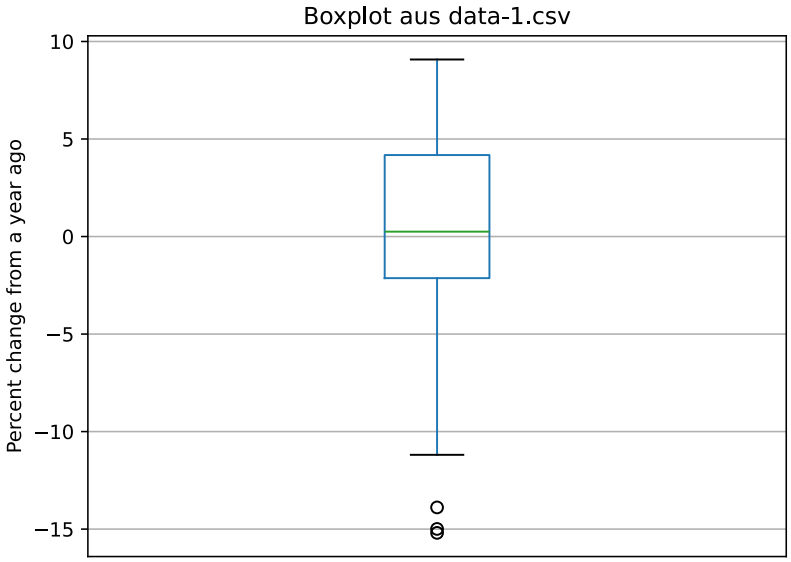
R1.8:
Die Spannweite beträgt: 24,27

R1.9:
Die mittlere Abweichung vom Median beträgt: 3,463.

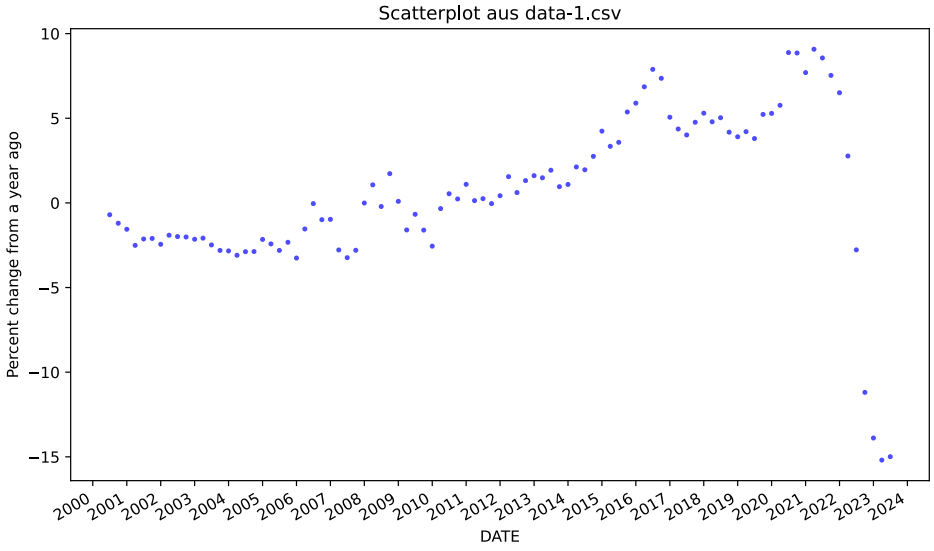
R1.10:
Die Stichprobenvarianz beträgt: 21,83.

R1.11:
Der Variationskoeffizient beträgt: 6,908.

R1.12:



R1.13:



R1.14:
Da alle Datenwerte gleichhäufig vorkommen, gibt es keinen Modus. Der Median liegt, wie man im Box-Plot erkennen kann, etwa um 0,4 unter dem arithmetischen Mittelwert und damit näher am Q(0,25) als an der oberen Grenze der Box.
Die Spannweite der Prozentsätze beträgt 24,27 und reicht aus dem negativen bis in den Positiven Bereich. Der Variationskoeffizient gibt das relative Streuungsmaß mit 6,908 an, wohingegen die Stichprobenvarianz die relative quadratische Abweichung vom Mittelwert mit 21,83 angibt.

