

Abgabe 1 für Computergestützte Methoden

Gruppe (36), (Johnnie Butros 3960396, Erick Lapp 4111943)

01.12.2024

Inhaltsverzeichnis

1	Der zentrale Grenzwertsatz	2
1.1	Aussage	2
1.2	Erklärung der Standardisierung	2
1.3	Anwendungen	2
2	Aufgabe 1: Datenhaltung & -aufbereitung	3
2.1	Thema Datenverarbeitung	3
2.1.1	3
2.1.2	3
2.1.3	3
2.2	Thema Datenhaltung	3
2.2.1	3
2.2.2	4
2.2.3	4
2.2.4	4
2.2.5	4
3	Quellenangaben	6

1 Der zentrale Grenzwertsatz

Der zentrale Grenzwertsatz (ZGS) ist ein fundamentales Resultat der Wahrscheinlichkeitstheorie, das die Verteilung von Summen unabhängiger, identisch verteilter (i.i.d.) Zufallsvariablen (ZV) beschreibt. Er besagt, dass unter bestimmten Voraussetzungen die Summe einer großen Anzahl solcher ZV annähernd normalverteilt ist, unabhängig von der Verteilung der einzelnen ZV. Dies ist besonders nützlich, da die Normalverteilung gut untersucht und mathematisch handhabbar ist.

1.1 Aussage

Sei X_1, X_2, \dots, X_n eine Folge von i.i.d. ZV mit dem Erwartungswert $\mu = \mathbb{E}(X_i)$ und der Varianz $\sigma^2 = \text{Var}(X_i)$, wobei $0 < \sigma^2 < \infty$ gelte. Dann konvergiert die standardisierte Summe Z_n dieser ZV für $n \rightarrow \infty$ in Verteilung gegen eine Standardnormalverteilung:

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (1)$$

Das bedeutet, dass für große n die Summe der ZV näherungsweise normalverteilt ist mit Erwartungswert $n\mu$ und Varianz $n\sigma^2$.

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2). \quad (2)$$

1.2 Erklärung der Standardisierung

Um die Summe der ZV in eine Standardnormalverteilung zu transformieren, subtrahiert man den Erwartungswert $n\mu$ und teilt durch die Standardabweichung $\sigma\sqrt{n}$. Dies führt zu der obigen Formel (1). Die Darstellung (2) ist für $n \rightarrow \infty$ nicht wohldefiniert.

1.3 Anwendungen

Der ZGS wird in vielen Bereichen der Statistik und der Wahrscheinlichkeitstheorie angewendet. Typische Beispiele sind:

- **Schätzen von Mittelwerten:** Beim Ziehen von Stichproben aus einer Grundgesamtheit kann der Stichprobenmittelwert durch den ZGS als näherungsweise normalverteilt betrachtet werden, wenn die Stichprobengröße groß genug ist.
- **Monte-Carlo-Simulationen:** Der ZGS wird verwendet, um Summen oder Durchschnittswerte von simulierten Zufallsvariablen zu approximieren.

Der zentrale Grenzwertsatz hat verschiedene Verallgemeinerungen. Eine davon ist der **Lindeberg-Feller-Zentrale-Grenzwertsatz** [1, Seite 328], der schwächere Bedingungen an die Unabhängigkeit und die identische Verteilung der ZV stellt.

2 Aufgabe 1: Datenhaltung & -aufbereitung

2.1 Thema Datenverarbeitung

2.1.1

Die Daten sind in 12 Spalten aufgeteilt die für group, station, date, day_of_year, day_of_week, month_of_year, precipitation, windspeed, min_temperature, average_temperature, max_temperature und count stehen und somit unsere Attribute sind. Unsere Gruppe ist die Nummer 36 und demnach ist die Station, die wir betrachten die Centre St & Chambers St. Wir betrachten den Zeitraum von 1 Jahr und starten am 01.01.2023 bis 31.12.2023. Wir erkennen auch das Daten wie day_of_week, day_of_month und month_of_year schon in date erkennbar sind.

2.1.2

Die CSV in Excel importieren, um durch Trennzeichen für jede Spalte ein Attribut zu bekommen. Dadurch wird es viel geordneter und wir können leichter unsere gesuchten Werte finden.

