Humboldt-Universität zu Berlin
Institut für Informatik
Prof. Dr. Ulf Leser
Arik Ermshaus, Hermann Stolte

Berlin, den 11.06.2024

Übungen zur Vorlesung
Statistik und Data Science für die Informatik

Übungsblatt 4

Abgabe: Bis Montag, den 24.06.2024, bis 23:59 Uhr über Moodle. Die Übungsblätter sind in Gruppen von drei (in Ausnahmefällen zwei) Studierenden zu bearbeiten. Die Lösungen sind, sofern nicht anders angegeben, auf nach Aufgaben getrennten **PDFs** über Moodle abzugeben. Alle Teilaufgaben einer Aufgabe sind in einem PDF hochzuladen. Pro Gruppe genügt es, wenn eine Person die Lösung der Gruppe abgibt. Zur Bewertung wird der zuletzt hochgeladene Stand herangezogen. Vermerken Sie auf allen Abgaben Ihre **Namen**, Ihre **CMS-Benutzernamen** und Ihre **Abgabegruppe (z.B. Gruppe 123)** aus Moodle. Benennen Sie die hochgeladenen PDF-Dateien nach dem Schema: A<Aufgabe>-<Person1>-<Person2>- <Person3>.pdf, bspw. A3-Musterfrau-Mustermann-Beispiel.pdf für Aufgabe 3 von Lisa Musterfrau, Peter Mustermann und Karla Beispiel. Die Auflistung der Namen kann in beliebiger Reihenfolge erfolgen. Beachten Sie die Informationen im Moodle-Kurs https://hu.berlin/sds24.

Aufgabe 1 (Diskrete Verteilungen) — 2 + 2 + 3 + 3 = 10 Punkte

In dieser Aufgabe sollen Sie die Wahrscheinlichkeiten für verschiedene Szenarien mithilfe diskreter Verteilungen berechnen. Geben Sie für jede Teilaufgabe Ihren Rechenweg an und benennen Sie die benutzte Verteilung.

(a) Eine Fabrik produziert Kugellager, von denen $n = 500$ in einem Lager liegen. Hiervon sind $d = 20$ defekt. Ein Qualitätskontrolleur zieht ohne Zurücklegen eine Stichprobe von $m = 50$ Kugellagern. Welcher Verteilung unterliegt die Anzahl dabei gefundener defekter Kugellager? Wie groß ist die Wahrscheinlichkeit, dass in seiner Stichprobe höchstens drei defekte Kugellager enthalten sind?

(b) Ein Ornithologe muss an einem bestimmten Spot maximal 84 Vögel beobachten, um eine seltene Vogelart mit einer Wahrscheinlichkeit von 68% mindestens einmal zu sehen. Wie heißt die Verteilung, der die Anzahl an Vögeln unterliegt, die der Ornithologe beobachten muss, um einen seltenen Vogel zu sehen? Wie hoch ist die Wahrscheinlichkeit, dass ein zufällig beobachteter Vogel von der seltenen Art ist?

(c) In einer klinischen Studie wird ein neues Medikament getestet, das bei 75% der Patienten wirksam ist. Welcher Verteilung unterliegt die Anzal an Patienten, bei denen das Medikament wirkt? Wie groß ist die Wahrscheinlichkeit, dass von 10 behandelten Patienten mindestens 5 eine positive Wirkung zeigen?

(d) An einer bestimmten Kreuzung in Berlin passieren pro Jahr im Durchschnitt 13 Fahrradunfälle. Wie heißt die Verteilung, der die jährliche Anzahl an Fahrradunfällen an der Kreuzung unterliegt? Wie groß ist die Wahrscheinlichkeit, dass in diesem Jahr zwischen 10 bis 15 Fahrradunfälle an der Kreuzung passieren?

Aufgabe 2 (Normalverteilung) **3 + 3 + 4 = 10 Punkte**

In einem Lebensmittelverarbeitungsbetrieb wird eine Produktionsanlage eingesetzt, um Kekse herzustellen. Um sicherzustellen, dass die Kekse den Anforderungen der Produktbeschreibung entsprechen, ist es wichtig, dass die Kekse ein bestimmtes Gewicht haben.

Die Produktionsanlage ist so konzipiert, dass sie Kekse produziert, deren Gewicht X normalverteilt ist $X \sim \mathcal{N}(\mu, \sigma^2)$. Das bedeutet, dass die meisten der produzierten Kekse ein Gewicht nahe dem Mittelwert von $\mu = 24$ Gramm aufweisen, aber es gibt auch einige, die etwas leichter oder schwerer sind.

