Gruppe 19

Paul Hendrik van Norden

Grundlagen und Anwendung der Wahrscheinlichkeitstheorie (GAWT)

2022 WS

Contents

[Hinweise 4](#_Toc127892794)

[Datensatz 1 5](#_Toc127892795)

[R1.1 Beschreibung des Datensatzes 5](#_Toc127892796)

[R1.2 Skalenvariante 5](#_Toc127892797)

[R1.3 Verwendete Software 5](#_Toc127892798)

[R1.4 Urliste 5](#_Toc127892799)

[R1.5 Rangliste 5](#_Toc127892800)

[R1.6 Excel-Datei 5](#_Toc127892801)

[R1.7 Modus, arithmetischer Mittelwert, Median 5](#_Toc127892802)

[R1.8 Spannweite 6](#_Toc127892803)

[R1.9 Mittlere Abweichung vom Median 6](#_Toc127892804)

[R1.10 Stichprobenvarianz 6](#_Toc127892805)

[R1.11 Variationskoeffizient 6](#_Toc127892806)

[R1.12 Box-Whisker-Plot 6](#_Toc127892807)

[R1.13 Scatterplot 7](#_Toc127892808)

[R1.14 Beschreibung der Daten 7](#_Toc127892809)

[R1.15 Quartile und Dezile 8](#_Toc127892810)

[R1.16 Quartilsabstand 8](#_Toc127892811)

[R1.17 Kovarianz 8](#_Toc127892812)

[R1.18 Korrelationskoeffizient 8](#_Toc127892813)

[R1.19 Einteilung Klassen und Histogramm 9](#_Toc127892814)

[R1.20 Kontingenztabelle 11](#_Toc127892815)

[R1.21 Rangkorrelationskoeffizient nach Spearman 11](#_Toc127892816)

[Datensatz 2 12](#_Toc127892817)

[R2.1 Beschreibung des Datensatzes 12](#_Toc127892818)

[R2.2 Daten bereinigt 12](#_Toc127892819)

[R2.3 Maßnahmen zur Datenbereinigung 12](#_Toc127892820)

[R2.4 Verwendete Software 12](#_Toc127892821)

[R2.5 Excel-Datei 12](#_Toc127892822)

[R2.6 Urliste 12](#_Toc127892823)

[R2.7 Rangliste 12](#_Toc127892824)

[R2.8 Modus, arithmetischer Mittelwert, Median 12](#_Toc127892825)

[R2.9 Spannweite 13](#_Toc127892826)

[R2.10 Mittlere Abweichung vom Median 13](#_Toc127892827)

[R2.11 Stichprobenvarianz 13](#_Toc127892828)

[R2.12 Variationskoeffizient 13](#_Toc127892829)

[R2.13 Box-Whisker-Plot 13](#_Toc127892830)

[R2.14 Scatterplot 14](#_Toc127892831)

[R2.15 Kreisdiagramm 15](#_Toc127892832)

[R2.16 Beschreibung der Daten 15](#_Toc127892833)

[R2.17 Quartile und Dezile 16](#_Toc127892834)

[R2.18 Quartilsabstand 16](#_Toc127892835)

[R2.18 Kovarianz 16](#_Toc127892836)

[R2.19 Korrelationskoeffizient 16](#_Toc127892837)

[Datensatz 3 17](#_Toc127892838)

[R3.1 Beschreibung des Datensatzes 17](#_Toc127892839)

[R3.2 Daten zusammengeführt 17](#_Toc127892840)

[R3.3 Daten bereinigt 17](#_Toc127892841)

[R3.4 Maßnahmen Datenbereinigung 17](#_Toc127892842)

[R3.5 Excel-Datei 17](#_Toc127892843)

[R3.6 Verwendete Software 17](#_Toc127892844)

[R3.7 Urliste 17](#_Toc127892845)

[R3.8 Rangliste 17](#_Toc127892846)

[R3.9 Modus, arithmetischer Mittelwert, Median 18](#_Toc127892847)

[R3.10 Spannweite 18](#_Toc127892848)

[R3.11 Mittlere Abweichung vom Median 18](#_Toc127892849)

[R3.12 Stichprobenvarianz 18](#_Toc127892850)

