Gruppe 19

Paul Hendrik van Norden

Grundlagen und Anwendung der Wahrscheinlichkeitstheorie (GAWT)

2022 WS

Inhaltsübersicht

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, Excel-Dateien, sowie die Rohdaten liegen in beigefügten Ordnern im GitHub-Repository vor.

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

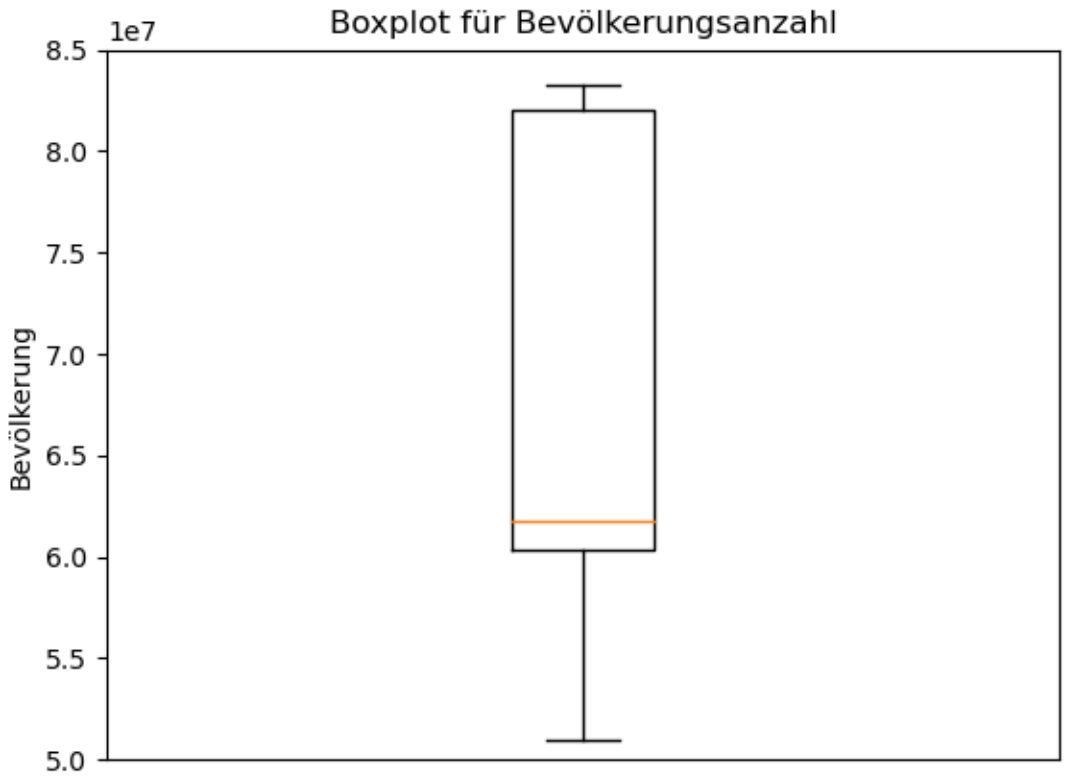
