IU Internationale Hochschule

B. Sc. Data Science

Workbook Statistik

Statistische Analyse des ALLBUS 2021 Teildatensatzes

Autor:	Vorname Nachname
Matrikelnummer:	Matrikelnummer
Anschrift:	Straße Hausnr., PLZ Ort
Betreuung:	Oliver Labs
Abgabedatum:	TT.MM.JJJJ

Erklärung / Sperrvermerk

Hier ggf. die Eigenständigkeits- und Sperrvermerkserklärung gemäß Vorgaben der Hochschule.

Danksagung

Optionaler Text für Danksagungen.

Abstract (Deutsch)

Kurzfassung der Arbeit (ca. 200 Wörter): Problemstellung, Methode, Ergebnisse, Implikationen.

Abstract (English)

Abstract (approx. 200 words): problem, method, results, implications.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ALLBUS Allgemeine Bevölkerungsumfrage der Sozialwissenschaften
ANOVA Analysis of Variance (Varianzanalyse)
CI Confidence Interval (Konfidenzintervall)
SD Standard Deviation (Standardabweichung)

1 Einleitung

Dieses Workbook dokumentiert die statistische Analyse des ALLBUS 2021 Teildatensatzes im Rahmen des Moduls „Deskriptive und Inferenzstatistikän der IU Internationale Hochschule im Wintersemester 2025.

Der ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) ist eine repräsentative Querschnittserhebung, die seit 1980 Einstellungen, Verhaltensweisen und Sozialstruktur der Bevölkerung in Deutschland erhebt. Der vorliegende Teildatensatz aus dem Jahr 2021 umfasst 29 ausgewählte Variablen.

1.1 Zielsetzung

Ziel dieser Arbeit ist die systematische Anwendung deskriptiver und inferenzstatistischer Methoden auf reale Umfragedaten. Im Einzelnen werden folgende Analysen durchgeführt:

1. **Univariate Deskription:** Analyse einzelner Variablen nach Skalenniveau
2. **Bivariate Deskription:** Untersuchung von Zusammenhängen zwischen zwei Variablen
3. **Inferenzstatistik:** Konfidenzintervalle und Hypothesentests
4. **Regression:** Modellierung linearer Zusammenhänge
5. **ANOVA:** Vergleich von Mittelwerten über mehrere Gruppen

1.2 Aufbau der Arbeit

Nach dieser Einleitung folgt in Kapitel 2 eine Beschreibung des Datensatzes und der verwendeten Variablen. Die Kapitel 3 bis 7 dokumentieren die fünf Analyseblöcke gemäß Aufgabenstellung. Kapitel 8 diskutiert die Ergebnisse kritisch, bevor Kapitel 9 die Arbeit zusammenfasst.

Alle statistischen Analysen wurden in Python durchgeführt (Version 3.12.6). Die verwendeten Pakete umfassen `pandas`, `numpy`, `matplotlib`, `scipy` und `statsmodels`.

2 Datenbeschreibung

2.1 Der ALLBUS 2021 Teildatensatz

Der verwendete Datensatz ist ein Auszug aus dem ALLBUS 2021 (ZA5284, Version 1.1.0) und enthält 29 Variablen aus verschiedenen Themenbereichen:

- Soziodemografische Merkmale (Alter, Geschlecht, Bildung, Einkommen)
- Politische und wirtschaftliche Einschätzungen
- Werthaltungen und Einstellungen
- Gesundheit und Lebenssituation

Die Stichprobe umfasst `N = [ANZAHL]` Befragte aus der deutschsprachigen Wohnbevölkerung ab 18 Jahren.

2.2 Verwendete Variablen

Gemäß der individuellen Variablenzuteilung werden in dieser Arbeit folgende Variablen analysiert:

2.2.1 Block 1: Univariate Deskription

- **mc04** (nominal): Ausländer im Freundeskreis
- **ep04** (ordinal): Wirtschaftslage in Deutschland in 1 Jahr
- **hhinc** (metrisch): Haushaltsnettoeinkommen in Euro

2.2.2 Block 2: Bivariate Deskription

- **ep01** (ordinal): Aktuelle Wirtschaftslage in Deutschland
- **fe14** (ordinal): Erziehungsziel „beliebt sein”
- **xt10** (metrisch): Interviewdauer in Minuten
- **age** (metrisch): Alter in Jahren

2.2.3 Block 3: Inferenzstatistik

- **lm02** (metrisch): Tägliche Fernsehdauer in Minuten
- **sex** (nominal): Geschlecht (für Gruppenvergleich)

