

Abgabe 1 für Computergestützte Methoden

Gruppe 50

Rebekka Ewert, Ekaterini Skampali, Eleonora Kamysni

01.12.2024

Inhaltsverzeichnis

1 Der zentrale Grenzwertsatz	2
1.1 Aussage	2
1.2 Erklärung der Standardisierung	2
1.3 Anwendungen	2
2 Datenhaltung & -aufbereitung	3
2.1 Thema Datenverarbeitung	3
2.1.1 Importieren des Datensatzes in einer Tabellenkalkulation	3
2.1.2 Entwurf des Datenbank-Schemas (1. und 2. Normalform)	4
2.1.3 Untersuchung der Daten	4
2.1.4 Berechnung der höchsten mittleren Temperatur	5
2.2 Thema Datenhaltung	6
2.2.1 Entwurf eines Datenbank-Schemas	6
2.2.2 Definition der Tabellen in SQL	7
2.2.3 Vorbereitung und Import des Datensatzes	7
2.2.4 SQL-Abfrage zu höchster mittlerer Temperatur	8

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 $\mu = E(X_i)$ 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:¹

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

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

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

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:

¹Der 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.

- **Schätzung von Mittelwerten in großen Stichproben:** Der Zentrale Grenzwertsatz ermöglicht es, die Verteilung des Stichprobenmittelwerts \bar{X} einer großen Zufallsstichprobe näherungsweise als normalverteilt anzunehmen, unabhängig von der Verteilung der Grundgesamtheit. Dies ist besonders nützlich in der statistischen Inferenz, beispielsweise beim Konstruieren von Konfidenzintervallen oder beim Testen von Hypothesen über den Mittelwert. ²
- **Qualitätssicherung in der Produktion:** In der Qualitätskontrolle werden Messwerte wie z. B. Gewicht, Länge oder Fehleranzahl von Produkten oft als Zufallsvariablen modelliert. Durch die Anwendung des ZGS kann man davon ausgehen, dass die Summe oder der Mittelwert der Messwerte einer großen Stichprobe näherungsweise normalverteilt ist. Dies erleichtert die Anwendung von Kontrollcharts oder Entscheidungsregeln, die auf der Normalverteilung basieren. ³

2 Datenhaltung & -aufbereitung

2.1 Thema Datenverarbeitung

2.1.1 Importieren des Datensatzes in einer Tabellenkalkulation

Zuerst haben wir die CSV-Datei, die uns zur Verfügung gestellt wurde, in Excel importiert. Dazu haben wir die Option **Daten > Aus Text/CSV** verwendet. Diese Datei enthielt Daten in einem textbasierten, durch Kommas getrennten Format. Während des Imports wurde überprüft, dass jede Spalte korrekt formatiert ist. Beispielsweise wurde die Spalte „date“ als Datumsformat erkannt, während numerische Spalten wie `precipitation`, `windspeed` oder `average_temperature` korrekt, als Zahlen importiert wurden. Danach haben wir unsere Datensatz **E 12 St & Ave C** isoliert. Fehlende Werte, wie das **NA**, wurden beibehalten, um diese später im Analyseprozess zu berücksichtigen.

Die erfolgreiche Datenübernahme in die Tabellenkalkulation ermöglichte es, Berechnungen wie die Ermittlung der höchsten mittleren Temperatur direkt durchzuführen. Diese Vorbereitung legt auch den Grundstein für die weitere Verarbeitung der Daten in einer Datenbank.

²<https://statologie.de/zentraler-grenzwertsatz/>

³<https://martin-grellmann.de/was-ist-der-zentrale-grenzwertsatz-und-warum-ist-der-wichtig>

2.1.2 Entwurf des Datenbank-Schemas (1. und 2. Normalform)

Die Kombination aus **Datum** und **Anzahl** identifiziert jede Zeile eindeutig und dient als Primärschlüssel. Der Datensatz erfüllt die 1. Normalform (1NF), da jede Spalte atomare Werte enthält und keine Mehrfachwerte oder sich wiederholenden Gruppen vorliegen. In der 2. Normalform (2NF) müssen alle Nicht-Prime-Attribute vollständig vom gesamten Primärschlüssel abhängen. Dies ist hier der Fall, da Attribute wie **precipitation**, **windspeed** oder **average_temperature** ausschließlich von **Datum** und **Anzahl** abhängen. Es gibt keine partiellen Abhängigkeiten, sodass die Tabelle bereits in der 2NF vorliegt.

