Abgabe 1 für Computergestützte Methoden

Gruppe (88), Anas El Fahsi (4242194), Arda Bostanci (4339434),
Felix Norbert Figge (4361841)

(2. Dezember 2024)

Inhaltsverzeichnis

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

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

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

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

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 $n \to \infty$ nicht wohldefiniert.

1.3 Anwendungen

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

- Berechnung von Konfidenzintervallen
- Durchführung von Hypothesentests

[1] 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] Bei der Schätzung der oberen und unteren Grenze wird eine Transformation in die Standardnormalverteilung mittels des zentralen Grenzwertsatzes verwendet [[2], Kapitel 3.5 unter „Konfidenzintervalle“, Seite 101-102]

[3] Zur Bestimmung der Irrtumswahrscheinlichkeit werden viele Stichproben gezogen bis nach dem zentralen Grenzwertsatz eine Normalverteilung vorliegt [[2], Kapitel 4 unter „Bestimmung der Irrtumswahrscheinlichkeit“, Seite 112]

2 Bearbeitung zur Aufgabe 1

In dieser Aufgabe sollen wir aus einem vorgegebenen Datensatz die höchste mittlere Temperatur berechnen – zum einen mithilfe einer Tabellenkalkulation und zum anderen durch das Entwerfen eines Datenbankschemas, das mit den importierten Daten in SQL umgesetzt wird. Anschließend berechnen wir die durchschnittliche Höchsttemperatur mit einer passenden SQL-Abfrage.

2.1 Berechnung per Tabellenkalkulation

Als erstes haben wir den Datensatz in Excel importiert und nach unserer Gruppennummer 88 gefiltert, damit wir nur mit den für uns relevanten Teil arbeiten. Anschließend haben wir auf die Spalte *average_ temperature* einen Filter angewendet, der die leeren Zellen ausschließt, damit wir bei der Abfrage nach der höchsten mittleren Temperatur keine Probleme haben. Dann haben wir mit dem MAX(...) Befehl in Excel den gewünschten Wert bestimmt [1].

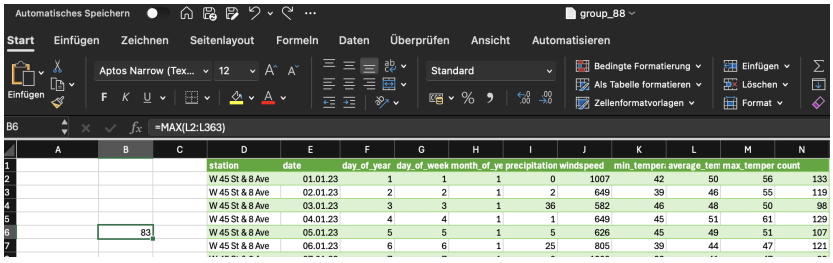

Abbildung 1: Zur Berechnung der höchsten mittleren Temperatur wurde über die Spalte L von den Zeilen 2 bis 363 der MAX(...) Befehl verwendet, wodurch wir auf den gesuchten Wert kamen.

Das Ergebnis sind 83° F. Die Einheit ist Fahrenheit, da dem Datensatz zu entnehmen ist, dass die Temperaturen in einer Straße in New York gemessen wurden. Um die Temperatur in Grad Celsius zu berechnen nehmen wir uns folgende Formel zur Hilfe:

$$^\circ\,\mathrm{C} = \frac{5}{9}(^\circ\,\mathrm{F} - 32). \tag{3}$$

Setzt man 83° F in die Formel ein erhält man als Ergebnis $28,3^\circ$ C.

[4]Zur Umrechnung von Grad Fahrenheit zu Grad Celsius wurde die oben stehende Formel aus dem Buch von Tipler herangezogen [[3], Kapitel 13.2, Seite 491, Formel 13.2.]

2.2 Berechnung mit SQL

Zuallererst haben wir die Zeilen entfernt, bei denen in der Spalte *average_ temperature* kein Wert vorhanden war, um die spätere Abfrage in SQL zu erleichtern. Basierend auf den Prinzipien der Datenbanknormalisierung haben wir folgendes Schema erstellt:

1. Eine dates-Tabelle mit einer id-Spalte als Primärschlüssel und den Spalten für Datum und Station.
2. Eine measurements-Tabelle, die alle Messungen (z. B. Temperaturen, Niederschläge) speichert und über einen Fremdschlüssel (die id) mit der dates-Tabelle verknüpft ist.

So schaffen wir es, Redundanzen zu minimieren und sicherzustellen, dass beide Tabellen die 1. und 2. Normalform erfüllen. Wir haben also zunächst die beiden CSV-Dateien mit den richtigen Spalten erstellt, die Tabellen in SQL definiert und die Daten mithilfe der SQL-Befehle aus dem Skript in die jeweiligen Tabellen importiert (siehe Abbildung 2).

Abbildung 2: Hier haben wir in Sqlite3 auf Mac die beiden Tabellen erstellt und die jeweiligen CSV-Dateien in die Tabellen importiert

Dann haben wir mithilfe der Abfrage nach der höchsten mittleren Temperatur den Wert 83 erhalten (Abbildung 3).

Abbildung 3: Abfrage nach der höchsten mittleren Temperature aus der measurements Tabelle

Wie bei der Berechnung mit Excel bereits erwähnt haben wir Temperaturen in Fahrenheit. Der Wert 83 ist derselbe, wie der den wir bei der Abfrage mit Excel erhalten haben und damit erhalten wir wieder eine Temperatur von $28,3°$ C.

3 Github

Wir haben unser Overleaf Dokument auf Github als ZIP-Datei und nochmal die einzelnen Dateien hochgeladen. Klicken Sie auf diesen Text um den Link zu öffnen.

[6]Das Datenbankschema wurde nach der 1. und 2. Normalform erstellt in der es heißt, dass alle Attribute atomar sein müssen und die Tabellen in mehrere Tabellen aufgetrennt werden müssen und mit Fremdschlüsseln in Beziehung gesetzt werden. [[4], unter dem Punkt „Vorüberlegungen: Normalformen“, S. 8.]

[5]Die DDL Befehle haben wir aus dem Skript [[4], unter dem Punkt „Schritt 1: Übersetzung in SQL“, S. 11.]

Literatur

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

[2] Jürgen Bortz. *Statistik für Human- und Sozialwissenschaften*. Springer, 6. edition, 2005.

[3] Paul A. Tipler and Gena Mosca. *Tipler Physik für Studierende der Naturwissenschaften und Technik*. Springer, 9. edition, 2024.

[4] Mike Teßmer. Praktische datenhaltung und -verarbeitung. Online, n.d. Zugriffsdatum: 28.11.2024.