

# **Datenprojekt Grundlagen und Anwendung der Wahrscheinlichkeitstheorie**

**Gruppe 5**

**Juriaan Muijsson**

**Friedo Schmitt**

**Jost Kracht**

**Wintersemester 2023/24**

### Quellenverzeichnis:

Oestreich, Markus/ Romberg, Oliver. (2014). Keine Panik vor Statistik! (5. Aufl.). SpringerSpektrum

Aufname der Daten für R4 mit phyphox: <https://phyphox.org/de/home-de/>

Zur Dateien-Struktur:

Es sind i.d.R. alle Punkte des Lastenhefts zu den 4 Datensätzen (R1 bis R4) in diesem Dokument aufgelistet.

Für erstellte csv-Dateien sind Verweise anhand der Dateinamen vorhanden, diese sowie alle anderen relevanten Dateien finden sich im Hauptordner zum jeweiligen Datensatz (R1, R2, R3, R4).

In den ignore-Ordnern befinden sich Dateien, die für die Abgabe nicht relevant sind, aber im Zuge der Bearbeitung entstanden sind, bspw. Zwischenergebnisse oder die Konsolenausgaben der Programme.

**R1.1:**

Daten des Statistischen Bundesamts, veröffentlicht am 12. September 2022

2 Spalten; Jahr (2015 bis 2021) und zugehörige Exportquote in Prozent

7 Datenpaare

Die Daten liegen im csv-Format vor und befinden sich in einer Datei.

**R1.2:**

Skalierung Jahreszahlen: Intervallskalierung

Skalierung Exportquote: Verhältnisskalierung

**R1.3**

Software und genutzte Funktionen:

Programmierung per PyCharm mit Python.

Bibliotheken:

- csv; Einlesen der csv-Dateien

- math; Verschiedene mathematische Operatoren

- os; Überprüfung ob angegebene Datei existiert

- matplotlib; Erstellen von Boxplot, Scatterplot...

**R1.4**

data-1-urliste.csv

**R1.5**

rangliste\_Jahr.csv

rangliste\_Exportquote.csv

**R1.6**

data-1.xlsx

**R1.7:**

Modus: -

Arithmetischer Mittelwert Jahr: 2018

Arithmetischer Mittelwert Exportquote: 46,314

Median Jahr: 2018

Median Exportquote: 46,9

**R1.8:**

Spannweite Jahr: 6

Spannweite Exportquote: 4,3

**R1.9:**

Mittlere Abweichung vom Median

Jahr: 1,714

Exportquote: 0,814

**R1.10:**

Stichprobenvarianz

Jahr: 4

Exportquote: 1,964

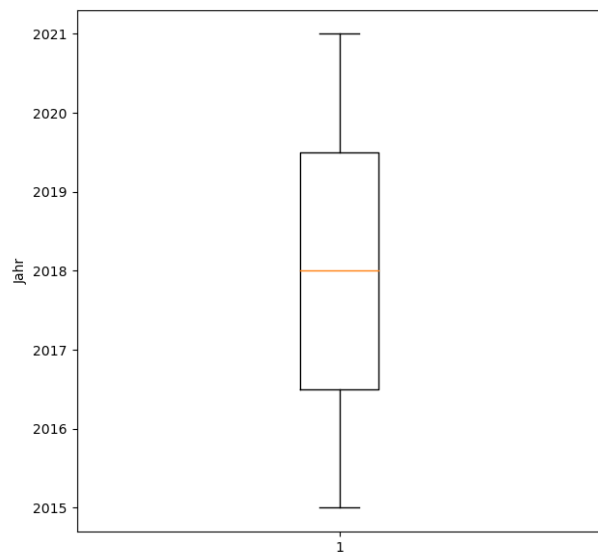
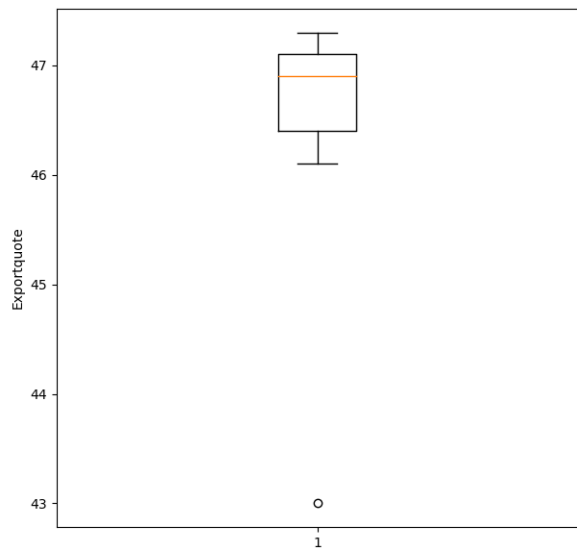
**R.11:**

