

# WORKBOOK ZUR ONLINE-STATISTIK-VERANSTALTUNG IM WINTERSEMESTER 2025: AUFGABENBESCHREIBUNG

## VORBEMERKUNGEN

Das Workbook Statistik dient zur Einübung und Vertiefung der statistischen Methoden, die im Kurs erarbeitet wurden. Am besten bearbeitest du jeden Aufgabenblock zügig, nachdem der entsprechende Inhalt in Vorlesung und Übung besprochen wurden. Lade das Workbook bis zur Abgabefrist als einzelnes **PDF-Dokument auf myCampus/Turnitin** hoch. Sende außerdem **entweder** eine Excel-Datei, die all deine im Workbook verwendeten Excel-Berechnungen und Excel-Visualisierungen enthält, oder einige **Python-Dateien** per E-Mail an [Oliver.Labs@iu.org](mailto:Oliver.Labs@iu.org). Die Bewertung erfolgt auf Basis des Workbooks und dies soll OHNE die Excel-Datei bzw. die Python-Dateien verständlich sein; die Excel-Datei bzw. Python-Dateien wird/werden nur in Zweifelsfällen zu Rate gezogen.

**Hinweise:** Alle Aufgabenstellungen beziehen sich auf einen Auszug der sozialwissenschaftlichen Erhebung **“ALLBUS”**. Den Datensatz (*DS-Statistik2025WiSe\_Workbook\_ALLBUS\_ZA5284\_version-1-1-0\_Teil Datensatz-aus-2021-nur-29-Variablen.xlsx*) sowie eine Beschreibung des Datensatzes (*DS-Statistik2025WiSe\_Workbook\_Datensatz-Beschreibung\_OL\_2025-11-01.docx*) findest du im myCampus-Kurs im „Workbook“-Verzeichnis. Die Aufgaben sind mithilfe von **Microsoft Excel** oder **Python** zu bearbeiten (außer beim Mosaikdiagramm und Post-Hoc-Test, wo Excel keine gute Möglichkeit bietet).

## ALLGEMEINES ZUR BEARBEITUNG ALLER AUFGABEN

- In den Aufgaben sollst du die dir zugeteilten Merkmale verwenden. **Die dir zugeteilten Merkmale** erhältst du, indem du in der Datei *DS-Statistik2025WiSe\_Workbook\_Zuteilung\_OL\_2025Q4.xlsx* (ebenfalls im „Workbook“-Verzeichnis auf myCampus liegend) die dortige Matrikelnummer 123456789 in *deine Matrikelnummer* änderst und mit RETURN bestätigst. In den folgenden Zeilen erscheint dann die von dir zu verwendende Variablen-Kombination.
- Eine Beschreibung des Vorgehens in Excel bzw. Python bitte kurzfassen und über geeignete Screenshots ergänzen; wichtiger sind: **klare Darstellung und Interpretation der Herangehensweise** (jeweils z.B.: **welches Verfahren, warum macht es hier Sinn?**) und der **Ergebnisse. Ordne auch alle Rechenergebnisse inhaltlich ein** (mit Bezug auf die zugewiesenen Variablen; vergleiche Deine Rechnungen gegebenenfalls auch mit weiteren Rechnungen, die zur Einschätzung helfen).

## FORMALE KRITERIEN

- Dein abgeschlossenes Workbook soll aus einem sinnvoll strukturierten, durchgehenden Text bestehen. Halte dich bei der Erstellung deines Workbooks an die **„Richtlinien für die Gestaltung wissenschaftlicher Arbeiten“** der IU. (Die Richtlinien findest du im myCampus-Kurs.) Verwende

wissenschaftliche Fachsprache und stelle sicher, dass dein Workbook frei von Rechtschreib-, Zeichensetzungs- und Grammatikfehlern ist.

- Beachte, dass dein Workbook ein **Quellenverzeichnis** enthalten muss; wenn du keine wissenschaftlichen Bücher oder Artikel verwendest, muss wenigstens auf die Vorlesungsunterlagen verwiesen werden. Gib bei den **einzelnen Aufgaben** an, auf welcher Quelle Deine Bearbeitung beruht, möglichst **mit Seiten/Folien-Angabe** oder wenigstens einer Kapitelangabe.
- Wichtig ist, dass du jede durchgeführte Analyse sinnvoll **theoretisch untermauert**, den zugehörigen **Output-Report** der jeweiligen Software eingefügt und diesen ausführlich **interpretiert** hast. Erläutere bei allen Software-Aufgaben auch knapp, **wie** du **in der Software vorgegangen** bist (inklusive sinnvoller Screenshots).
- Verwende für den **Hauptteil zwischen 10 und 20 Seiten** (grobe Anhaltspunkte siehe die Aufgaben; die dortigen %-Zahlen geben den Anteil an der Gesamtpunktzahl an).