Eine Optimierung wäre, die Spalte **station** in eine separate Tabelle auszulagern, da diese Informationen unabhängig von den Wetterdaten sind und sich wiederholen. Die Wetterdaten könnten dann über einen Fremdschlüssel mit einer neuen Stations-Tabelle verknüpft werden. Dadurch wird die Datenbank strukturierter, wartbarer und redundanzfrei.

2.1.3 Untersuchung der Daten

Zu Beginn haben wir den für unsere Gruppe relevanten Datensatz, **E 12 St & Ave C** untersucht und uns mit seinem Aufbau vertraut gemacht. Im Folgenden beschreiben wir unsere Erkenntnisse. Der Datensatz umfasst Daten im Zeitraum vom 01.01. bis zum 31.12.2023 mit insgesamt 365 Zeilen und 12 Spalten. Neben den Spalten **Gruppe**, **Station** und **Datum**, gibt es noch die Spalten **Tag im Jahr**, **Tag in der Woche** und **Monat im Jahr**, sowie weitere sechs Spalten zu verschiedenen Abfragen. In diesen sind die Informationen Windgeschwindigkeit, Niederschlag, maximale und minimale Temperatur, Durchschnittstemperatur und Zählung enthalten.

In der Spalte **day_of_year** (Tag im Jahr) steht am 01.08. und 06.08.2023 die Bezeichnung **NA** in den Zellen. Dieses Problem ist einfach zu beheben, da man sich an den vorangegangenen Daten orientieren kann. Somit ergibt sich für den 01.08.2023 der Tag 213 im Jahr und für den 06.08.2023 der Tag 218 im Jahr.

In der Spalte **day_of_week** (Tag in der Woche) steht am 23.09.2023 ebenfalls die Bezeichnung **NA** in der Zelle. Dieses Problem ist ebenfalls einfach zu beheben, da man sich am vorangegangenen Datum orientieren kann, wodurch sich für den 23.09.2023 der Tag 7 der Woche ergibt.

In der Spalte **precipitation** (Niederschlag) wurde am 24.03.2023 ein negativer Wert **-1** verzeichnet, was jedoch falsch ist, da es keinen Niederschlag > 0 gibt. Des Weiteren ist die Zelle am 10.11.2023 leer, daher ergänzen wir hier die Bezeichnung **NA**, da der Wert 0 nicht korrekt wäre. Es wäre nicht

korrekt, weil der Wert 0 tatsächlich eine spezifische Aussage darstellt, nämlich dass an diesem Datum ein Messwert von 0 vorlag. Da jedoch nicht eindeutig feststeht, ob an diesem Datum keine Messung durchgeführt wurde oder ob der Wert schlicht vergessen wurde, ist NA die passendere Wahl.

In der Spalte `windspeed` (Windgeschwindigkeit) wurde am 28.05.2023 ein negativer Wert -1 verzeichnet, was jedoch falsch ist, da es keine Windgeschwindigkeit > 0 gibt. Des Weiteren ist die Zelle am 15.09.2023 leer, daher ergänzen wir hier Bezeichnung NA, da der Wert 0 nicht korrekt wäre. Dies würde nämlich suggerieren, dass an dem Datum nicht gemessen wurde, was möglicherweise sein kann, jedoch könnte es auch der Fall sein, dass bloß vergessen wurde einen Wert einzutragen.

In der Spalte `min.temperature` (Minimale Temperatur) steht die Bezeichnung NA am 03.10.2023 in der Zelle. Dennoch wird die `average.temperature` (Durchschnittstemperatur) mit 65 angegeben.

In der Spalte `average.temperature` (Durchschnittstemperatur) steht am 20.01.2023 der Wert -1, welcher nicht korrekt sein kann. Dieses Problem lässt sich jedoch beheben, indem man die Durchschnittstemperatur eigenständig berechnet mithilfe der gegebenen maximalen und minimalen Temperatur, also

$$=([\text{@}[\text{min.temperature}]] + [\text{@}[\text{max.temperature}]])/2,$$

d.h. wir rechnen $(39+50)/2 = 45$ und ergänzen diesen Wert in der Zelle.