Variationskoeffizient

Jahr: 0,001

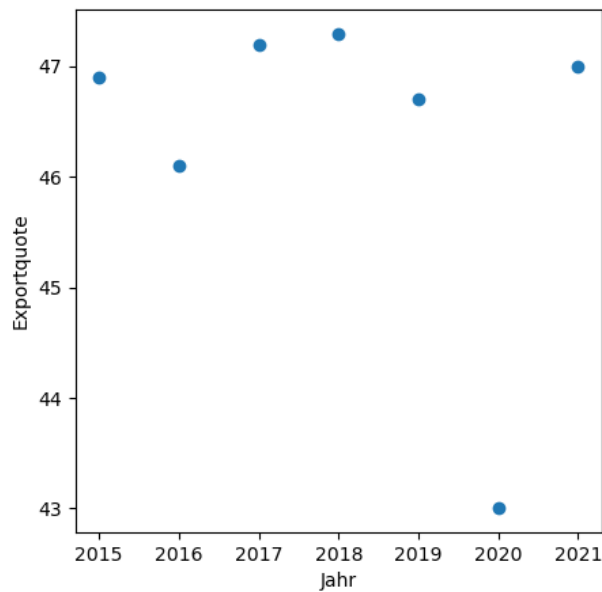
Exportquote: 0,03

### R1.12:



(boxplot\_Exportquote.png

boxplot\_Jahr.png)

**R1.13:**

(Scatterplot\_data\_1.png)

**R1.14:**

Textuelle Zusammenfassung:

Da die Daten nur 7 Datenpunkte umfassen, lässt sich zu den Daten wenig sinnvolles bzw. aussagekräftiges zusammenfassen.

Der Ausreißer-Wert 43.0 verzerrt allgemein die Datenauswertung.

Insb., weil so wenig Datenpunkte vorhanden sind, hat dieser eine Ausreißer eine relativ große Auswirkung.

Bis auf 43.0 ist die Exportquote relativ eng um den Mittelwert/ Median verteilt, bzw. relativ glatt verteilt.

Die Spannweite ohne diesen ist 1.2, mit diesem 4,3.

Auch die negative Korrelation lässt sich vor allem durch den Ausreißer-Wert 43.0 erklären.

**R1.15:**

Dezile Jahr:

Q0.1: 2015

Q0.2: 2016

Q0.3: 2017

Q0.4: 2017

Q0.5: 2018

Q0.6: 2019

Q0.7: 2019

Q0.8: 2020

Q0.9: 2021

Quartile Jahr:

Q0.25: 2016

Q0.5: 2018

Q0.75: 2020

Dezile Exportquote:

Q0.1: 43.0

Q0.2: 46.1

Q0.3: 46.7

Q0.4: 46.7

Q0.5: 46.9

Q0.6: 47.0

Q0.7: 47.0

Q0.8: 47.2

Q0.9: 47.3

Quartile Exportquote:

Q0.25: 46.1

Q0.5: 46.9

Q0.75: 47.2



**R1.16:**

Quartilsabstand Jahre: 4

Quartilsabstand Exportquote: 1.1

**R1.17:**

Kovarianz: -1.07

**R1.18:**

Korrelationskoeffizient: -0.38

**R1.19, R1.20, R1.21:**

Im Zuge der „Kann“-Kennzeichnung haben wir eine Klasseneinteilung bei 7 Datenpunkten als wenig sinnvoll erachtet.

**R2.1:**

Daten des Statistischen Bundesamts, veröffentlicht am 12. September 2022

2 Spalten; Jahr (2015 bis 2021) und zugehörige Exportquote in Prozent

7 Datenpaare

Fehlerhafte/ unvollständige Daten vorhanden

Die Daten liegen im csv-Format vor und befinden sich in einer Datei.

**R2.2:**

data-2-bereinigt.csv

**R2.3:**

Maßnahmen zur Datenbereinigung:

Im Code werden Werte falschen Datentyps mithilfe linearer Interpolation ersetzt.