R1.15:

Quartile:	Q(0,25): -2,13345
	Q(0,5): 0,25056
	Q(0,75): 4,17768
Dezile:	D(0,1): -2,83
	D(0,2): - 2,43598
	D(0,3): -1,94102
	D(0,4): -0,6789
	D(0,5): 0,2506
	D(0,6): 1,4968
	D(0,7): 3,4379
	D(0,8): 4,7804
	D(0,9): 6,3843

R1.16:
Der Quartilsabstand R_Q0.5 beträgt: 6,31113

R1.17:
Die Kovarianz beträgt: 4 216,675

R1.18:
Der Korrelationskoeffizient: 0,3662

R1.19:
Klasse 1: [-15,19326 ; -10)

Klasse 2: [-10 ; -5)

Klasse 3: [-5 ; 0)

Klasse 4: [0 ; 5)

Klasse 5: [5 ; 9,07686]

R1.20:

Kontingenztabelle					
Jahre/Klas	Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5
2000			2		
2001			4		
2002			4		
2003			4		
2004			4		
2005			4		
2006			4		
2007			4		
2008			2	2	
2009			3	1	
2010			2	2	
2011			1	3	
2012				4	
2013				4	
2014				4	
2015				3	1
2016					4
2017				3	1
2018				2	2
2019				3	1
2020					4
2021					4
2022	2		1	1	
2023	2				

R1.21:
Der Rangkorrelationskoeffizient nach Spearman beträgt etwa 0.14157.

R2.1:

Dieser Datensatz besteht wieder aus einer Tabelle mit zwei Spalten, welche fehlerbehaftet ist sowohl spalten- als auch zeilenweise. Teils fehlen Dateneinträge, Jahre sind nicht fortlaufend, sondern springen um 100 Jahre, oder ähnliches.
Grundlegend sind die Lebendgeburtenzahlen jährlich auf die Zeit von 1950 bis 2022 aufgetragen. Die Daten stammen vom Statistischen Bundesamt aus der Statistik „Genesis Tabelle 12612-0001“ (https://www-genesis.destatis.de/datenbank/online/statistic/12612/table/12612-0001).
Der Datensatz liegt wieder in UTF8-Kodierung vor.
Der Quellcode der CSV-Datei besteht aus 75 Zeilen Code.

R2.3:
Zunächst wurde ein Python-Programm geschrieben, um Zeileneinträge anzupassen. Außerdem wurden die Spaltennamen angepasst und von Hand Jahreszahlen gelöscht, die Fehlerbehaftet waren und das Plotten gestört hätte.

R2.4:
Es wurde Python in VS-Code verwendet mit folgenden Bibliotheken:

- Pandas
- Scipy
- Numpy
- Matplotlib
- Statistics

Zudem wurde auch Excel, seine gängigen Funktionen sowie Statistik Add-ins verwendet.

R2.8:

Modus:	kein Modus, weil sich kein Wert wiederholt
Arithmetischer Mittelwert:	911 790,8254
Median:	812 292