**ACHTUNG: Eine Nicht-Einhaltung der formalen Kriterien kann zu Punktabzügen in den einzelnen Aufgabenblöcken führen.**

## **AUFGABENBLOCK 1/5: UNIVARIATE DESKRIPTIVE STATISTIK (20%, CA. 2-5 SEITEN)**

- a) Beschreibe mit eigenen Worten kurz den **gesamten für das Workbook in Form der Excel-Tabelle zur Verfügung gestellten Datensatz, d.h. alle dortigen Variablen**, nicht nur die Dir zugeteilten. Gehe dabei wenigstens auf folgende Punkte ein (ergänzt durch eine knappe Einleitung zur Herkunft der Daten und dem Datensatz):
- Was ist der **Stichprobenumfang** des Workbook-Datensatzes?
  - Wie häufig sind **fehlende Angaben** und wie planst du, in diesem Workbook mit diesen umzugehen (Eliminierungsverfahren, Imputationsverfahren)?
  - Welche **Merkmale** haben **welches Skalenniveau** (nur sehr knapp erläutern warum)?
  - Welche Variablen wurden **Dir für die einzelnen Aufgabenblöcke 1-5 zugeteilt**? Füge auch einen Screenshot der Zuteilungs-Excel-Tabelle ein. (5%)
- b) Für diese Aufgabe (Teil 1b) bis 1d)) wurden dir drei Variablen zugeteilt: **eine nominal skalierte Variable, eine ordinal skalierte Variable und eine metrische Variable**. Erläutere anhand der ausgewählten Variablen die Bedeutung von **Skalenniveaus** ausführlich. (5%)
- c) Erstelle eine Tabelle für das *ordinal skalierte* Merkmal, welche die **absoluten und relativen Häufigkeiten** auflistet. Erstelle ein entsprechendes **Säulen- / Balken- oder Kreisdiagramm** und erläutere knapp, welches warum geeignet ist. Berechne den **Modus** und den **Median** und erläutere deine Ergebnisse. (5%)
- d) Wähle eine sinnvolle **Klasseneinteilung** für das *metrisch skalierte* Merkmal, erläutere diese und liste die absoluten und relativen Häufigkeiten in einer Tabelle auf. Erstelle ein entsprechendes **Histogramm** und erläutere an deiner Variablen, wann die Verwendung der Häufigkeitsdichten auf der y-Achse sehr wichtig ist und wann weniger. Berechne das **arithmetische Mittel** und die **Standardabweichung** und interpretiere all deine Ergebnisse und Visualisierungen. (5%)

## AUFGABENBLOCK 2/5: BIVARIATE DESKRIPTIVE STATISTIK (20%, 2-5 SEITEN)

- a) Für diese Aufgabe sind dir **zwei Variablen** zugeteilt worden. Stelle deren gemeinsame Häufigkeitsverteilung anhand einer **Kontingenztafel** dar und erstelle ein **Mosaikdiagramm** (bei zu vielen Merkmalsausprägungen, bilde zunächst geeignete Gruppen). Erläutere, *welches* der Zusammenhangsmaße „Phi-Koeffizient“, „Cramers V“, „Rangkorrelation von Spearman“, „Korrelation von Pearson“ für Deine Variablen geeignet berechnet werden kann. **Berechne dieses Zusammenhangsmaß** und interpretiere das Ergebnis. (10%)
- b) Für diese Aufgabe sind dir **zwei metrische Variablen** zugewiesen worden. Erstelle ein **Streudiagramm** der Variablen. Erläutere am Beispiel der zwei Variablen, wie der **Korrelationskoeffizient nach Bravais-Pearson** berechnet wird und was dieser aussagt. Berechne den Korrelationskoeffizienten, vergleiche dessen Ergebnis mit dem Streudiagramm und interpretiere deine Ergebnisse. (5%)
- c) Erkläre - *optional* am Beispiel der beiden dir zugeteilten Variablen aus Teil b) - den Unterschied zwischen den Begriffen „**Korrelation**“ und „**Kausalität**“. (5%)