Dies wurde als sinnvoll erachtet, da die übrigen Exportquoten auf einer glatten Kurve verlaufen bzw. wenig Abweichung vorhanden ist.

Ist der erste oder letzte Wert fehlerhaft, wird der Wert danach bzw. davor übernommen.

**R2.4:**

Software und genutzte Funktionen:

Programmierung per PyCharm mit Python.

Bibliotheken:

csv; Einlesen der csv-Dateien

math; Verschiedene mathematische Operatoren

os; Überprüfung ob angegebene Datei existiert

matplotlib; Erstellen von Boxplot, Scatterplot...

**R2.7:**

Ranglisten:

rangliste\_Exportquote.csv

rangliste\_Jahr.csv

**R2.8:**

Jahr:

Modus: -

Arithmetische Mittelwert: 2018

Median: 2018

Exportquote:

Modus: -

Median Exportquote: 47.0

Arithmetische Mittelwert Exportquote: 47.0

**R2.9:**

Spannweite Jahr: 6.0

Spannweite Exportquote: 0.6

**R2.10:**

Mittlere Abweichung vom Median Jahr: 1.71

Mittlere Abweichung vom Median Exportquote: 0.16

**R2.11:**

Stichprobenvarianz Jahr: 4.0

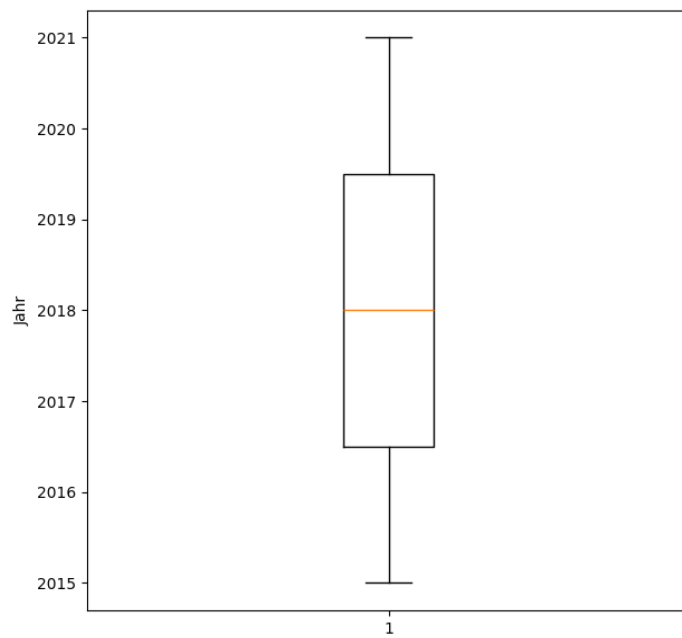
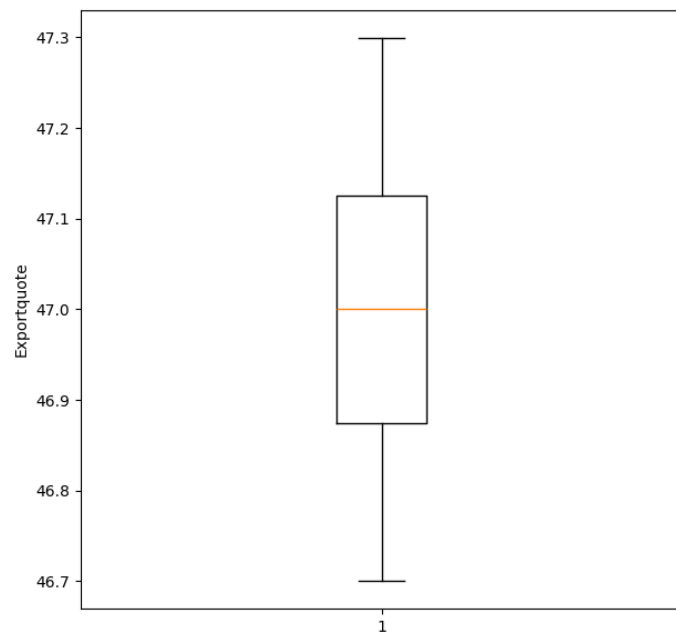
Stichprobenvarianz Exportquote: 0.04