2.2.4 Block 4: Regression

- **age** (metrisch): Alter in Jahren (Prädiktor)
- **hhinc** (metrisch): Haushaltsnettoeinkommen (Zielvariable)

2.2.5 Block 5: ANOVA

- **hs01** (ordinal): Gesundheitszustand (Gruppierungsfaktor)
- **gd02** (metrisch): Wohndauer im Ort in Jahren

2.3 Datenbereinigung

Vor den Analysen wurden folgende Bereinigungsschritte durchgeführt:

1. Behandlung fehlender Werte (Sondercodes wie -99, -98, -42 etc.)
2. Konvertierung von Datentypen (numerisch vs. kategorial)
3. Plausibilitätsprüfungen (Wertebereich, Ausreißer)
4. Speicherung als bereinigte CSV-Datei (`allbus_clean.csv`)

Details zur Datenbereinigung finden sich im Python-Skript `src/data_prep.py`.

3 Univariate Deskription und Skalenniveaus

3.1 Aufgabe 1b: Skalenniveaus

Die Wahl geeigneter statistischer Methoden hängt wesentlich vom Skalenniveau der Variablen ab. Man unterscheidet drei grundlegende Skalenniveaus:

3.1.1 Nominalskalenniveau

Definition: Variablen mit Nominalskalenniveau besitzen Kategorien ohne natürliche Ordnung. Es kann nur Gleichheit oder Verschiedenheit festgestellt werden.

Beispiel: Variable `mc04` – „Haben Sie Ausländer im Freundeskreis?"mit den Kategorien „Ja", „Nein", „Weiß nicht". Diese Kategorien haben keine sinnvolle Reihenfolge.

Zulässige Statistiken:

- Häufigkeiten (absolut und relativ)
- Modus (häufigste Kategorie)
- Balken- oder Kreisdiagramme

3.1.2 Ordinalskalenniveau

Definition: Variablen mit Ordinalskalenniveau besitzen geordnete Kategorien. Die Abstände zwischen den Kategorien sind jedoch nicht interpretierbar.

Beispiel: Variable `ep04` – „Wie wird sich die wirtschaftliche Lage in Deutschland in einem Jahr entwickeln?"mit Kategorien von „wesentlich besser"bis „wesentlich schlechter". Die Kategorien sind klar geordnet, aber der Abstand zwischen „besseründ „wesentlich besserïst nicht quantifizierbar.

Zulässige Statistiken:

- Alle Statistiken des Nominalniveaus
- Median (mittlere Kategorie)
- Perzentile
- Rangkorrelationen (Spearman, Kendall)

3.1.3 Metrisches Skalenniveau

Definition: Variablen mit metrischem Skalenniveau (Intervall- oder Verhältnisskala) besitzen gleichabständige Einheiten. Abstände und Verhältnisse sind interpretierbar.

Beispiel: Variable `hhinc` – Haushaltsnettoeinkommen in Euro. Die Differenz zwischen 2000€ und 3000€ ist genauso groß wie zwischen 4000€ und 5000€.

Zulässige Statistiken:

- Alle Statistiken der vorherigen Niveaus
- Mittelwert
- Standardabweichung, Varianz
- Pearson-Korrelation
- Regression, t-Tests, ANOVA

3.2 Aufgabe 1c: Univariate Analyse von `ep04` (ordinal)

3.2.1 Häufigkeitsverteilung

Interpretation: [TODO: Beschreibung der Verteilung]

3.2.2 Grafische Darstellung

Abb. 1: Häufigkeitsverteilung der Variable `ep04` – Erwartete Wirtschaftslage in 1 Jahr.

Eigene Darstellung auf Basis ALLBUS 2021.

3.2.3 Lagemaße

- **Modus:** [TODO: Häufigste Kategorie]
- **Median:** [TODO: Mittlere Kategorie]

Interpretation: [TODO: Interpretation der Lagemaße]

3.3 Aufgabe 1d: Univariate Analyse von `hhinc` (metrisch)

3.3.1 Klassenbildung

Für die Darstellung der Einkommensverteilung wurden folgende Klassen gebildet:

3.3.2 Histogramm

Abb. 2: Histogramm des Haushaltsnettoeinkommens (`hhinc`).

Eigene Darstellung auf Basis ALLBUS 2021.