[R3.13 Variationskoeffizient 18](#_Toc127892851)

[R3.14 Box-Whisker-Plot 19](#_Toc127892852)

[R3.15, R3.16, R3.18 Scatterplot 20](#_Toc127892853)

[R3.17 Funktion Curvefitting 20](#_Toc127892854)

[R3.19 Beschreibung der Daten 21](#_Toc127892855)

[R3.20 Quartile und Dezile 21](#_Toc127892856)

[R3.21 Quartilsabstand 21](#_Toc127892857)

[R2.22 Kovarianz 21](#_Toc127892858)

[R2.23 Korrelationskoeffizient 21](#_Toc127892859)

[Datensatz 4 22](#_Toc127892860)

[R4.1 Selbst zusammengestellter Datensatz 22](#_Toc127892861)

[R4.2 Daten bereinigt 22](#_Toc127892862)

[R4.3 Maßnahmen zur Datenbereinigung 22](#_Toc127892863)

[R4.4 Verwendete Software 22](#_Toc127892864)

[R4.5 Modus, arithmetischer Mittelwert, Median 23](#_Toc127892865)

[R4.6 Stichprobenvarianz 23](#_Toc127892866)

[R4.7 Boxplot 23](#_Toc127892867)

[R4.8 Beschreibung der Daten 24](#_Toc127892868)

# Hinweise

1)

In den ersten 3 Datensätzen liegt jeweils ein Eintrag für den Stichtag und ein Eintrag für den Bevölkerungsstand vor. Bei der Berechnung von Lageparametern, Streuungsparametern wird nur die Variable für den Bevölkerungsstand einbezogen, da die Berechnung jener Parameter für die Stichtage sinnlos und/oder rechnerisch nicht durchführbar ist. Dies gilt auch für die Erstellung von bestimmten Graphen (Boxplot, Histogramm, …). Desweitern ist die Berechnung der Kovarianz und des Korrelationskoeffizienten nach Pearson nicht möglich, da die beiden Skalentypen nicht vereinbar sind. Wenn also nur ein Wert angegeben ist, bezieht dieser sich automatisch auf die Variable, die den Bevölkerungsstand wiedergibt. Ist dies nicht der Fall wird dies explizit angegeben.

2)

Alle berechneten Werte werden auf maximal drei Nachkommastellen gerundet.

3)

Jegliche Csv-Dateien, sowie die Rohdaten liegen in entsprechenden beigefügten Ordnern vor oder sind über folgendes GitHub-Repository auffindbar:

# Datensatz 1

### R1.1 Beschreibung des Datensatzes

Der Datensatz gibt die Fortschreibung des Bevölkerungsstandes in Deutschland von 1950 bis 2021 an. Dabei werden bis 1989 nur die Bevölkerung der Westdeutschen Länder einbezogen. Ab 2011 sind die Ergebnisse auf Grundlage des Zensus 2011. Die Daten stammen vom statischen Bundesamt (Destatis) und wurden auf dem Datenportal „Genesis“ (Link: <https://www-genesis.destatis.de>) hochgeladen. Die Genesis Tabellennummer lautet: 12411-0001. Die Daten sind vom Stand 10.10.2022 / 11:01:39. Das verwendete Datenset lag in einer csv-Datei vor.

### R1.2 Skalenvariante

Den Stichtagen liegt eine Intervallskala und der Stichprobe für den Bevölkerungsstand eine Verhältnisskala zugrunde.

### R1.3 Verwendete Software

Zur Analyse der Daten wurde die webbasierte interaktive Entwicklungsumgebung Jupyter-Lab benutzt. Die dazugehörige verwendete Programmiersprache zur Analyse und Manipulation der Daten war Python in Kombination mit den Bibliotheken Pandas, Numpy und Matplotlib. Zur Verwaltung und Installation der Tools wurde die Plattform Anaconda genutzt.