**R2.12:**

Variationskoeffizient Jahr: 0.0

Variationskoeffizient Exportquote: 0.0

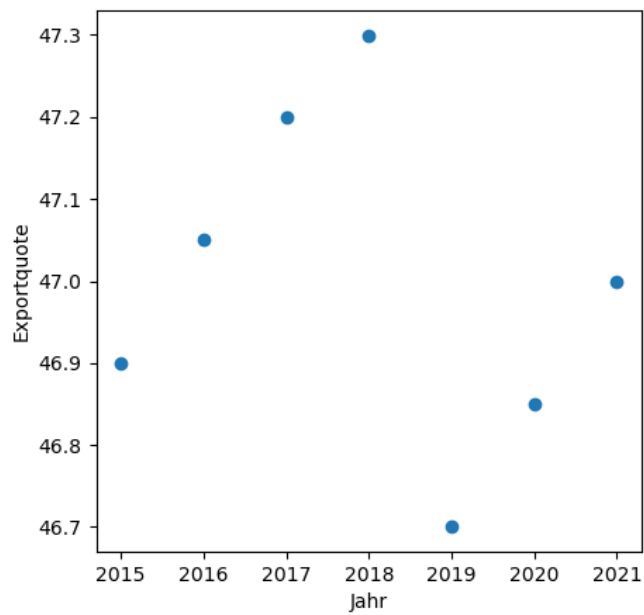
**R2.13:**



(boxplot\_Exportquote.png

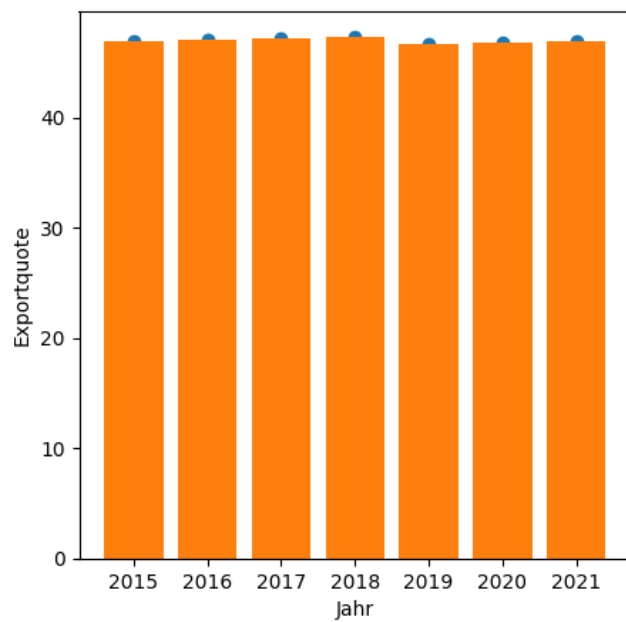
boxplot\_Jahr.png)

**R2.14:**



(scatterplot\_data\_2.png)

**R2.15:**



(plot.png)

Aufgrund des Offset ist das Balkendiagramm zwar nicht präzise, zeigt aber gut die „Daten-Glätte“

**R2.16:**

Die Exportquote über die Jahre verläuft sehr glatt; Spannweite und Streuung sind gering.

Zudem scheinen die Werte recht symmetrisch um den Median verteilt zu sein, da der Median gleich dem arithmetischen Mittel ist.

Es liegt eine leicht negative Korrelation vor.

Insgesamt sind die Aussagen über die Daten mittelmäßig aussagekräftig, da nur wenige Datenpunkte vorliegen und 2 Datenpunkte fehlerhaft sind.

**R2.17:**

Quartile Jahr:

Q0.25: 2016

Q0.5: 2018

Q0.75: 2020

Dezile Jahr:

Q0.1: 2015

Q0.2: 2016

Q0.3: 2017

Q0.4: 2017

Q0.5: 2018

Q0.6: 2019

Q0.7: 2019

Q0.8: 2020

Q0.9: 2021

Quartile Exportquote:

Q0.25: 46.85

Q0.5: 47.0

Q0.75: 47.2

Dezile Exportquote:

Q0.1: 46.7

Q0.2: 46.85

Q0.3: 46.9

Q0.4: 46.9

Q0.5: 47.0

Q0.6: 47.05

Q0.7: 47.05

Q0.8: 47.2

Q0.9: 47.3

**R2.18:**

Quartilsabstand Jahr: 4.0

Quartilsabstand Exportquote: 0.35

**R2.19:**

Kovarianz: -0.1

**R.2.20:**

Korrelationskoeffizient: -0.26

**R3.1:**

Daten des Statistischen Bundesamts, veröffentlicht am 12. September 2022

2 Spalten; Jahr (2015 bis 2021) und zugehörige Exportquote in Prozent

7 Datenpaare

Es sind fehlerhafte/ unvollständige Daten vorhanden

Die Daten liegen im csv-Format vor und befinden sich in zwei Dateien.