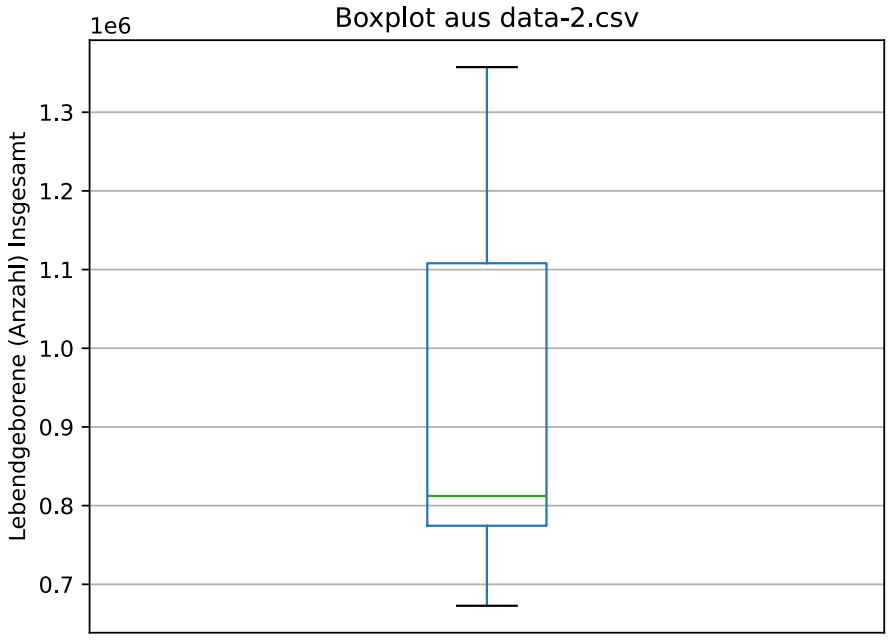
R2.9:
Die Spannweite beträgt: 684 580

R2.10:
Die mittlere Abweichung vom Median beträgt: 157 733,1

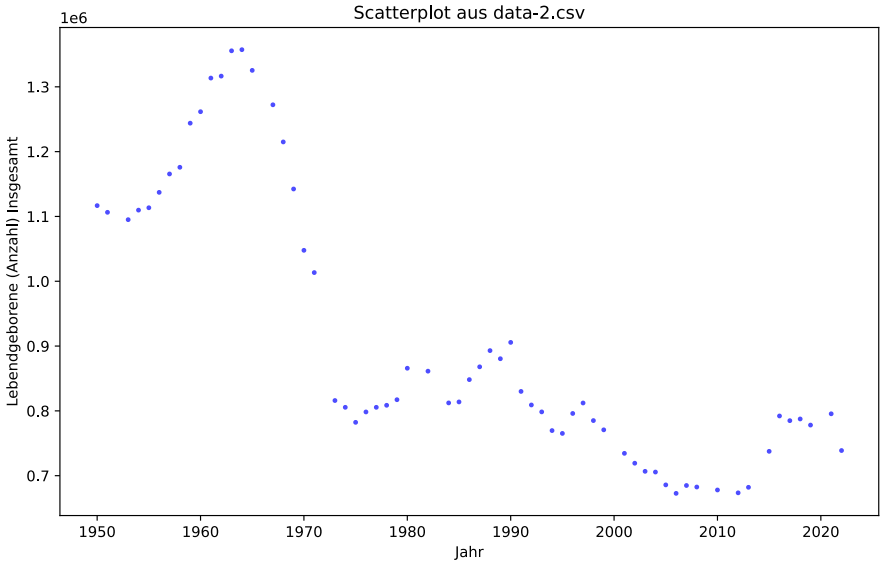
R2.11:
Die Stichprobenvarianz beträgt: 43 824 145 922,308