2.1.3

Wir gehen in Excel auf die Spalte group und wählen nur die mit der 36 aus, um nur die Daten zu bekommen die für unsere Gruppe relevant sind. Mit derselben Methode können wir die anderen Spalten von fehlenden oder unrealistischen Werten bereinigen. Die höchste mittlere Temperatur in Excel berechnen durch =MAX(J12750:J13112) welcher 83 ist und da es in Fahrenheit noch ist, müssen wir umrechnen in Celsius $(83-32) \cdot (5/9) = 28,33333333$.

2.2 Thema Datenhaltung

2.2.1

SQLite ist flexibel, was die Definition von Datentypen betrifft. Im Gegensatz zu vielen anderen relationalen DBMS erlaubt SQLite eine breite Flexibilität bei der Wahl der Datentypen, die bei der Erstellung von Tabellen verwendet werden.

Hier sind einige der gängigsten Datentypen in SQLite:

NULL:

- Repräsentiert einen NULL-Wert (keinen Wert oder unbekannten Wert).

INTEGER:

- Wird verwendet, um Ganzzahlen zu speichern.
- SQLite verwendet in der Regel die INTEGER-Typen für Datentypen wie int, smallint oder bigint in anderen Systemen.
- Es können Ganzzahlen unterschiedlicher Größe gespeichert werden, wobei SQLite den Typ dynamisch anpasst.

REAL:

- Wird verwendet, um Fließkommazahlen (mit Dezimalstellen) zu speichern.
- Entspricht den Datentypen float oder double in anderen Systemen.

TEXT:

- Wird verwendet, um Textzeichenfolgen zu speichern.
- Speichert beliebige Textdaten, von kurzen Zeichenketten bis hin zu großen Textblöcken.

BLOB:

- Wird verwendet, um Binärdaten zu speichern (z. B. Bilder oder Dateien).
- BLOBs können beliebige binäre Daten speichern und werden als Rohdaten behandelt.

2.2.2

Datenbank Schema entwerfen

1.Normalform = Trennung nicht-atomarer Attribute

Station	Date	Day_of_Year	Day_of_Week	Month_of_Year	Precipitation	Windspeed	Min_Temp.	Avg_Temp.	Max_Temp.	Count
Centre St & Chambers St	01.01.2023	1	1	1	0	1007	42	50	56	164

2.Normalform= Auftrennung in mehrere Tabellen und Fremdschlüssel-Beziehungen mit passenden Abhängigkeiten

Tabelle: Fahrradverleih

ID#	StationID#	Date	Precipitation	Windspeed	Min_Temp	Avg_Temp	Max_Temp	Count
-----	------------	------	---------------	-----------	----------	----------	----------	-------

Tabelle: Station

StationID#	Station
------------	---------

Primärschlüssel in Blau

Schlüsselattribute in Grün

2.2.3

```
sql = '''CREATE TABLE STATION(  
    ID INTEGER PRIMARY KEY,  
    Station TEXT NOT NULL  
)'''  
sql = '''CREATE TABLE FAHRRADVERLEIH(  
    ID INTEGER PRIMARY KEY,  
    StationID INTEGER,  
    date TEXT NOT NULL,  
    Percipitation INTEGER NOT NULL,  
    windspeed INTEGER NOT NULL,  
    min_temperature INTEGER NOT NULL,  
    average_temperature INTEGER NOT NULL,  
    max_temperature INTEGER NOT NULL,  
    count INTEGER NOT NULL,  
    FOREIGN KEY(StationID) REFERENCES STATION(ID)  
)'''
```

2.2.4

Durch diesen Befehl können wir unsere csv Datei importieren und durch unsere definierten Tabellen wiedergeben lassen.

```
import pandas as pd  
df = pd.read_csv('bike_sharing_data_(with_NAs).csv')  
print(df)
```

2.2.5

Literatur

- [1] Achim Klenke. Wahrscheinlichkeitstheorie. Springer, 3. edition, 2013.

3 Quellenangaben

<https://github.com/JohnnieButros/Abgabe-1-Computergestuetzte-Methoden.git>