## AUFGABENBLOCK 3/5: HYPOTHESENTESTS (20%, 2-4 SEITEN)

- a) Erläutere am Beispiel der ALLBUS-Umfrage den Unterschied zwischen **Grundgesamtheit** und **Stichprobe**. (5%)
- b) Für diese Aufgabe ist dir **eine metrische Variable** zugewiesen worden. Berechne das **symmetrische zweiseitige 95%-Konfidenzintervall** um deren Mittelwert und erläutere dein Ergebnis. (5%)
- c) Teile den Datensatz in zwei Stichproben vergleichbarer Größe, die zwei Teilgesamtheiten deiner Wahl repräsentieren, z. B. männliche und nicht-männliche Befragte (verwende hierfür eine, von Dir selbst ausgewählte, Variable, z.B. das Geschlecht, aber gerne auch eine andere geeignete). Wende einen **t-Test für zwei unabhängige Stichproben** auf einem Konfidenzniveau von 95% an, um für deine metrische Variable die Mittelwerte in den Teilgesamtheiten zu vergleichen. Erläutere knapp, warum die Stichproben unabhängig sind. Entscheide bei dem t-Test selbst, ob du einen einseitigen oder zweiseitigen Test durchführen möchtest, formuliere zunächst die Null- und Alternativhypothese, führe die Berechnung durch und interpretiere zuletzt deine Ergebnisse. (10%)

### AUFGABENBLOCK 4/5: ZUSAMMENHANGSANALYSE (20%, 2-3 SEITEN)

- a) Für diese Aufgabe sind dir **zwei metrische Variablen** aus dem Datensatz zugewiesen worden. Berechne dafür den **Korrelationskoeffizient** nach Bravais-Pearson und führe einen **Signifikanztest** durch. Erläutere, wie immer, kurz, wie du dabei mit der Software vorgegangen bist und interpretiere deine Ergebnisse. **(5%)**
- b) Nimm an, dass ein kausaler Zusammenhang zwischen deinen beiden Variablen besteht. Triff dabei eine Annahme, welche der Variablen die Ursache und welche die Wirkung darstellt (auch, wenn es nicht so viel Sinn macht) und erlautere diese knapp. Erstelle ein **Streudiagramm** beider Variablen, das auch die **Regressionsgerade** zeigt. **(5%)**
- c) Führe dann eine **einfache lineare Regression** durch, um die Beziehung zwischen den beiden Variablen zu modellieren. Schreibe die Gleichung der Ausgleichsgeraden konkret hin (Steigung und y-Achsenabschnitt). Berechne das **Bestimmtheitsmaß  $R^2$**  und führe einen **Signifikanztest** durch. Wähle außerdem einen geeigneten Beispiel-x-Wert, für den Du die lineare Regression für eine **Vorhersage** eines y-Wertes nutzt. Interpretiere das Ergebnis, auch im Hinblick auf Deine vorige Wahl der Ursache-Variable und der Wirkung-Variable und im Hinblick auf den Vergleich von Streudiagramm und Rechnung. **(10%)**

### AUFGABENBLOCK 5/5: ANOVA (20%, 2-3 Seiten)

- a) Für diesen Aufgabenblock hast Du **eine ordinale und eine metrische Variable** zugeteilt bekommen. Teile die Zeilen so in Gruppen auf (z.B. durch Sortieren des Datensatzes), dass innerhalb jeder Gruppe jeweils das Dir zugeteilte ordinale Merkmal identisch ist. Wie viele Gruppen sind es und wie groß sind die einzelnen Gruppen? **(5%)**
- b) Erläutere, warum die Gruppen unabhängig sind, so dass damit eine ANOVA durchgeführt werden kann, um zu prüfen, ob sich Mittelwerte der *metrischen* Variable der einzelnen Gruppen in der Grundgesamtheit unterscheiden? **(5%)**
- c) Führe die in b) erwähnte ein-faktorielle ANOVA durch, und zwar für das Signifikanzniveau  $\alpha = 0,20$ . Führe, auch wenn es für dieses Signifikanzniveau KEINEN Sinn macht, danach einen Post-Hoc-Test durch. Interpretiere Deine Ergebnisse; vergleiche dazu mit den Ergebnissen für  $\alpha = 0,05$  und  $\alpha = 0,40$ . **(10%)**