R2.12:
Der Variationskoeffizient beträgt: 0,22959452

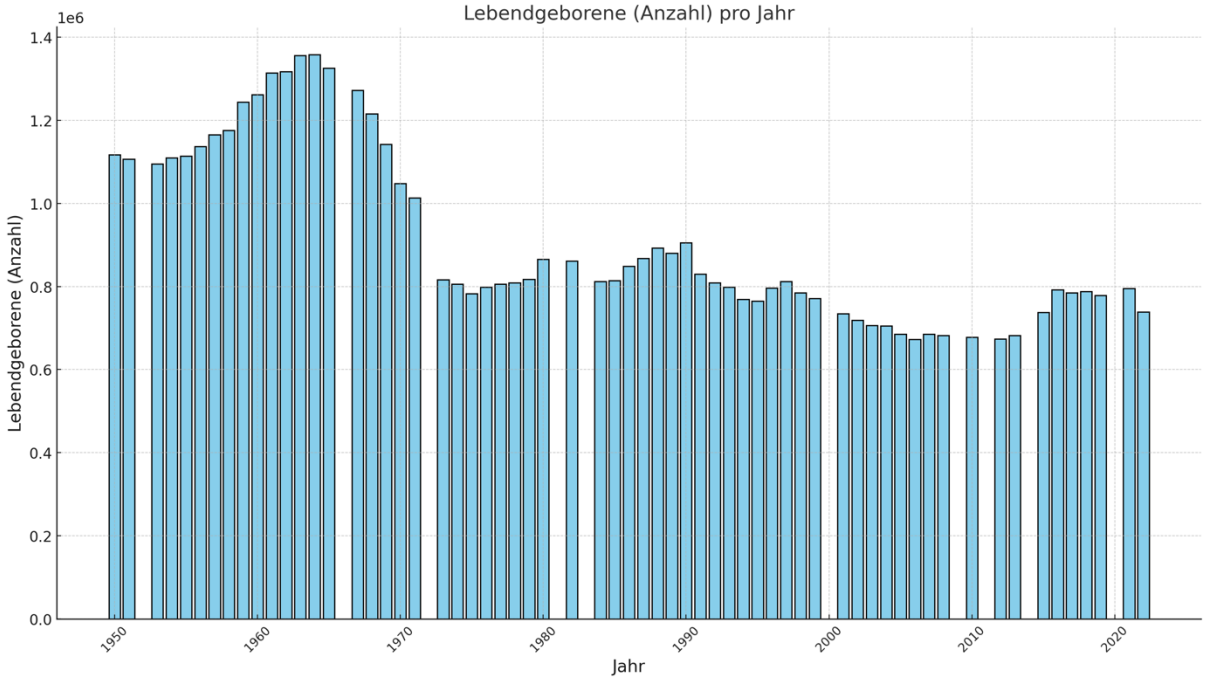
R2.13:



R2.14:



R2.15:



Die X-Achse trägt die Jahre gegenüber der Anzahl von Lebendgeborenen auf.

R2.16:

Es liegt auch hier kein Modus vor, da alle Werte nur einzeln vorkommen.
Der arithmetische Mittelwert beläuft sich auf 911 790,8254 und liegt über dem Median, welcher den Wert 812 292 annimmt.
Die Spannweite beläuft sich auf 684 580 und die mittlere Abweichung vom Median beträgt 157 733,1. Die Stichprobenvarianz liegt bei (schnallen Sie sich an) 43 824 145 922,308.
Der Variationskoeffizient beläuft sich hierbei auf 0,22959452.

R2.17:

Quartile: Q(0,25): 774 417
Q(0,5): 812 292
Q(0,75): 1 108 061,5

Dezile: D(0,1): 689 760,4
D(0,2): 749 379,8
D(0,3): 784 980,8
D(0,4): 798 424,4
D(0,5): 812 292,0
D(0,6): 862 177,8
D(0,7): 1 027 132,4
D(0,8): 1 128 981,8
D(0,9): 1 258 075,6

R2.18:
Der Quartilsabstand R_Q0,5 beträgt: 333 644,5

R2.19:
Die Kovarianz beträgt: -3 607 257,51

R2.20:
Der Korrelationskoeffizient beträgt: -0,8212.

R3.1:

Der dritte zweigeteilte Datensatz besteht aus den Dateien „data-3-a.csv“ sowie aus „data-3-b.csv“. Im A-Teil befindet sich eine Tabelle mit zwei Spalten ohne jegliche Bezeichnung. Hierbei sind in der ersten Spalte acht-stellige Kombinationen aus Großbuchstaben und Ziffern eingetragen, wohingegen in der zweiten Spalte die Jahreszahlen von 1950 bis 2022 aufgeführt sind. Der Teil B des dritten Datensatzes enthält ebenso eine solche zweispaltige Tabelle, welche eben solche Buchstaben-Ziffern-Kombinationen gegenüber einer Anzahl von Gestorbenen aufträgt. Die Anzahl an Zeilen ist hierbei gleich der Anzahl des A-Teils.
Die Quelle ist hier wieder das Statistische Bundesamt mit seiner Satistik „Genesis Tabelle 12613-0001“ über die Sterbefälle von 1950 bis heute (https://www-genesis.destatis.de/datenbank/online/table/12613-0001/search/s/MTI2MTMtMDAwMQ==).
Der Datensatz liegt auch hier in UTF8-Kodierung vor und beinhaltet 75 Zeilen Code.

R3.4:
Es wurde ein Programm geschrieben, um die Datensätze miteinander zu verknüpfen.
Die Buchstaben-Ziffern-Codes haben jeweils angegeben, welches Jahr mit welchem Wert verbunden ist.

R3.6:
Es wurde Python in VS-Code verwendet mit folgenden Bibliotheken:

- Pandas
- Scipy
- Numpy
- Matplotlib
- Statistics

Zudem wurde auch Excel, seine gängigen Funktionen sowie Statistik Add-ins verwendet.

R3.9:

Modus:	kein Modus, da alle Werte gleich häufig vorkommen
arithmetischer Mittelwert:	892 568,3836
Median:	895 070

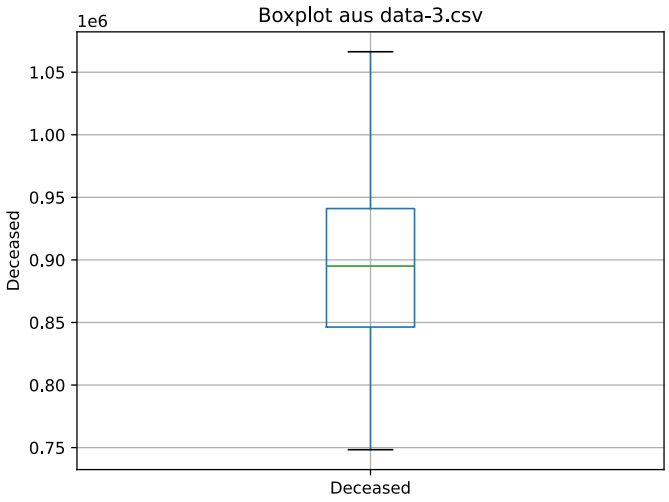
R3.10:
Die Spannweite beträgt: 318 012

R3.11:
Die mittlere Abweichung vom Median beträgt: 52 593,2877

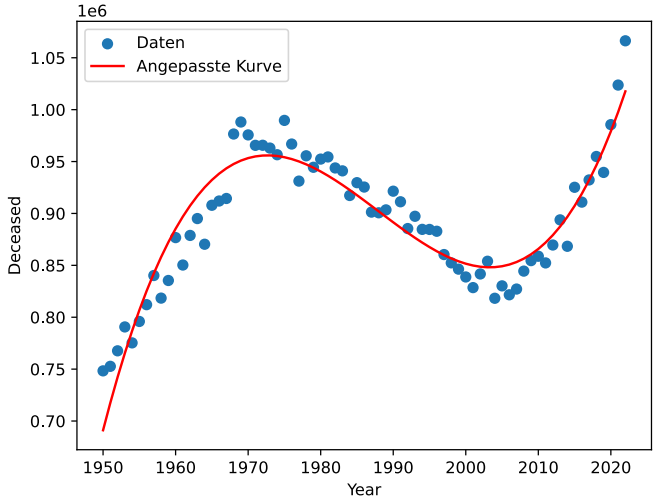
R3.12:
Die Stichprobenvarianz beträgt: 4 189 857 093,379

R3.13:
Der Variationskoeffizient beträgt: 0,07252

R3.14:



R3.15:



R3.17:
$7.505471447409998*X^3+-855.4214093339821*X^2+27216.86531511555*X+691116.8417620478$

R3.19:
Es liegt auch hier kein Modus vor, da alle Werte nur einzeln vorkommen.
Der arithmetische Mittelwert beläuft sich auf 892 568,3836 und liegt leicht unter dem Median, welcher den Wert 895 070 annimmt.
Die Spannweite beläuft sich auf 318 012 und die mittlere Abweichung vom Median beträgt 52 593,2877. Die Stichprobenvarianz liegt bei (schnallen Sie sich wieder ab) 4 189 857 093,379.
Der Variationskoeffizient beläuft sich hierbei auf 0,07252.