R1.10 Stichprobenvarianz

Stichprobenvarianz: 142802788477530.6

R1.11 Variationskoeffizient

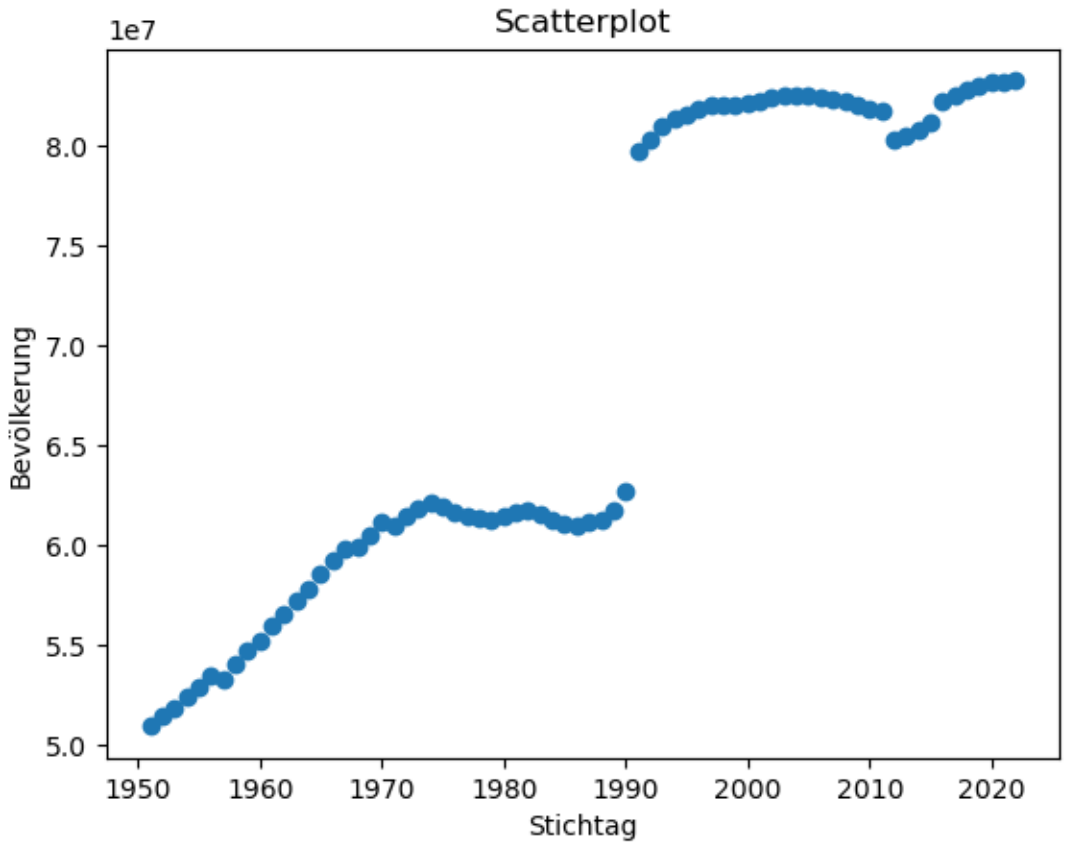
Variationskoeffizient: 2068850.585

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

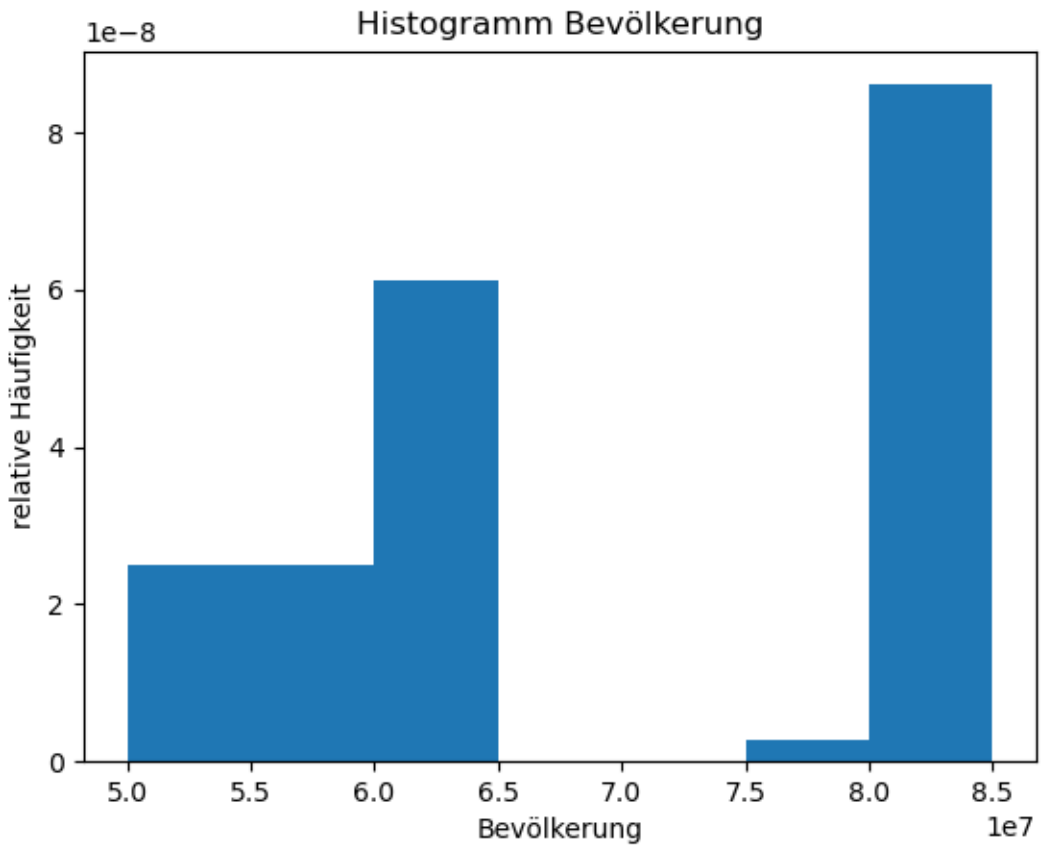
Einteilung der Stichprobe für die Bevölkerung in folgende Klassen:

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

date population	1950-1960	1960-1970	1970-1980	1980-1990	1990-2000	2000-2010	2010-2021	Total
5 - 5.5	9	0	0	0	0	0	0	9
5.5 - 6	2	7	0	0	0	0	0	9
6 - 6.5	0	3	10	9	0	0	0	22
6.5 - 7	0	0	0	0	0	0	0	0
7 - 7.5	0	0	0	0	0	0	0	0
7.5 - 8	0	0	0	1	0	0	0	1
8 - 8.5	0	0	0	0	10	10	11	31
Total	11	10	10	10	10	10	11	72

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

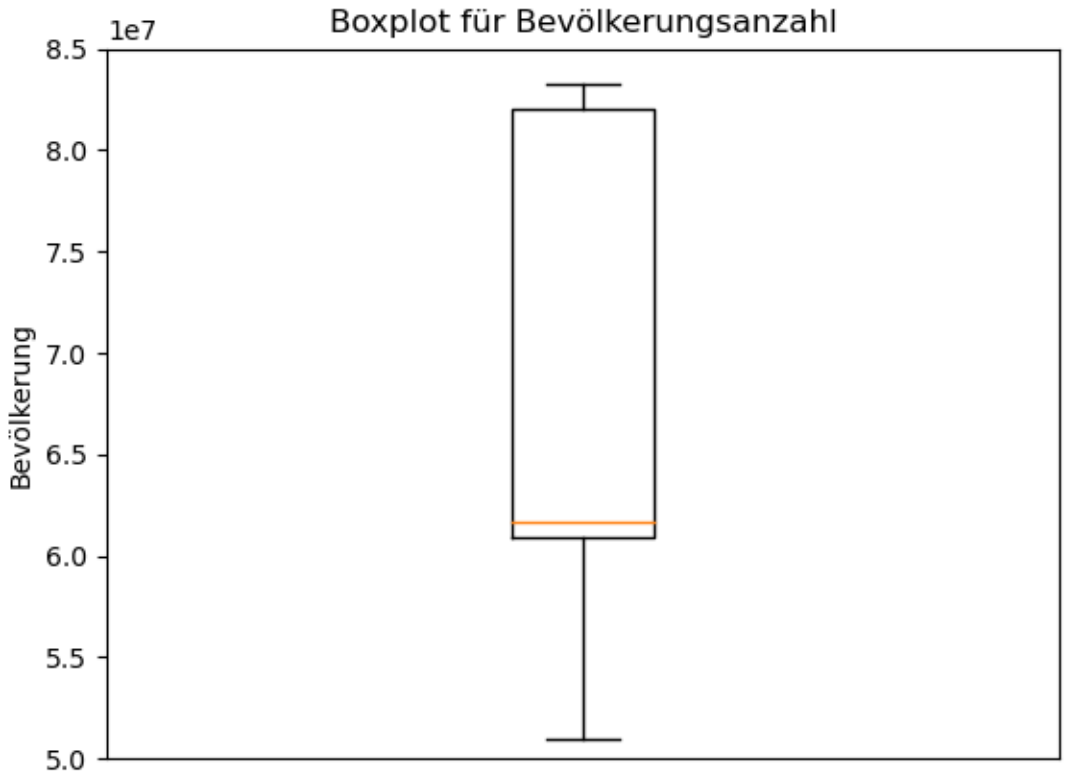
R2.11 Stichprobenvarianz

Stichprobenvarianz: 137735543999416.7

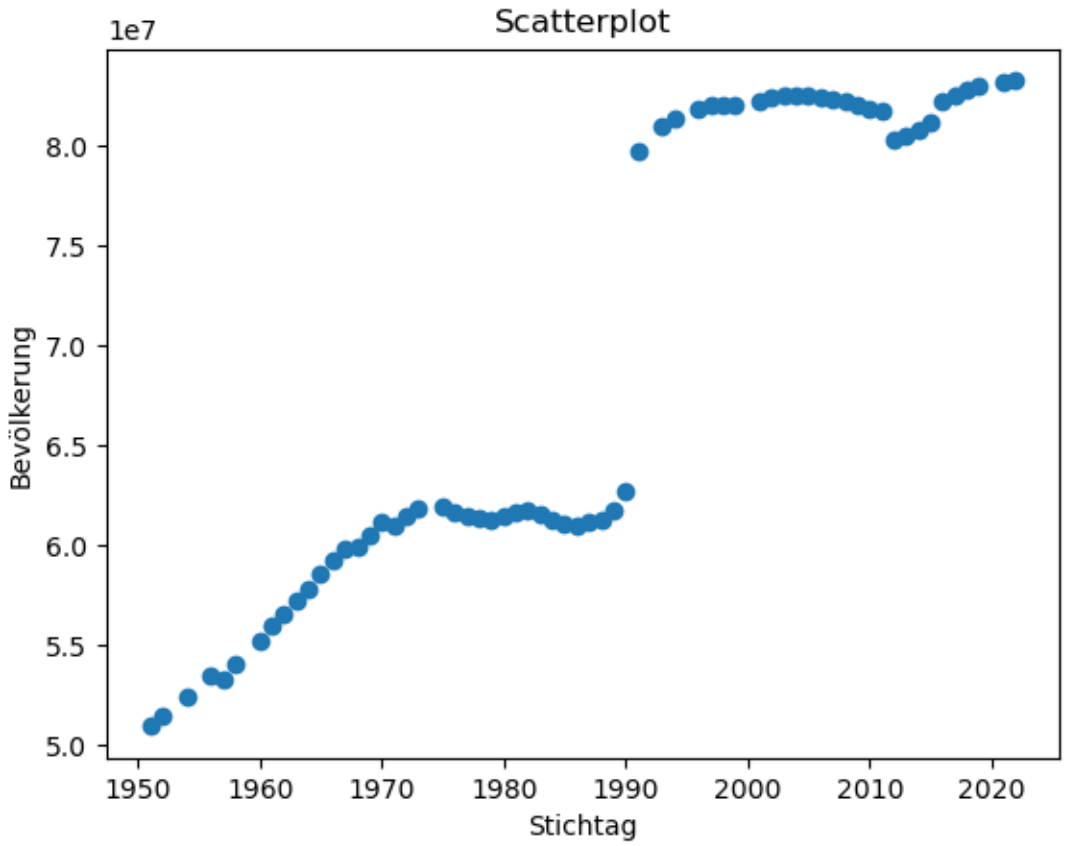
R2.12 Variationskoeffizient

Variationskoeffizient: 1993890.685

R2.13 Box-Whisker-Plot

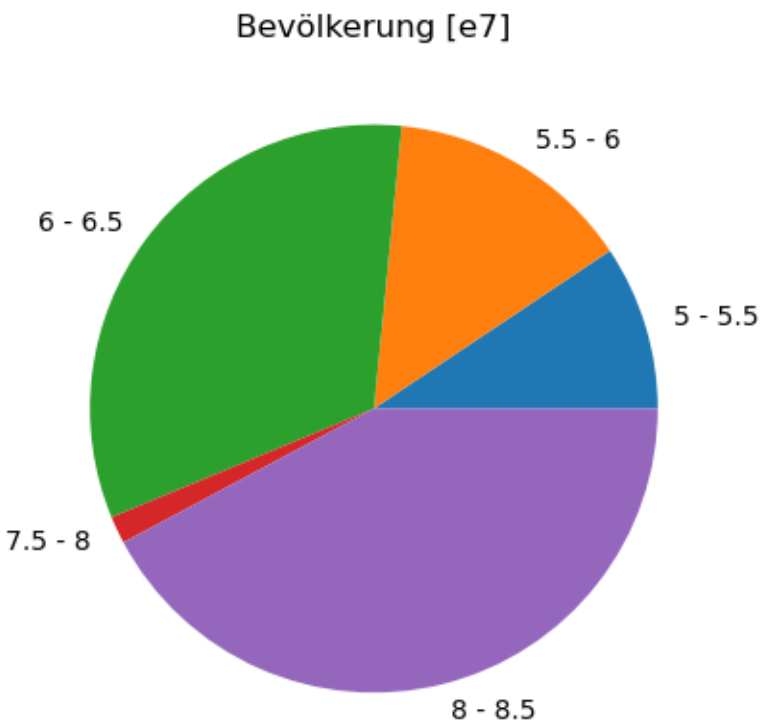


R2.14 Scatterplot



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

