Grundlagen und Anwendung der

Wahrscheinlichkeitstheorie WS 24/25

Datenprojekt
Gruppe 11

Joan Kemeni (2. Semester), Moritz Semmelmann, Naresh Sivayogan,
Sam Fischer (3.Semester)

R1.1:
Der Datensatz enthält eine Tabelle mit dem Titel „Household prizes“. Diese besteht aus zwei Spalten: Die erste Spalte enthält Datumsangaben im Format „DD-MM-YYYY“, und die zweite Spalte gibt die „prozentuale Veränderung im Vergleich zum Vorjahr“ an. Dabei ist zu beachten, dass die Datumsangaben nicht täglich fortschreiten, sondern in Quartalsabständen angegeben sind. Das bedeutet, dass jeweils der erste Tag eines neuen Quartals direkt auf den ersten Tag des vorherigen Quartals folgt.
Der Zeitraum des Datensatzes erstreckt sich von 2000 bis 2023. Die Werte der zweiten Spalte, welche vermutlich die prozentuale Veränderung der Haushaltspreise im Vergleich zum Vorjahr beschreiben, können sowohl positive als auch negative Werte annehmen. Sie bewegen sich in einem Bereich von etwa -15 % bis +9 %.
Dieser Datensatz wurde aus der Online-Datenbank FRED (Federal Reserve Economic Data) der Federal Reserve Bank of St. Louis entnommen. Eine entsprechende Referenz findet sich unter der URL: https://fred.stlouisfed.org/series/QDER628BIS#0.
Im Quelltext ist die Datei mit dem Namen „data-1.csv“ gespeichert. Sie umfasst 94 Zeilen und ist im UTF8-Format kodiert. Jede Zeile enthält ein Datum und den dazugehörigen Prozentsatz.

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

R1.3:
Der Python und matlab Code wurde mit folgenden Programmen bearbeitet: SciPy, matlab , pandas, excel, matplotlib
Zudem wurde auch Excel, seine gängigen Funktionen sowie Statistik Add-ins verwendet.

R1.7:
Modus: es gibst keinen Modus, denn wenn es mehrere gleich häufig vorkommende Wert gibt, gibt mode den kleinsten dieser Werte zurück.
(Quelle : https://www.mathworks.com/help/matlab/ref/double.mode.html)