Dabei liegen die Jahreszahlen in Datei a in codierter Form vor. Die Codierung kann Datei b entnommen werden.

**R3.2:**

data3.csv

**R3.3:**

data-3-bereinigt.csv

**R3.4:**

Die Daten wurden manuell in eine für die Auswertung mit dem Python Programm notwendige Form gebracht.

Bei Werten mit falschem Datentyp findet im Python Programm lineare Interpolation statt.

Ist der erste oder letzte Wert fehlerhaft, wird der Wert danach bzw. davor übernommen.

Die lineare Interpolation wurde als sinnvoll erachtet, da die sonstigen Daten recht glatt verlaufen.

**R3.5:**

data-3-a-bereinigt.xlsx

**R3.6:**

Software und genutzte Funktionen:

Programmierung per PyCharm mit Python.

Bibliotheken:

csv; Einlesen der csv-Dateien

math; Verschiedene mathematische Operatoren

os; Überprüfung ob angegebene Datei existiert



matplotlib; Erstellen von Boxplot, Scatterplot, Curvefitting...

numpy; Allgemeine Berechnungen

**R3.7:**

data-3-urliste-jahr.csv

data-3-urliste-exportquote.csv

**R3.8:**

rangliste\_Exportquote.csv

rangliste\_Key.csv

rangliste\_Jahr.csv

**R3.9:**

Modus Exportquote (nach linearer Interpolation): 47.2

Modus Jahr: -

Arithmetische Mittelwert Jahr: 2018.0

Arithmetische Mittelwert Exportquote: 47.15

Median Jahr: 2018.0

Median Exportquote: 47.2

**R3.10:**

Spannweite Jahr: 6.0

Spannweite Exportquote: 1.0

**R3.11:**

Mittlere Abweichung von Median Jahr: 1.71

Mittlere Abweichung von Median Exportquote: 0.22

**R3.12:**

Stichprobenvarianz Jahr: 4.0

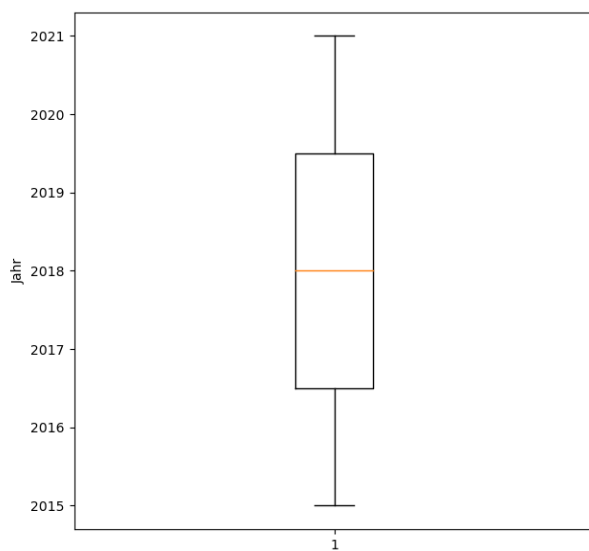
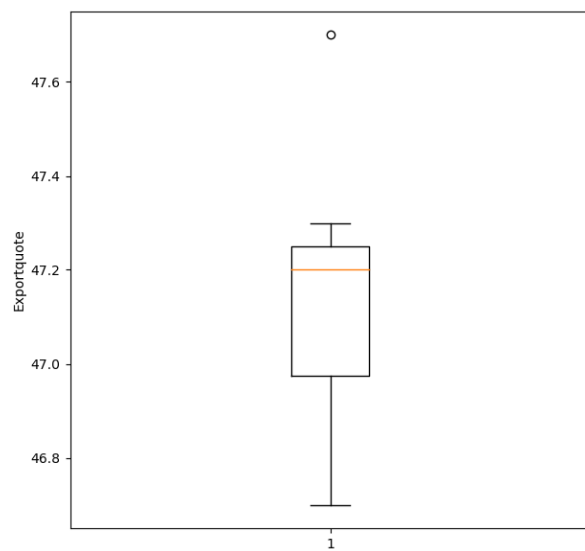
Stichprobenvarianz Exportquote: 0.09