### 

### R1.4 Urliste

* Datei: csv\_data1\_urliste.csv

### 

### R1.5 Rangliste

* Datei: csv\_data1\_rangliste.csv

### 

### R1.6 Excel-Datei

* Datei: excel\_data\_1.xlsx

### 

### R1.7 Modus, arithmetischer Mittelwert, Median

Angabe von Modus, arithmetischem Mittelwert und Median:

Modus: Jeder Wert in der Stichprobe kommt genau einmal vor

Arithmetischer Mittelwert: 69025182.139

Median: 61762240.5

### R1.8 Spannweite

Spannweite: 32278999

### R1.9 Mittlere Abweichung vom Median

Mittleren Abweichung vom Median: 10703382.222

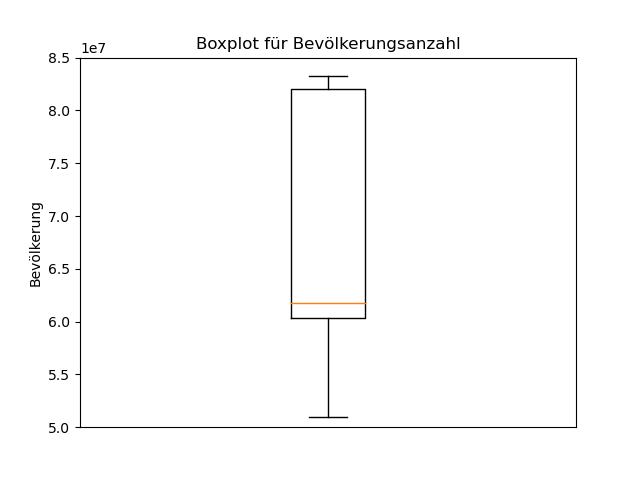
### R1.10 Stichprobenvarianz

Stichprobenvarianz: 142802788477530.6

### R1.11 Variationskoeffizient

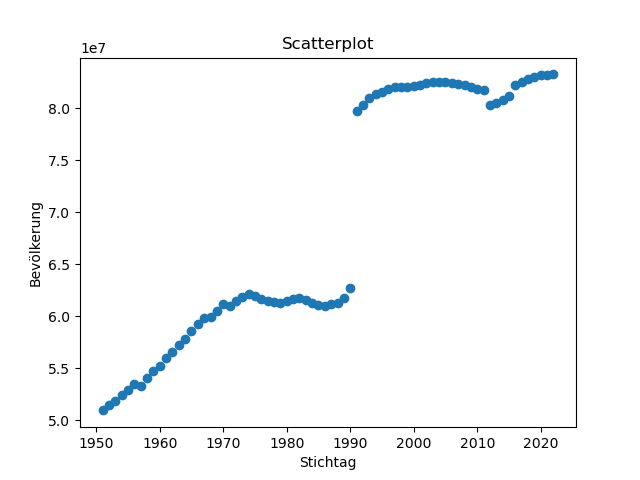
Variationskoeffizient: 2068850.585

### R1.12 Box-Whisker-Plot



### R1.13 Scatterplot

Scatterplot für die beiden Variablen:



### R1.14 Beschreibung der Daten

Der Datensatz zeigt eine allgemeine Tendenz zur Steigerung der Bevölkerungszahl im Zeitraum von 1950 bis 2021. Zu Beginn liegt die Bevölkerungszahl bei ungefähr 51 Millionen. Bis Mitte der 1960er Jahre steigt die Bevölkerungszahl jährlich um etwa 5 Millionen. Ab den 1970er Jahren nimmt die jährliche Steigerungsrate ab, bleibt aber weiterhin positiv. Mitte der 1970er Jahre bleibt die Bevölkerungszahl weitestgehend konstant, wobei ein sehr leichter Abfall erkennbar ist. Um 1990 ist wieder ein deutlicher Anstieg der Bevölkerungszahl erkennbar und kurz nach 1990 findet ein enormer Sprung der Bevölkerungszahl von ungefähr 63 Millionen auf 78 Millionen Menschen. Danach steigt der die Bevölkerungszahl weiter leicht an. Es ist jedoch zu beachten, dass es von 2008 bis 2011 kurzzeitige Rückgänge in der Bevölkerungszahl gibt, die jedoch von den übrigen Jahren übertroffen werden. Im letzten Jahr erreicht der Bevölkerungsstand sein Maximum mit ungefähr 83 Millionen Menschen. Die durchschnittliche Bevölkerungsanzahl liegt insgesamt bei ca. 69 Millionen.