Arithmetischer Mittelwert: 0,67639

Median: 0,25056

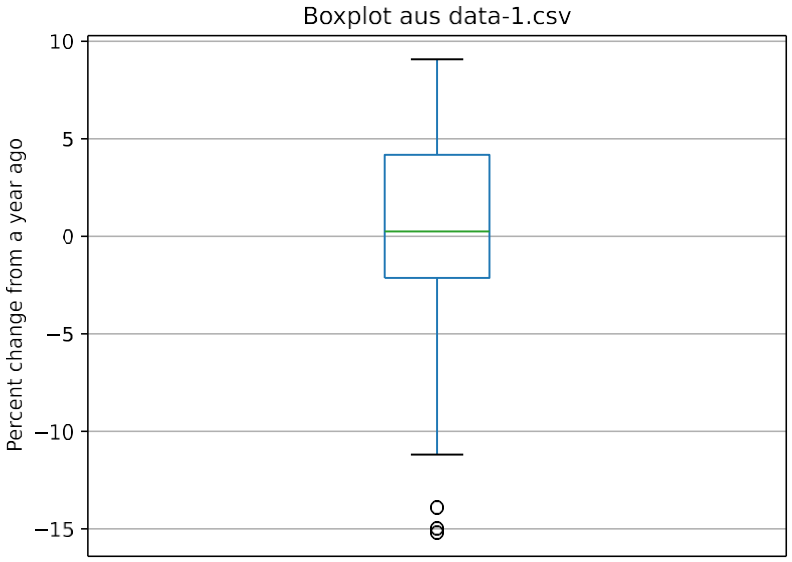
R1.8: *Die Spannweite:* 24,27

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

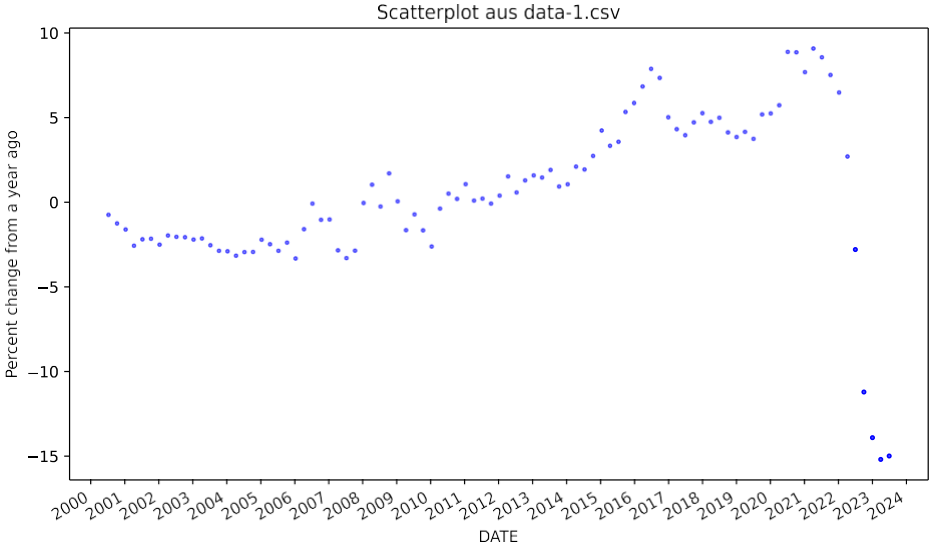
R1.10: *Die Stichprobenvarianz*: 21,5956.

R1.11: *Der Variationskoeffizient:* 6,9077.

R1.12:



R1.13:



R1.14:

Da alle Datenwerte nur einmal vorkommen, existiert kein Modus. Der Median liegt, wie im Box-Plot erkennbar, etwa 0,4 unter dem arithmetischen Mittelwert und befindet sich somit näher bei Q(0,25) als an der oberen Grenze der Box.
Die Spannweite der Prozentsätze beträgt 24,27 und erstreckt sich vom negativen in den positiven Bereich. Der Variationskoeffizient gibt ein relatives Streuungsmaß von 6,908 an, während die Stichprobenvarianz die durchschnittliche quadratische Abweichung vom Mittelwert mit einem Wert von 21,83 beschreibt.

R1.15:

Quartile: Q(0,25): -2,13345 ; Q(0,5/median): 0,25056 ; Q (0,75): 4,17768

Dezile:
D(10%) -2,8433
D(20%) - 2,4432
D(30%) -1,9570
D(40%) -0,68156
D(50%) 0,2506
D(60%) 1,5035
D(70%) 3,4849
D(80%) 4,7888
D(90%) 6,5771

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

R1.17:
Die Kovarianz= 4216,675

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:

R1.21:
Der Rangkorrelationskoeffizient laut Spearman ist ungefähr 0.14158

R2.1:
Dieser Datensatz besteht aus einer Tabelle mit zwei Spalten, die sowohl spalten- als auch zeilenweise fehlerhaft ist. In einigen Fällen fehlen Dateneinträge, oder die Jahre sind nicht fortlaufend, sondern springen beispielsweise um 100 Jahre. Grundsätzlich enthalten die Daten die jährlichen Lebendgeburtenzahlen für den Zeitraum von 1950 bis 2022.

Die Quelle der Daten ist das Statistische Bundesamt, entnommen aus der Statistik „Genesis Tabelle 12612-0001“ (https://www-genesis.destatis.de/datenbank/online/statistic/12612/table/12612-0001).

Der Datensatz ist in UTF8-Kodierung gespeichert, und der Quellcode der zugehörigen CSV-Datei umfasst 75 Zeilen.

R2.3:
Zuerst wurde ein Python-Programm erstellt, um die Zeileneinträge zu korrigieren. Zusätzlich wurden die Spaltennamen angepasst, und fehlerhafte Jahreszahlen, die das Plotten beeinträchtigt hätten, wurden manuell entfernt.

R2.4:
Der Python und matlab Code wurde mit folgenden Programmen bearbeitet: SciPy, matlab , pandas, excel, matplotlib

R2.8:
Modus: kein Modus, weil sich kein Wert wiederholt.
Arithmetischer Mittelwert= 911 790,8254
Median= 812 292