R3.20:

Quartile:	Q(0,25):	846 330
	Q(0,5):	895 070
	Q(0,75):	941 032
Dezile:	D(0,1):	818 300,4
	D(0,2):	839 356,2
	D(0,3):	853 320,4
	D(0,4):	8754 40,6
	D(0,5):	895 070,0
	D(0,6):	911 392,8
	D(0,7):	930 251,4
	D(0,8):	953 610,0
	D(0,9):	966 636,2

R3.21:
Der Quartilsabstand R_Q0,5 beträgt: 94 702

R3.22:
Die Kovarianz beträgt: 362 236,94

R3.23:
Der Korrelationskoeffizient beträgt:0,263757

R4.3:
Die Datenmenge wurde in ihren Werten reduziert.

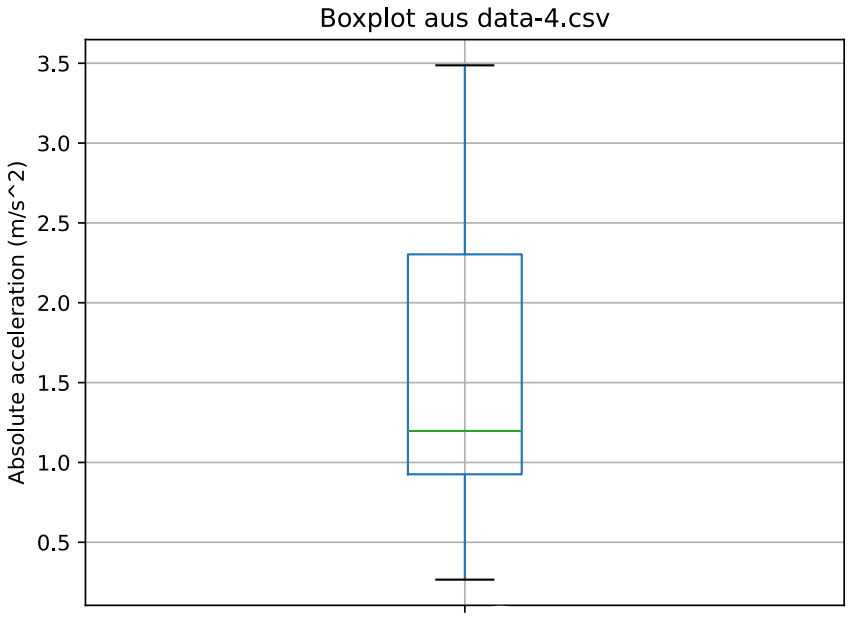
R4.4:
- Phyphox (Beschleunigung (ohne g))
- VS-Code
- Microsoft-Excel

R4.5:

Modi:	1,105; 1,106; 1,135; 1,149, 1,197
Arithmetischer Mittelwert:	1,578
Median:	1,197

R4.6:
Die Stichprobenvarianz beträgt: 0,8461

R4.7:



R4.8:

Es liegen zwei Modi (1,105 und 1,135) vor, welche jeweils zweimal in den Messwerten vorkommen. Der Arithmetische Mittelwert liegt mit etwa 1,4 m/s^2 unter dem Median von etwa 1,75 m/s^2. Die Stichprobenvarianz beschreibt die mittlere Abweichung der gemessenen Werte vom empirischen Mittelwert und beträgt in dieser Datenerhebung etwa 0,94.