Dies ist auf eine Reihe von Faktoren zurückzuführen, einschließlich kleiner Abweichungen in der Zusammensetzung der Grundzutaten, mechanischen Toleranzen in der Produktionsanlage selbst und anderen Faktoren im Verarbeitungsprozess. Die Standardabweichung dieser Gewichtsverteilung beträgt $\sigma = 2$ Gramm.

(a) Wie viel Prozent der Kekse sind leichter als 22g? Geben Sie Ihren Rechenweg an.

(b) Wie viel Prozent der Kekse sind schwerer als 30g? Geben Sie Ihren Rechenweg an.

(c) Wie groß müsste μ bei gleichem σ sein, damit nur 1% der Kekse leichter als 22g sind? Geben Sie Ihren Rechenweg an.

Aufgabe 3 (Zentraler Grenzwertsatz) **3 + 3 + 4 = 10 Punkte**

Analysieren Sie die beobachteten Verteilungen verschiedener Statistiken zu einer Menge von Stichproben durch Implementierungen in NumPy und Matplotlib. Nehmen Sie dabei Bezug auf die Übereinstimmung mit ihren Erwartungen gemäß des zentralen Grenzwertsatzes.

(a) Simulieren Sie das folgende Experiment: Sie bestimmen die Augensumme von zwei gleichzeitig geworfenen, fairen, sechsseitigen Würfeln für $n = 200$ Wiederholungen. Anschließend berechnen Sie den Sample Mean, also den Mittelwert der beobachteten Augensummen. Das gesamte Experiment wird $m = 10000$ mal wiederholt.

Visualisieren Sie die Häufigkeitsverteilung des Sample Means für alle Durchläufe des Experiments in einem Histogramm. Beschriften Sie die Achsen und zeichnen Sie ein zu den Achsen passendes Gitter (Grid) als Diagrammhintergrund ein. Markieren Sie den theoretischen Erwartungswert des Sample Means als vertikale, grau gestrichelte Linie. Beschreiben Sie kurz und stichhaltig die beobachtete Verteilung im Vergleich mit der zu erwartenden Verteilung des Sample Means laut des zentralen Grenzwertsatzes.

(b) Analysieren Sie die Gültigkeit des zentralen Grentzwertsatzes für verschiedene Datenverteilungen. Simulieren Sie dazu $m = 10000$ Wiederholungen mit je $n = 50$ zufälligen Samples aus (1) der hypergeometrischen Wahrscheinlichkeitsverteilung $h(x|49; 6; 6)$, die modelliert, x Richtige im Lotto-Spiel "6 aus 49" zu erzielen, und (2) der Poisson-Verteilung mit $\lambda = 3.17$, die die erwartete Anzahl von Toren je Spiel in der Fußball-Bundesliga beschreibt. Berechnen Sie für jede Wiederholung die Summe und den Mittelwert und visualisieren Sie deren absolute Häufigkeiten in einem Histogramm. Kommentieren Sie kurz und stichhaltig, ob die beobachteten Verteilungen mit der Erwartung nach dem zentralen Grenzwertsatz übereinstimmen.

(c) In Moodle finden Sie die Datei "pokemon.csv". Dieser Datensatz umfasst 721 Pokémon, einschließlich ihrer Nummer, ihres Namens, ihres ersten und zweiten Typs sowie grundlegender Eigenschaften: HP, Angriff, Verteidigung, Spezialangriff, Spezialverteidigung und Geschwindigkeit.[1]

Visualisieren Sie jeweils den Sample Mean der Attribute HP, Angriff und Geschwindigkeit für $m \in \{10, 100, 1000, 10000\}$ Stichproben von $n = 6$ zufällig ausgewählten Pokemon. Für jedes Attribut soll es eine Figure, und darin für jedes m eine Zeile mit einem Histogramm geben, das die relativen Häufigkeiten des z-transformierten Sample Means darstellt. Zeigen Sie zusätzlich die Standardnormalverteilung als Linie über jedem Histogramm. Kommentieren Sie kurz und stichhaltig in Bezug auf die Stichprobenanzahl m, wie die beobachteten Verteilungen mit der Erwartung nach dem zentralen Grenzwertsatz übereinstimmen.

Nutzen Sie zur Beantwortung der Fragen aus (a), (b) und (c) das in Moodle bereitgestellte Python-Template "aufgabe3.py". Im Template gibt es jeweils eine Funktion je Teilaufgabe, in welcher die Aufgabe programmatisch gelöst werden soll. Die erwarteten Datentypen der Rückgabe entnehmen Sie der Beschreibung zu Beginn jeder Funktion.

Verändern Sie **nicht** die Signatur der definierten Funktionen. Darüber hinaus können Sie beliebige Hilfsfunktionen definieren. Sie dürfen lediglich die Open-Source-Bibliotheken NumPy, Pandas und Matplotlib nutzen.

[1] Datensatz: https://www.kaggle.com/datasets/abcsds/pokemon