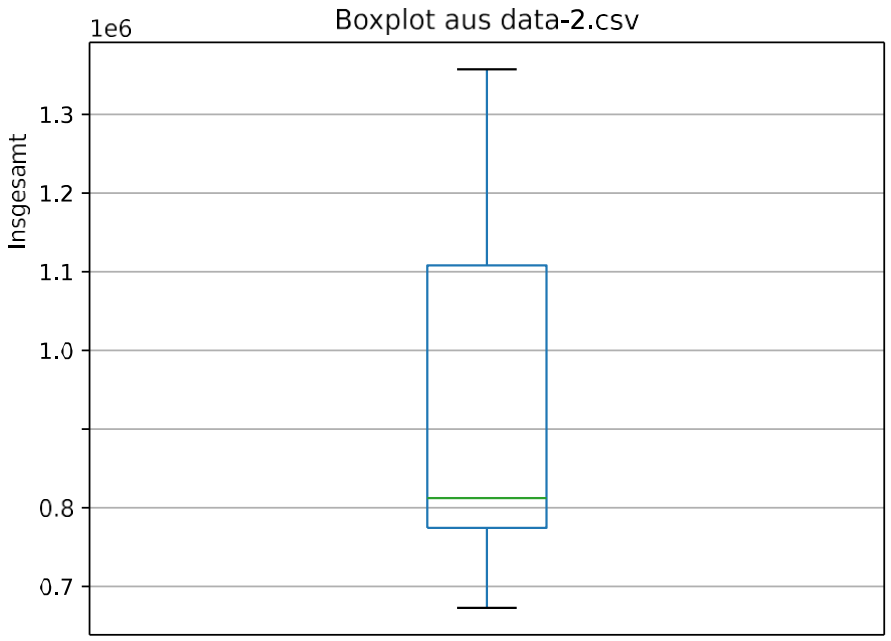
R2.9:
Die Spannweite = 684 580

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

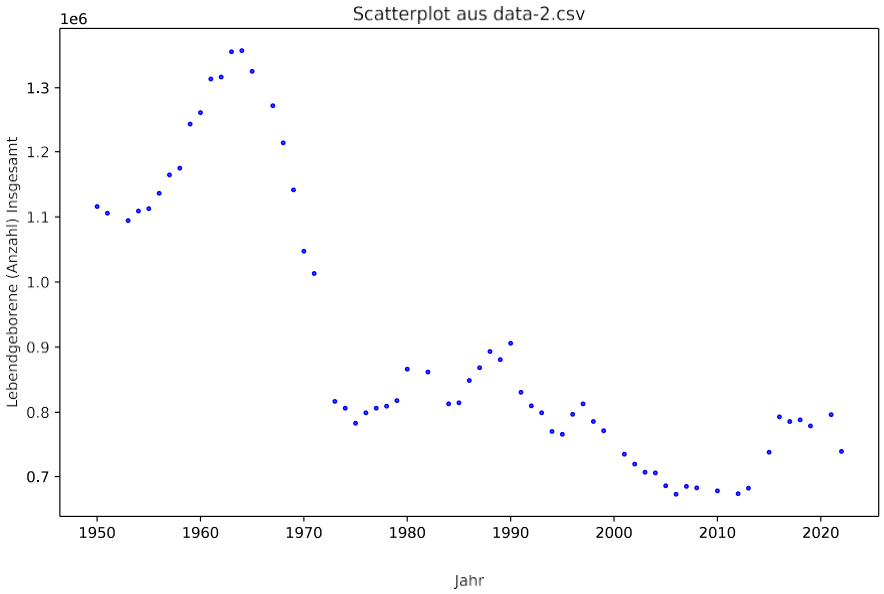
R2.11:
Die Stichprobenvarianz = 43 824 145 922,308