**R3.13:**

Variationskoeffizient Jahr: 0.0

Variationskoeffizient Exportquote: 0.01

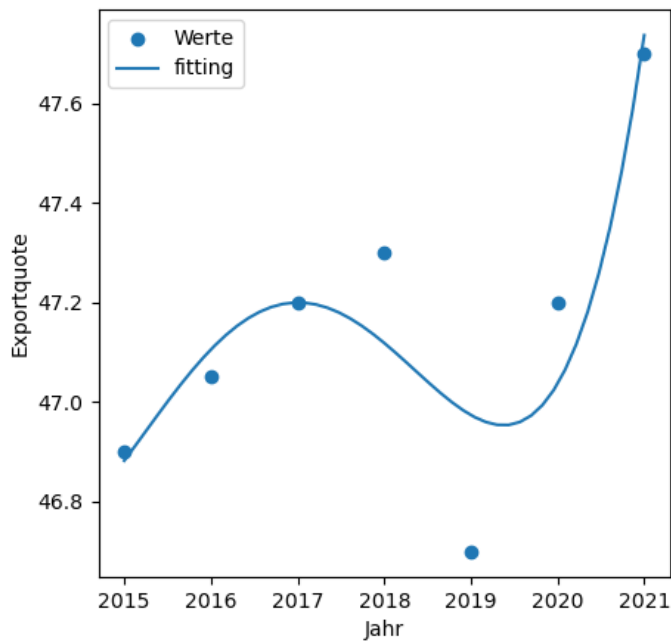
**R3.14:**



(boxplot\_Exportquote.png

boxplot\_Jahr.png)

**R3.15, R3.16:**



(scatterplot-data-3.png)

**R3.17:**

Eine Polynomfunktion 5. Grades hat sich als Kompromiss aus Fehler und Anschaulichkeit der Regression ergeben:

$$y = 0.01 * x^5 + -0.13 * x^4 + 0.84 * x^3 + -2.4 * x^2 + 2.78 * x + 45.9$$

**R3.18:**

Legende siehe Scatterplot

**R3.19:**

Die Exportquote über die Jahre verläuft sehr glatt; Spannweite und Streuung sind gering.

Zudem scheinen die Werte recht symmetrisch um den Median verteilt zu sein, da der Median und arithmetisches Mittel fast gleich sind.

Es liegt eine mittelmäßige positive Korrelation vor.

Insgesamt sind die Aussagen über die Daten mittelmäßig aussagekräftig, da nur wenige Datenpunkte vorliegen und 2 Datenpunkte fehlerhaft sind.

**R3.20:**

Quartile Jahr:

Q0.25: 2016

Q0.5: 2018

Q0.75: 2020

Dezile Jahr:

Q0.1: 2015

Q0.2: 2016

Q0.3: 2017

Q0.4: 2017

Q0.5: 2018

Q0.6: 2019

Q0.7: 2019

Q0.8: 2020

Q0.9: 2021

Quartile Exportquote:

Q0.25: 46.9

Q0.5: 47.2

Q0.75: 47.3

Dezile Exportquote:

Q0.1: 46.7

Q0.2: 46.9

Q0.3: 47.05

Q0.4: 47.05

Q0.5: 47.2

Q0.6: 47.2

Q0.7: 47.2

Q0.8: 47.3

Q0.9: 47.7

**R3.21:**

Quartilsabstand Jahr: 4

Quartilsabstand Exportquote: 0.4

**R3.22:**

Kovarianz: 0.37

**R3.23:**

Korrelationskoeffizient: 0.63

**R4.1:**

data-4-1D.csv

**R4.2:**

Data-4-bereinigt.csv

(Die Daten sind hier wieder 2D, damit sie für das universelle Python-Programm geeignet sind)

**R4.3:**

Dimensionsanpassung der ursprünglich 3-D-Daten mittels LibreOffice Calc

Die Daten wurden nicht weiter angepasst. Zwar verzerren die geringen Messwerte bei Phasen in Ruhe die Datenauswertung, andererseits ist es schwierig, eine Messwertgrenze für "In Ruhe" zu definieren.