### R1.15 Quartile und Dezile

Angabe der Quartile und Dezile:

Quartile:

1. Quartil: 60334393.25
2. Quartil: 61762240.50
3. Quartil: 82018374.25

Dezile:

1. Dezil: 54129844.4
2. Dezil: 58729279.0
3. Dezil: 61076617.5
4. Dezil: 61440403.6
5. Dezil: 61762240.5
6. Dezil: 80445407.6
7. Dezil: 81787060.5
8. Dezil: 82173242.2
9. Dezil: 82519572.6

### R1.16 Quartilsabstand

Quartilsabstand R\_Q0.5: 21683981

### R1.17 Kovarianz

Die Kovarianz ist mathematisch nicht berechenbar.

### R1.18 Korrelationskoeffizient

Der Korrelationskoeffizient ist mathematisch nicht berechenbar.

### R1.19 Einteilung Klassen und Histogramm

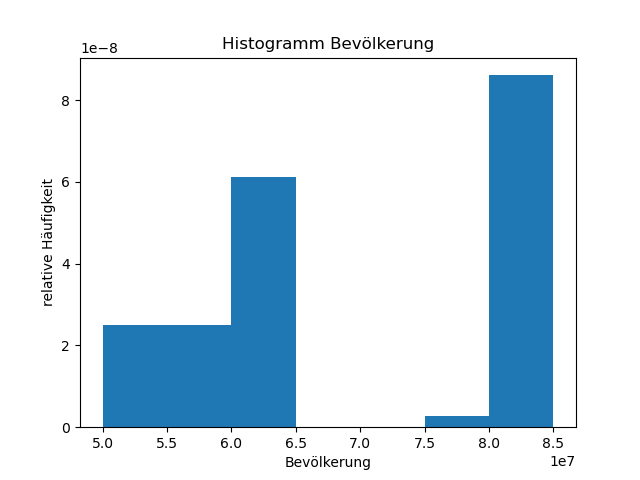
Einteilung der Stichprobe für die Bevölkerung in folgende Klassen:

|  |  |  |
| --- | --- | --- |
| Klassen - Nummer | Klassen - Bezeichnung | Definition (Werte -> von – bis) |
| 1 | 5 – 5.5 | [5e7 – 5.5e7] |
| 2 | 5.5 – 6 | ]5.5e7 – 6e7] |
| 3 | 6 - 6.5 | ]6e7 – 6.5e7] |
| 4 | 7 - 7.5 | ]6.5e7 – 7e7] |
| 5 | 6.5 – 7 | ]7e7 – 7.5e7] |
| 6 | 7.5 – 8 | ]7.5e7 – 8e7] |
| 7 | 8 - 8.5 | ]8e7 – 8.5e7] |

Einteilung der Stichtage in folgende Klassen:

|  |  |  |
| --- | --- | --- |
| Klassen - Nummer | Klassen - Bezeichnung | Definition (Werte -> von – bis) |
| 1 | 1950 - 1960 | [1950/12/31 – 1960/12/31] |
| 2 | 1960 - 1970 | ]1960/12/31 – 1970/12/31] |
| 3 | 1970 - 1980 | ]1970/12/31 – 1980/12/31] |
| 4 | 1980 - 1990 | ]1980/12/31 – 1990/12/31] |
| 5 | 1990 -2000 | ]1990/12/31 – 2000/12/31] |
| 6 | 2000 - 2010 | ]2000/12/31 – 2010/12/31] |
| 7 | 2010 - 2021 | ]2010/12/31 – 2021/12/31] |

Histogramm für Verteilung der Bevölkerungsanzahl:



### R1.20 Kontingenztabelle

Table

Description automatically generated

### R1.21 Rangkorrelationskoeffizient nach Spearman

Rangkorrelationskoeffizient nach Spearman: 0.931

# Datensatz 2

### R2.1 Beschreibung des Datensatzes

Der Datensatz gibt die Fortschreibung des Bevölkerungsstandes in Deutschland von 1950 bis 2021 an. Dabei werden bis 1989 nur die Bevölkerung der Westdeutschen Länder einbezogen. Ab 2011 sind die Ergebnisse auf Grundlage des Zensus 2011. Die Daten stammen vom statischen Bundesamt (Destatis) und wurden auf dem Datenportal „Genesis“ (Link: <https://www-genesis.destatis.de>) hochgeladen. Die Genesis Tabellennummer lautet: 12411-0001. Die Daten sind vom Stand 10.10.2022 / 11:01:39. Das verwendete Datenset lag in einer csv-Datei vor.