R2.12:
Der Variationskoeffizient = 0,22959452

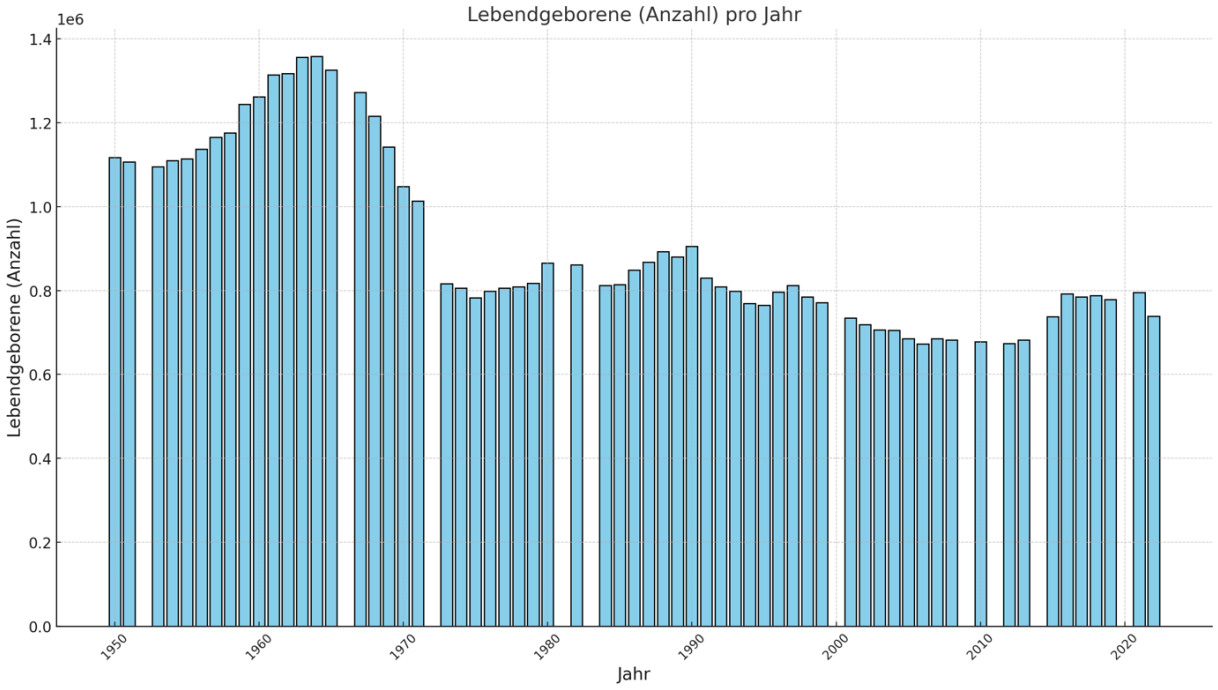
R2.13:



R2.14:



R2.15:



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

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 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= 333 644,5

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

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

R3.1:

Der dritte Datensatz ist in zwei Teile aufgeteilt und besteht aus den Dateien „data-3-a.csv“ und „data-3-b.csv“. Der A-Teil enthält eine Tabelle mit zwei Spalten, die keine Bezeichnungen aufweisen. In der ersten Spalte sind achtstellige Kombinationen aus Großbuchstaben und Ziffern verzeichnet, während in der zweiten Spalte die Jahreszahlen von 1950 bis 2022 angegeben sind.
Der B-Teil des dritten Datensatzes umfasst ebenfalls eine zweispaltige Tabelle. Hier werden dieselben Buchstaben-Ziffern-Kombinationen der ersten Spalte mit einer Anzahl von Gestorbenen in der zweiten Spalte gegenübergestellt. Die Anzahl der Zeilen im B-Teil entspricht dabei derjenigen im A-Teil.
Auch diese Daten stammen vom Statistischen Bundesamt und basieren auf der Statistik „Genesis Tabelle 12613-0001“ zu den Sterbefällen von 1950 bis 2024 (https://www-genesis.destatis.de/datenbank/online/table/12613-0001/search/s/MTI2MTMtMDAwMQ==).
Der Datensatz liegt im UTF8-Format vor und umfasst 75 Zeilen Code.

R3.4:
Ein Programm wurde entwickelt, um die Datensätze miteinander zu verbinden. Die Buchstaben-Ziffern-Codes geben jeweils an, welches Jahr mit welchem Wert verknüpft ist.

R3.6:
Der Python und matlab Code wurde mit folgenden Programmen bearbeitet: SciPy, matlab , pandas, excel, matplotlib

R3.9:
Modus: kein Modus, da alle Werte gleich häufig vorkommen
Arithmetischer Mittelwert= 892 568,3836
Median= 895 070