**R4.4:**

Software und genutzte Funktionen:

Programmierung per PyCharm mit Python.

Bibliotheken:

csv; Einlesen der csv-Dateien

math; Verschiedene mathematische Operatoren

os; Überprüfung ob angegebene Datei existiert

matplotlib; Erstellen von Boxplot

scipy; Berechnung des Modus

**R4.5:**

Modus (mit 1 Nachkommastelle Genauigkeit): 0.3 (mit  $n = 13$ )

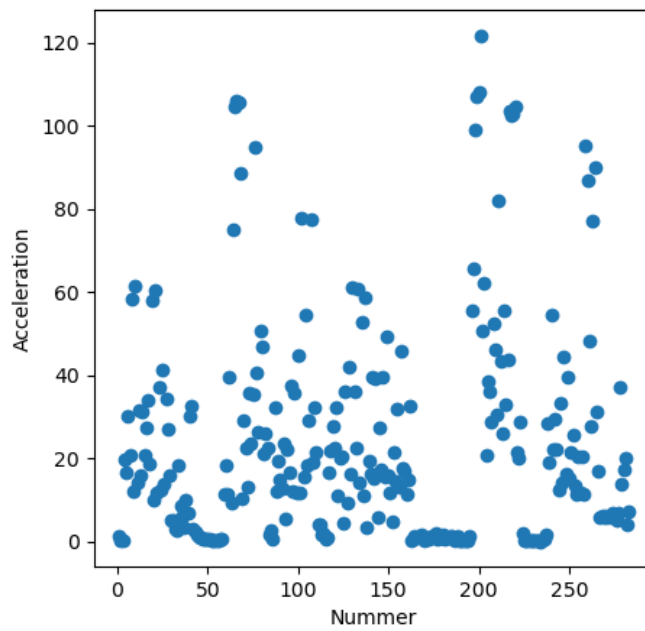
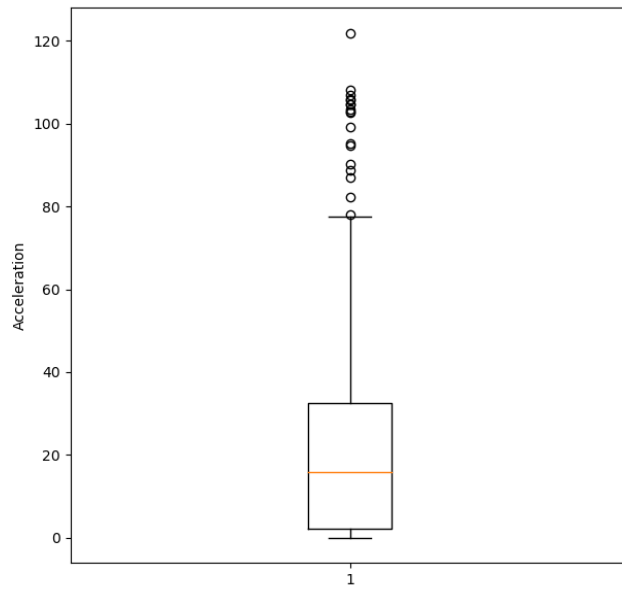
Arithmetischer Mittelwert: 23.42

Median: 15.83

**R4.6:**

Stichprobenvarianz: 691.18

**R4.7:**



(boxplot\_Acceleration.png)

(scatterplot\_data\_4.png)

**R4.8:**

Bei den Daten handelt es sich um 1D Zentripetalbeschleunigungs-Daten des iPhone-Beschleunigungssensors, aufgezeichnet bzw. generiert mit der Phyphox-App.

Es wurden 283 Daten aufgezeichnet.

Die Daten sind in  $\text{m/s}^2$  angegeben.

Die Beschleunigungsdaten wurden durch die Drehung eines Arms aufgenommen.

Dementsprechend lässt sich die hohe Stichprobenvarianz erklären (der Arm war zwischendurch in Ruhe, und die Winkelgeschwindigkeit bei der Drehung überhaupt nicht konstant)

Der Mittelwert ist wenig aussagekräftig, da die Zeiten des Handys in Ruhe mit einbezogen werden.

Hier ist die Spannweite der Daten von  $121.81 \text{ m/s}^2$  interessanter, da dies aufgrund des Nullpunkts in Ruhe die maximale momentane Beschleunigung darstellt.

$121.81 \text{ m/s}^2$  entspricht etwa  $12,4g$