### R2.2 Daten bereinigt

* Datei: csv\_data2\_urliste\_bereinigt.csv

### R2.3 Maßnahmen zur Datenbereinigung

Zur Bereinigung der Daten wurden folgende Schritte vorgenommen:

* Korrektur falscher Datumsangaben
* Entfernung von leeren Zeilen und Zeilen mit NaN-Values
* Entfernung von Zeilen mit sinnlosen Einträgen (Angabe eines Namens anstelle der Bevölkerungsanzahl)

### R2.4 Verwendete Software

* Selbe Programme verwendet wie in R1.3

### R2.5 Excel-Datei

* Datei: excel\_data\_2.xlsx

### R2.6 Urliste

* Datei: csv\_data2\_urliste\_unbereinigt.csv

### R2.7 Rangliste

* Datei: csv\_data2\_rangliste.csv

### R2.8 Modus, arithmetischer Mittelwert, Median

Angabe von Modus, arithmetischem Mittelwert und Median:

Modus: Jeder Wert in der Stichprobe kommt genau einmal vor

Arithmetischer Mittelwert: 69078784.017

Median: 61713896.0

### R2.9 Spannweite

Spannweite: 32278999

### R2.10 Mittlere Abweichung vom Median

Mittleren Abweichung vom Median: 10380550.297

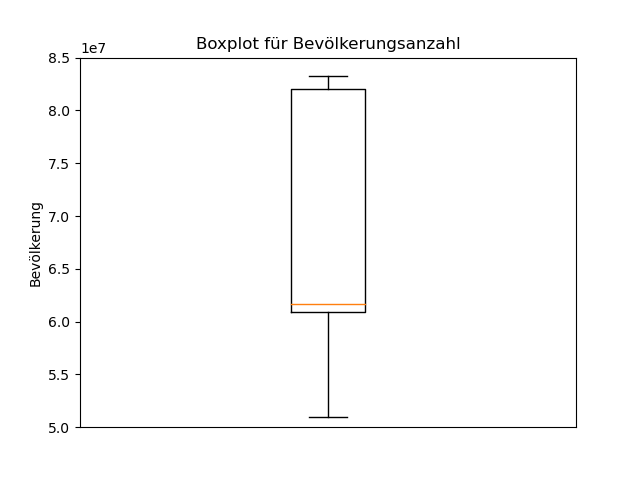
### R2.11 Stichprobenvarianz

Stichprobenvarianz: 137735543999416.7

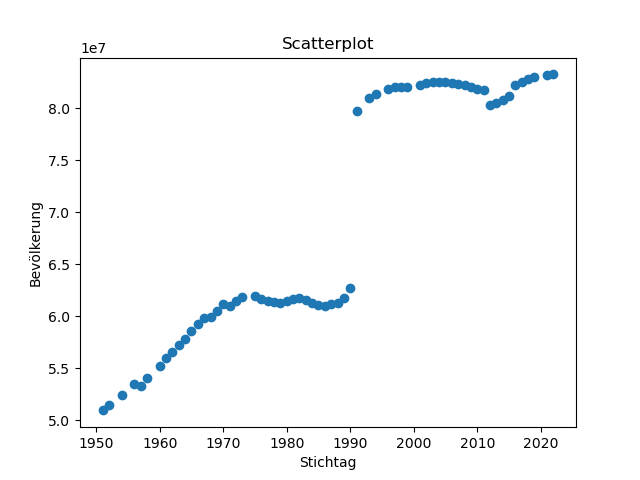
### R2.12 Variationskoeffizient

Variationskoeffizient: 1993890.685

### R2.13 Box-Whisker-Plot

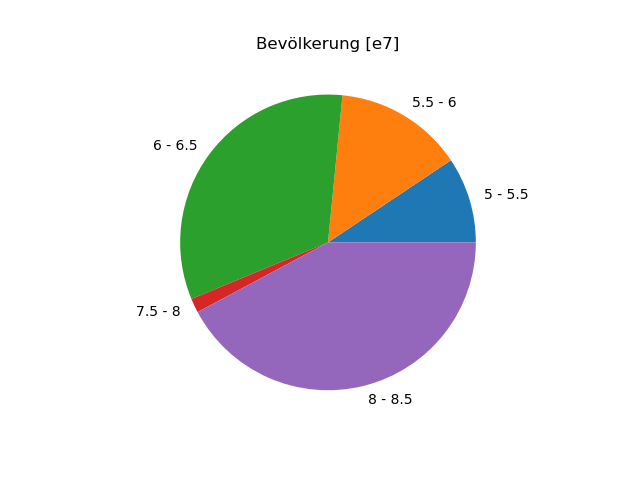


### R2.14 Scatterplot



### R2.15 Kreisdiagramm

Kreisdiagramm für die Bevölkerung:



Das Kreisdiagramm zeigt die relative Häufigkeit der in Klassen eingeteilten Bevölkerungsanzahl an.

### R2.16 Beschreibung der Daten

Da dieser Datensatz fast identisch mit Datensatz 1 ist, wird an dieser Stelle wieder auf die Beschreibung in R1.14 verwiesen.

### R2.17 Quartile und Dezile

Angabe der Quartile und Dezile:

Quartile:

1. Quartil: 60866631.25
2. Quartil: 61713896.0
3. Quartil: 82018374.25

Dezile:

1. Dezil: 55467457.9