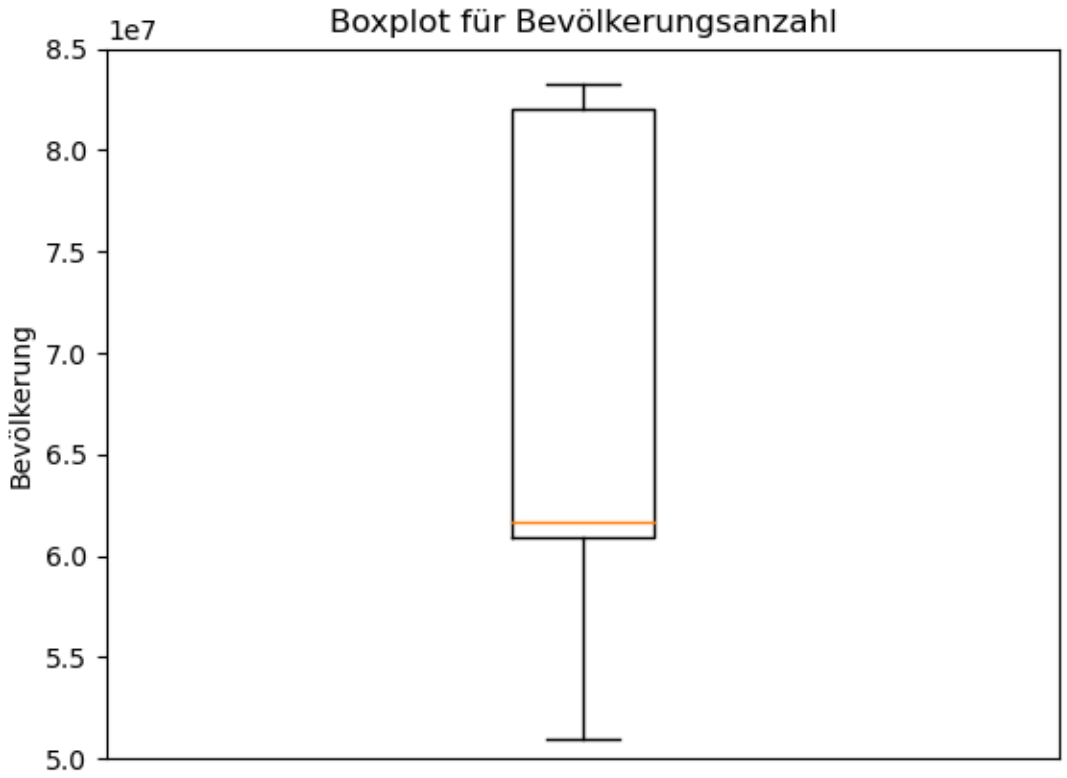
R3.12 Stichprobenvarianz

Stichprobenvarianz: 137735543999416.7

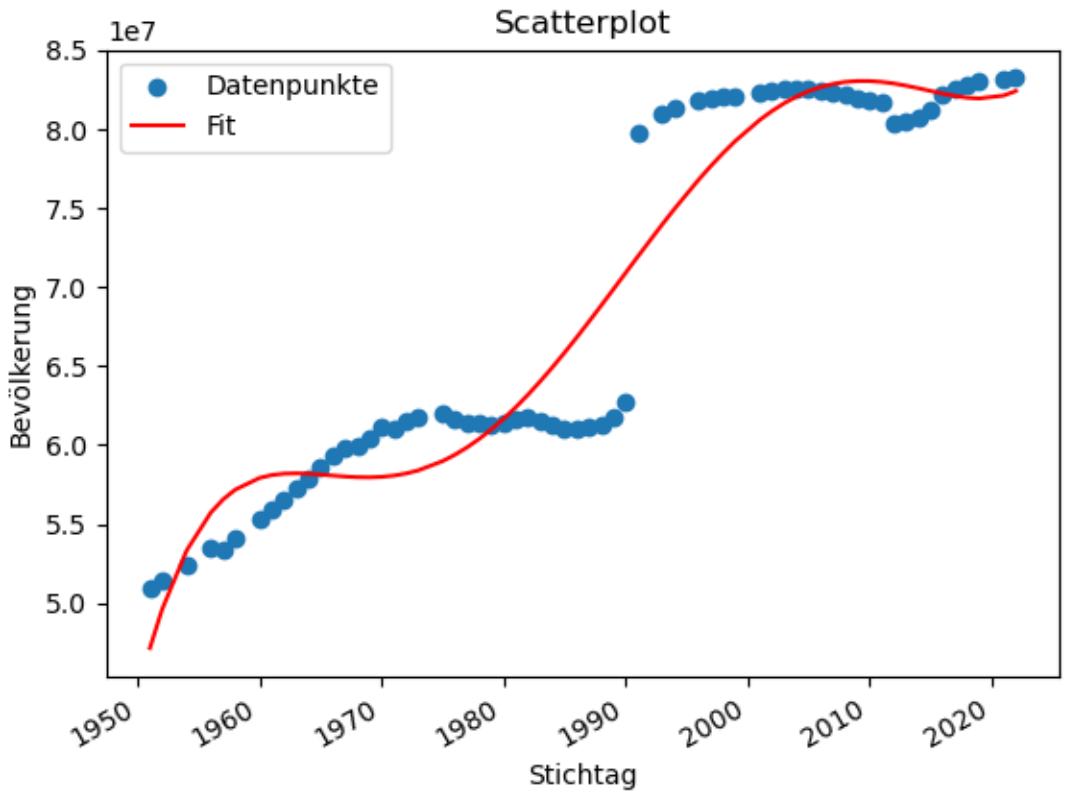
R3.13 Variationskoeffizient

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:

$9.77e\text{-}14\ x^5 - 3.579e\text{-}09\ x^4 + 2.626e\text{-}05\ x^3 + 0.1832\ x^2 + 145.2\ x + 5.799e{+}07$

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