Außerdem steht in der Spalte `average.temperature` (Durchschnittstemperatur) am 26.02.2023 die Bezeichnung NA in der Zelle. Dieses Problem lässt sich jedoch ebenfalls beheben, indem man die Mitteltemperatur eigenständig berechnet, wie bereits oben geschehen, d.h. wir rechnen $(23+50)/2 = 36,5$ und ergänzen diesen Wert in der Zelle.

In der Spalte `max.temperature` (maximale Temperatur) fehlt der Wert am 12.05.2023, d.h. die Zelle ist leer. Dennoch ist eine Durchschnittstemperatur gegeben. In der Spalte `count` wurde am 28.06.2023 eine Zählung von -1 erfasst, was faktisch nicht korrekt sein kann, da eine Zählung keinen Wert > 0 annehmen kann.

Des Weiteren ist uns aufgefallen, dass die Berechnung der Durchschnittstemperatur in der Spalte `average.temperature` nur in 82 von 365 Zeilen korrekt war. In den restlichen Zeilen gab es Abweichungen von -46 bis +65 Grad Fahrenheit. Diese Abweichungen sind uns aufgefallen, da wir eine Spalte ergänzt haben, in der wir die Durchschnittstemperatur nachgerechnet haben.

2.1.4 Berechnung der höchsten mittleren Temperatur

Um die gegebenen Durchschnittstemperaturen von Fahrenheit in Grad Celsius umzurechnen, ergänzen wir eine neue Spalte (N). Wir nutzen für die

Umrechnung die Formel

$$= ([@average_temperature] - 32) * 5/9.$$

Um nun die höchste mittlere Temperatur herauszufinden, haben wir den Befehl = Max [] auf die Spalte angewendet. Dabei kam der Wert 28,33 Grad Celsius raus. Wenn wir den Befehl = Max [] ebenfalls auf unsere Spalte mit den neu berechneten Durchschnittstemperaturen anwenden, erhalten wir den Wert 28,1 Grad Celsius.

2.2 Thema Datenhaltung

2.2.1 Entwurf eines Datenbank-Schemas

Der Datensatz wurde analysiert, um eine geeignete Struktur in der Datenbank zu entwickeln, die den Anforderungen der 1. und 2. Normalform entspricht. D.h. in der ersten Normalform enthalten alle Spalten atomare Werte und es gibt keine Mehrfachwerte in einer Zelle. In der zweiten Normalform wurde die Tabelle zur Vermeidung von Redundanzen in zwei logische Tabellen aufgeteilt. Eine Tabelle für die Stationsdaten und eine Tabelle für die Wetterdaten.

Tabelle Stations:

- station_id: Eindeutige ID für jede Station (Primärschlüssel)
- station_name: Name der Station (z. B. E 12 St & Ave C)
- group: Gruppenzugehörigkeit der Station (z. B. 1)

Tabelle WeatherData:

- id: Eindeutige ID für jeden Datensatz (Primärschlüssel)
- station_id: Verweis auf die Station (Fremdschlüssel)
- date: Datum im Format YYYY-MM-DD
- count: Anzahl der Einträge pro Tag
- precipitation: Niederschlagsmenge in Millimetern
- windspeed: Windgeschwindigkeit in km/h
- min_temperature: Minimale Temperatur des Tages
- average_temperature: Durchschnittstemperatur des Tages
- max_temperature: Maximale Temperatur des Tages

2.2.2 Definition der Tabellen in SQL

```
1 SELECT * FROM bike_sharing_data_with_NAs
2 WHERE Station = 'E 12 St & Ave C';
```

Abbildung 1: Import und Isolierung unseres Datensatzes

SQL DDL-Befehle für das Schema:

```
1 CREATE TABLE Station (
2     station_id INTEGER PRIMARY KEY AUTOINCREMENT,
3     station_name TEXT NOT NULL
4 );
```