2. Dezil: 59594396.8
3. Dezil: 61131340.5
4. Dezil: 61439872.8
5. Dezil: 61713896.0
6. Dezil: 80484576.8
7. Dezil: 81803781.2
8. Dezil: 82192545.2
9. Dezil: 82515411.8

### R2.18 Quartilsabstand

Quartilsabstand R\_Q0.5: 21151743.0

### R2.18 Kovarianz

Die Kovarianz ist mathematisch nicht berechenbar.

### R2.19 Korrelationskoeffizient

Der Korrelationskoeffizient ist mathematisch nicht berechenbar.

# Datensatz 3

### R3.1 Beschreibung des Datensatzes

Der Datensatz gibt die Fortschreibung des Bevölkerungsstandes in Deutschland von 1950 bis 2021 an. Dabei werden bis 1989 nur die Bevölkerung der Westdeutschen Länder einbezogen. Ab 2011 sind die Ergebnisse auf Grundlage des Zensus 2011. Die Daten stammen vom statischen Bundesamt (Destatis) und wurden auf dem Datenportal „Genesis“ (Link: <https://www-genesis.destatis.de>) hochgeladen. Die Genesis Tabellennummer lautet: 12411-0001. Die Daten sind vom Stand 10.10.2022 / 11:01:39. Es lagen insgesamt 2 csv-Dateien vor.

### R3.2 Daten zusammengeführt

* Datei: csv\_data3\_zusammengeführt.csv

### R3.3 Daten bereinigt

* Datei: csv\_data3\_urliste\_bereinigt.csv

### R3.4 Maßnahmen Datenbereinigung

Zur Bereinigung der Daten wurden folgende Schritte vorgenommen:

* Korrektur falscher Datumsangaben
* Entfernung von leeren Zeilen und Zeilen mit NaN-Values
* Entfernung von Zeilen mit sinnlosen Einträgen (Angabe eines Namens anstelle der Bevölkerungsanzahl)

### R3.5 Excel-Datei

* Datei: excel\_data\_3.xlsx

### R3.6 Verwendete Software

* Selbe Programme verwendet wie in R1.3

### R3.7 Urliste

* Datei: csv\_data3\_urliste\_unbereinigt.csv

### R3.8 Rangliste

* Datei: csv\_data3\_rangliste.csv

### R3.9 Modus, arithmetischer Mittelwert, Median

Angabe von Modus, arithmetischem Mittelwert und Median:

Modus: Jeder Wert in der Stichprobe kommt genau einmal vor

Arithmetischer Mittelwert: 69078784.016

Median: 61713896.0

### R3.10 Spannweite

Spannweite: 32278999

### R3.11 Mittlere Abweichung vom Median

Mittleren Abweichung vom Median: 10380550.297

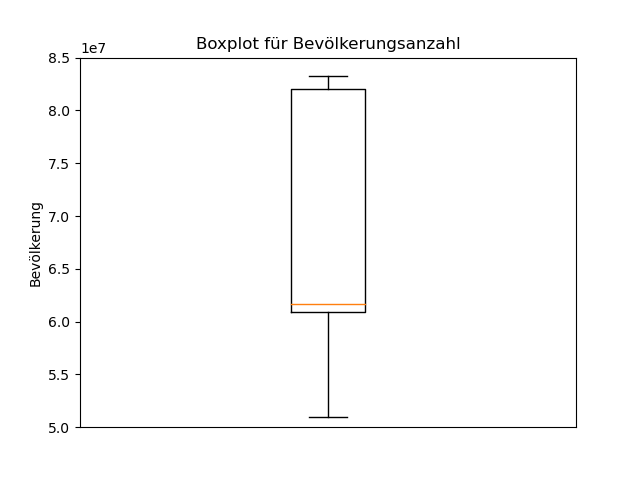
### R3.12 Stichprobenvarianz

Stichprobenvarianz: 137735543999416.7

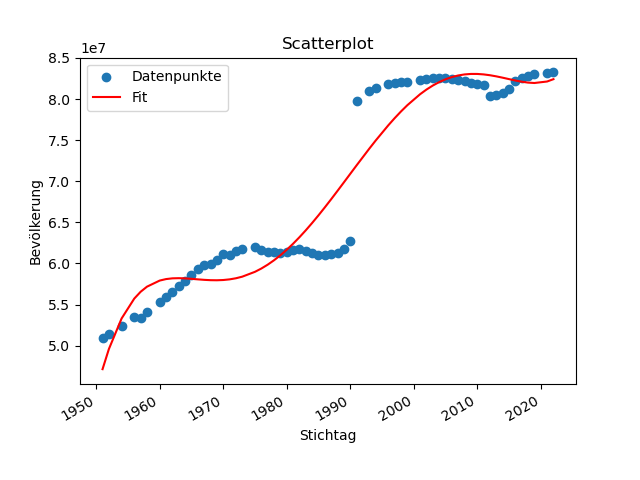
### R3.13 Variationskoeffizient

Variationskoeffizient: 1993890.685

### R3.14 Box-Whisker-Plot



### R3.15, R3.16, R3.18 Scatterplot



### R3.17 Funktion Curvefitting

Da es sich bei den Werten auf der x-Achse nicht um eine echte metrische Skala handelt, ist die direkte Berechnung eines Polynoms für ein Curvefitting nicht möglich. Die Daten müssen erst in numerische Werte umgewandelt werden. Wenn das Polynom berechnet wurde, werden die numerischen Werte für das Datum wieder zum jeweiligen Datum umgewandelt. Der Vorgang wird vom benutzten Programm erledigt.

Formel für Curvefitting-Polynom:

9.77e-14 x5 – 3.579e-09 x4 + 2.626e-05 x3 + 0.1832 x2 + 145.2 x + 5.799e+07

Man muss bei der Formel jedoch beachten, dass sich die x-Werte auf die zum Datum äquivalenten numerischen Werte beziehen.

### R3.19 Beschreibung der Daten

Da dieser Datensatz fast identisch mit Datensatz 1 ist, wird an dieser Stelle wieder auf die Beschreibung in R1.14 verwiesen.

### R3.20 Quartile und Dezile

Angabe der Quartile und Dezile:

Quartile:

1. Quartil: 60866631.25
2. Quartil: 61713896.0
3. Quartil: 82018374.25

Dezile:

1. Dezil: 55467457.9
2. Dezil: 59594396.8
3. Dezil: 61131340.5
4. Dezil: 61439872.8
5. Dezil: 61713896.0
6. Dezil: 80484576.8
7. Dezil: 81803781.2
8. Dezil: 82192545.2
9. Dezil: 82515411.8

### R3.21 Quartilsabstand

Quartilsabstand R\_Q0.5: 21151743.0

### R2.22 Kovarianz

Die Kovarianz ist mathematisch nicht berechenbar.