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

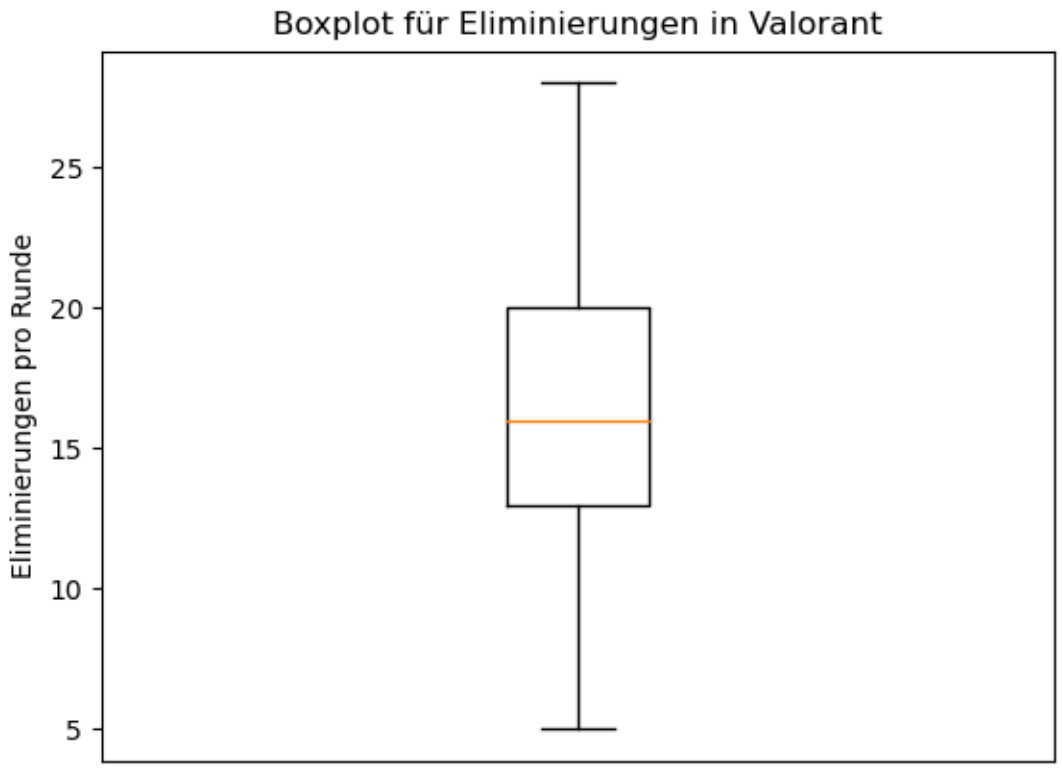
Angabe von Modus, arithmetischem Mittelwert und Median:

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 Eliminierungen pro Runde häufen sich im Bereich von 13 bis 20.

Quellen

Bücher:

Keine Panik vor Statistik – Markus Oestreich, Oliver Romberg

Sonstiges:

Vorlesungsmaterial