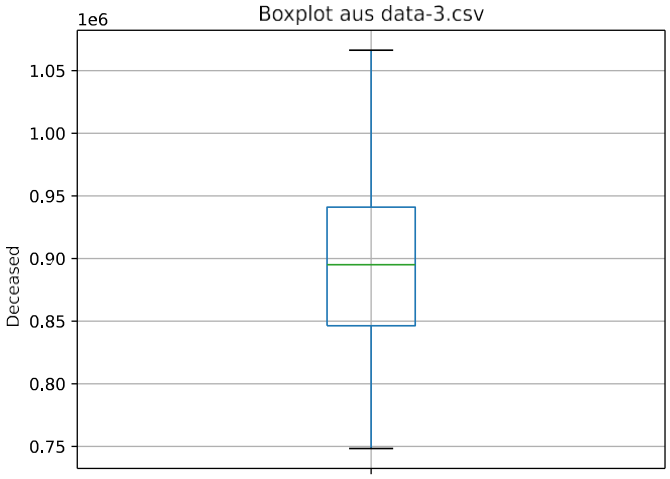
R3.10:
Die Spannweite= 318 012

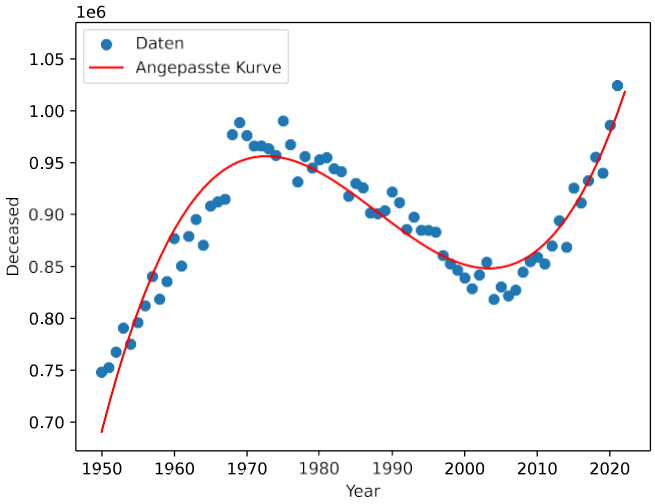
R3.11:
Die mittlere Abweichung vom Median = 52 593,2877

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

R3.13:
Der Variationskoeffizient = 0,07252

R3.14:





R3.17:

$(7.505471447409998 * X^{3}) + (-855.4214093339821 * X^{2}) + (27216.86531511555 * X) + (691116.8417620478)$

R.3.19:

Auch in diesem Fall existiert kein Modus, da alle Werte nur einmal auftreten. Der arithmetische Mittelwert beträgt 892 568,3836 und liegt geringfügig unter dem Median, der bei 895 070 liegt.
Die Spannweite beträgt 318 012, während die mittlere Abweichung vom Median 52 593,2877 beträgt. Die Stichprobenvarianz beläuft sich auf 4 189 857 093,379.
Der Variationskoeffizient liegt bei 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 = 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.

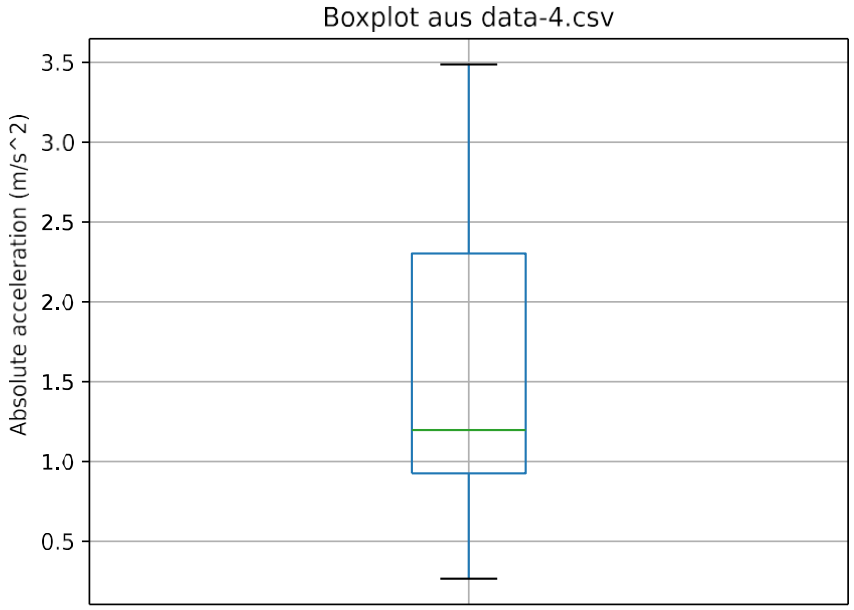
R4.4:
- Phyphox
- VS-Code
- Microsoft-Excel

R4.5:

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

R4.6:
Die Stichprobenvarianz = 0,8461

R.4.7:



R4.8:

Es gibt zwei Modi (1,105 und 1,135), die jeweils zweimal in den Messwerten auftreten. Der arithmetische Mittelwert liegt mit ungefähr 1,4 m/s² unter dem Median, der bei etwa 1,75 m/s² liegt. Die Stichprobenvarianz, die die mittlere Abweichung der gemessenen Werte vom empirischen Mittelwert beschreibt, beträgt in dieser Datenerhebung etwa 0,94.