### R2.23 Korrelationskoeffizient

Der Korrelationskoeffizient ist mathematisch nicht berechenbar.

# Datensatz 4

### R4.1 Selbst zusammengestellter Datensatz

Informationen zum vorliegenden Datenset:

Die Werte sind auf die Anzahl der Eliminierungen bezogen, die ich jeweils in den letzten 100 Runden in dem Online-Videospiel Valorant gemacht habe. Valorant ist ein taktischer First-Person-Shooter, der von Riot Games entwickelt wurde. In dem Spiel treten zwei Teams mit jeweils fünf Spielern gegeneinander an und versuchen, ein Ziel zu erreichen, indem sie entweder eine Bombe legen oder entschärfen. Es werden nur Runden des Spielmodus „Competitive“ einbezogen.

Die Daten können folgender Seite entnommen werden:

[Ubberx#EUW's Valorant Competitive Match History - Valorant Tracker](https://tracker.gg/valorant/profile/riot/Ubberx%23EUW/matches?playlist=competitive)

Dez 21, 2022 – Feb 16, 2023

Csv Datei -> csv\_data4\_urliste.csv

### R4.2 Daten bereinigt

* Datei: csv\_data4\_urliste\_bereinigt.csv

### R4.3 Maßnahmen zur Datenbereinigung

Zur Bereinigung der Daten wurden folgende Schritte vorgenommen:

* Überprüfung ob Zeilen ohne Eintrag oder sinnlosem Eintrag vorliegen und Entfernung dieser.
* Werte, die kleiner als 5 und größer als 50 sind, werden entfernt.
  + Werte die kleiner als 5 sind werden entfernt, da es sich hierbei höchstwahrscheinlich um sehr früh abgebrochene Spiele handelt (zum Beispiel aufgrund von früher Aufgabe einer der Teams), die nicht repräsentativ sind.
  + Werte die größer als 50 sind werden entfernt, da das Erzielen von 50 Eliminierungen in einer Runde sehr schwierig ist und es sich somit um einen falsch eingetragenen Wert handeln muss.

### R4.4 Verwendete Software

* Selbe Programme verwendet wie in R1.3.

### R4.5 Modus, arithmetischer Mittelwert, Median

Angabe von Modus, arithmetischem Mittelwert und Median:

Modus: 15, 20 (2 Werte)

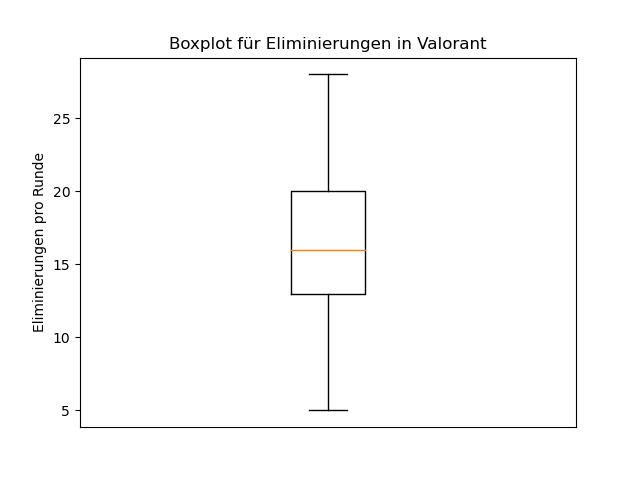
Arithmetischer Mittelwert: 16.224

Median: 16

### R4.6 Stichprobenvarianz

Stichprobenvarianz: 25.186

### R4.7 Boxplot



### R4.8 Beschreibung der Daten

Den Daten kann man entnehmen, dass ich im Durschnitt ungefähr 16 Eliminierungen erzielen. Meine niedrigste Eliminierungsanzahl in einer Runde liegt bei 6 und meine höchste bei 20. Die Vielzahl meiner Kills häufen sich im Bereich von 13 bis 20.