Data Science Praktikum 2
Korrelation und Stochastik
Sommersemester 2024

Prof. Dr. Marina Tropmann-Frick
Tobias Schreier

25. April 2024

Wichtig:

Die Aufgaben 1 bis 3 sollen auf einem Zettel per Hand ohne Nutzung von Python oder Excel erledigt werden (Klausurrelevant), andernfalls wird die Lösung nicht akzeptiert!

Aufgabe 1: Korrelation zwischen Videoaufrufen und Likes

Ein YouTube-Analyse-Team hat eine Stichprobe von 15 zufällig ausgewählten Gaming-Videos aus den deutschen Trends analysiert. Sie möchten herausfinden, ob es einen Zusammenhang zwischen der Anzahl der Videoaufrufe und der Anzahl der Likes gibt. Die Ergebnisse ihrer Analyse sind in Tabelle 1 dargestellt:

(a) Was sind die statistischen Einheiten dieser Untersuchung und welche Merkmale wurden an ihnen gemessen? Gebt das Skalenniveau der Merkmale an.

(b) Zeichnet ein Streudiagramm der Daten und interpretiert das resultierende Bild. Welche Art von Beziehung scheint zwischen den beiden Variablen zu bestehen?

(c) Berechnet den Pearson-Korrelationskoeffizienten zwischen den Videoaufrufen und Likes. Was sagt dieser Wert über die Stärke und Richtung der Beziehung zwischen den beiden Variablen aus?

(d) Auf Basis eures Ergebnisses: Wenn ein Video eine außergewöhnlich hohe Anzahl an Aufrufen hat, erwartet ihr dann auch eine außergewöhnlich hohe Anzahl an Likes? Begründet eure Antwort.

Video	Videoaufrufe (in Tsd.)	Likes (in Tsd.)
1	734	25
2	609	25
3	679	13
4	242	2
5	885	39
6	813	40
7	757	24
8	409	47
9	59	25
10	773	18
11	327	38
12	804	26
13	854	14
14	649	9
15	120	10

Tabelle 1: Daten für zufällig ausgewählte Gaming-Videos

Aufgabe 2: Einfluss der Hintergrundmusik auf die Zuschauerzufriedenheit

YouTube möchte untersuchen, inwieweit die Art der Hintergrundmusik in einem Video die Zufriedenheit der Zuschauer beeinflusst. In einer Pilotstudie wurden 9 Zuschauer gebeten, ein bestimmtes Video anzusehen. Das Video wurde mit drei verschiedenen Hintergrundmusikarten gezeigt: drei Zuschauern mit klassischer Musik, drei mit Popmusik und drei mit Ambient-Klängen. Den Zuschauern wurde nicht mitgeteilt, dass es Unterschiede in der Hintergrundmusik gibt, sondern sie wurden lediglich gebeten, ihre Zufriedenheit mit dem Video anhand einer 5-Punkte-Skala zu bewerten:

5-Punkte-Skala:

1. sehr unzufrieden
2. eher unzufrieden
3. neutral
4. zufrieden
5. sehr zufrieden

Die Bewertungen sind gemeinsam mit der Art der Hintergrundmusik in Tabelle 2 wiedergegeben:

Zuschauer Nr.	Hintergrundmusik	Zufriedenheitsbewertung
1	*Ambient*	1
2	*Ambient*	2
3	*Ambient*	2
4	*Pop*	2
5	*Pop*	2
6	*Pop*	2
7	*Rock*	3
8	*Rock*	1
9	*Rock*	5

Tabelle 2: Bewertungen und Art der Hintergrundmusik

Für die Zwecke dieser Analyse ordnet ihr bitte die Musikarten in Bezug auf ihre Intensität wie folgt:

1. Rang: *Ambient* (am wenigsten intensiv)
2. Rang: *Pop* (mittel)
3. Rang: *Rock* (am intensivsten)

(a) Berechnet den Rangkorrelationskoeffizienten nach Spearman (einfache Formel) und interpretiert das Ergebnis.

(b) Welche potenziellen Probleme könnten beim Verwenden der Formel in Bezug auf diese Daten auftreten?

(c) Wie könnten solche Probleme in einer umfangreicheren oder komplexeren Analyse angegangen oder vermieden werden?

Aufgabe 3: Stochastik

Ihr seid die Datenanalysten eines großen Werbeunternehmens, das mit YouTube-Influencern zusammenarbeitet. Ein Unternehmen möchte wissen, wie oft bestimmte Influencer in den YouTube-Empfehlungen auftauchen. Ihr habt den Auftrag, dies anhand von 1.000 zufälligen YouTube-Nutzern zu erheben. Hierzu bittet ihr eure Mitarbeiter, zufällige Nutzer zu kontaktieren und nach ihren Empfehlungen zu fragen. Ein Mitarbeiter schafft es im Schnitt pro Stunde, 10 Nutzer zu befragen.

(a) Erläutert, welches Wahrscheinlichkeitsmodell die beschriebene Situation am besten abbildet. Welche Parameter benötigt ihr für dieses Modell und welchen Wert würdet ihr diesen Parametern hier zuweisen?

(b) Bestimmt unter Rückgriff auf das von euch gewählte Modell die Wahrscheinlichkeit, dass ein Mitarbeiter innerhalb einer Stunde genau 8 Nutzer befragt.

(c) Wie groß ist die Wahrscheinlichkeit, dass ein Mitarbeiter innerhalb einer Stunde höchstens 8 Nutzer befragt?

(d) Wie groß ist die Wahrscheinlichkeit, dass ein Mitarbeiter innerhalb einer Stunde mindestens 8 Nutzer befragt?

Tipps und Anregungen:

- Handelt es sich um einen diskreten oder einen kontinuierlichen Prozess?
- Zählt ihr die Anzahl der Ereignisse in einem bestimmten Zeit- oder Raumintervall, oder messt ihr die Zeit oder den Raum bis zum ersten Auftreten eines Ereignisses?
- Sind die Ereignisse unabhängig und treten sie mit einer konstanten durchschnittlichen Rate auf?

Aufgabe 4: Korrelationsanalyse der YouTube Daten mit Python

Siehe Jupyter Notebook `02_Korrelation.ipynb`.