Abbildung 2: Tabelle Station

```
1 CREATE TABLE Wetterdaten (
2     weather_id INTEGER PRIMARY KEY AUTOINCREMENT,
3     station_id INTEGER NOT NULL,
4     date DATE NOT NULL,
5     day_of_year INTEGER,
6     day_of_week INTEGER,
7     month_of_year INTEGER,
8     precipitation REAL,
9     windspeed REAL,
10    min_temperature REAL,
11    average_temperature REAL,
12    max_temperature REAL,
13    COUNT INTEGER,
14    FOREIGN KEY (station_id) REFERENCES Station(station_id)
15 );
```

Abbildung 3: Tabelle Wetterdaten

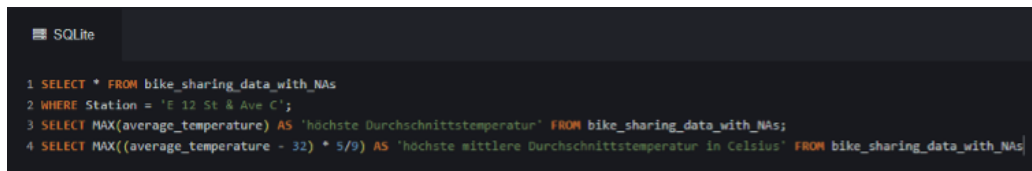
2.2.3 Vorbereitung und Import des Datensatzes

Der vollständige Datensatz wurde in zwei Teile aufgeteilt, um den Tabellen `stations` und `temperatures` zu entsprechen. Die Spalte `station` wurde extrahiert und jeder `station` eine eindeutige `station_id` zugeordnet.

Vorgehen: Der Datensatz wurde in SQLite importiert und die relevanten Daten E 12 St & Ave C durch Filtern extrahiert. Es wurden zwei separate CSV-Dateien erstellt. Die `stations.csv`-Datei mit `station_id` und `station_name` und die `temperatures.csv`-Datei mit den übrigen Attributen und der `station_id`. Diese Dateien wurden mit den Befehlen `.mode csv` und `.import` in SQLite geladen.

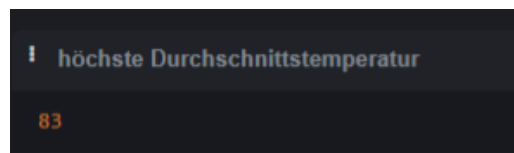
2.2.4 SQL-Abfrage zu höchster mittlerer Temperatur

Formulierung einer SQL-Abfrage, um die höchste mittlere Temperatur in Grad Celsius aus den Ihrer Gruppe zugeordneten Daten zu ermitteln.



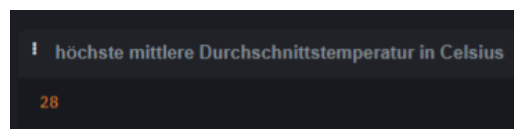
```
SQLite
1 SELECT * FROM bike_sharing_data_with_NAs
2 WHERE Station = 'E 12 St & Ave C';
3 SELECT MAX(average_temperature) AS 'höchste Durchschnittstemperatur' FROM bike_sharing_data_with_NAs;
4 SELECT MAX((average_temperature - 32) * 5/9) AS 'höchste mittlere Durchschnittstemperatur in Celsius' FROM bike_sharing_data_with_NAs;
```

Abbildung 4: SQL-Abfrage



```
! höchste Durchschnittstemperatur
83
```

Abbildung 5: Höchste Durchschnittstemperatur in Grad Fahrenheit



```
! höchste mittlere Durchschnittstemperatur in Celsius
28
```

Abbildung 6: Höchste Durchschnittstemperatur in Grad Celsius

Literatur

- [1] Achim Klenke. *Wahrscheinlichkeitstheorie*. Springer, 3. Auflage, 2013.
- [2] Statologie. *Der zentrale Grenzwertsatz: Grundlagen und Anwendungen*. Online verfügbar unter <https://statologie.de/zentraler-grenzwertsatz>.
- [3] Martin Grellmann. *Was ist der zentrale Grenzwertsatz und warum ist er wichtig?*. Online verfügbar unter <https://martin-grellmann.de/was-ist-der-zentrale-grenzwertsatz>