Häufigkeitsdichte vs. absolute Häufigkeit: Das Histogramm zeigt die Häufigkeitsdichte, sodass die Fläche jedes Balkens proportional zur relativen Häufigkeit der entsprechenden Klasse ist. Dies ist bei unterschiedlich breiten Klassen wichtig für eine korrekte Interpretation.

3.3.3 Statistische Kennwerte

- **Mittelwert:** [TODO] Euro

- **Standardabweichung:** [TODO] Euro
- **Median:** [TODO] Euro
- **Minimum:** [TODO] Euro
- **Maximum:** [TODO] Euro

Interpretation: [TODO: Interpretation der Kennwerte, Schiefe der Verteilung, Vergleich Mittelwert vs. Median]

4 Bivariate Deskription

4.1 Aufgabe 2a: Zusammenhang `ep01` × `fe14` (ordinal)

4.1.1 Kreuztabelle

Interpretation: [TODO: Beschreibung der gemeinsamen Verteilung]

4.1.2 Grafische Darstellung

Abb. 3: Mosaikdiagramm: Zusammenhang zwischen aktueller Wirtschaftseinschätzung (`ep01`) und Erziehungsziel „beliebt sein"(`fe14`).

Eigene Darstellung auf Basis ALLBUS 2021.

4.1.3 Zusammenhangsmaß

Für den Zusammenhang zweier ordinaler Variablen eignet sich die Spearman-Rangkorrelation:

- **Spearman-Korrelation:** $r_s =$ [TODO]
- **p-Wert:** [TODO]

Interpretation: [TODO: Stärke und Richtung des Zusammenhangs, Signifikanz]

4.2 Aufgabe 2b: Korrelation `xt10` und `age` (metrisch)

4.2.1 Streudiagramm

Abb. 4: Streudiagramm: Zusammenhang zwischen Alter (`age`) und Interviewdauer (`xt10`).

Eigene Darstellung auf Basis ALLBUS 2021.

4.2.2 Pearson-Korrelation

- **Pearson-Korrelation:** $r =$ [TODO]
- **p-Wert:** [TODO]
- **Stichprobengröße:** $n =$ [TODO]

Interpretation: [TODO: Stärke und Richtung des linearen Zusammenhangs, statistische Signifikanz]

4.3 Aufgabe 2c: Korrelation ist nicht Kausalität

Die gefundene Korrelation zwischen Alter und Interviewdauer [TODO: positiv/negativ/nicht signifikant] bedeutet **nicht**, dass das Alter die Interviewdauer verursacht (oder umgekehrt).

4.3.1 Mögliche alternative Erklärungen

1. **Drittvariablen:** Eine nicht beobachtete Variable könnte beide beeinflussen. Beispielsweise könnte Bildungsniveau sowohl mit Alter als auch mit der Ausführlichkeit der Antworten korrelieren.
2. **Interviewer-Effekte:** Die Interviewdauer hängt stark vom Interviewer ab (Fragetempo, Gesprächsführung). Wenn bestimmte Interviewer bevorzugt ältere oder jüngere Personen befragen, entsteht eine Scheinkorrelation.
3. **Umgekehrte Kausalität:** Auch wenn eine kausale Beziehung bestünde, ist unklar, in welche Richtung sie verläuft.
4. **Zufälligkeit:** Bei einem Signifikanzniveau von 5% ist jede 20. Korrelation zufällig signifikant, selbst wenn kein echter Zusammenhang besteht.

4.3.2 Bedingungen für Kausalität

Um eine kausale Aussage zu rechtfertigen, wären erforderlich:

- **Zeitliche Reihenfolge:** Die Ursache muss der Wirkung vorausgehen
- **Ausschluss von Drittvariablen:** Kontrolle konfundierender Faktoren
- **Experimentelles Design:** Randomisierte Kontrollstudien (hier nicht möglich)
- **Theoretische Plausibilität:** Mechanismus muss erklärbar sein

Fazit: Beobachtete Korrelationen in Querschnittsdaten sind wichtige Hinweise auf Zusammenhänge, erlauben aber *per se* keine kausalen Schlussfolgerungen. Weiterführende Analysen (z. B. Längsschnittstudien, Experimente) wären nötig, um Kausalität zu belegen.

5 Inferenzstatistik mit `lm02`

5.1 Aufgabe 3a: Grundgesamtheit vs. Stichprobe

5.1.1 Grundgesamtheit (Population)

Die **Grundgesamtheit** im Kontext des ALLBUS 2021 umfasst die deutschsprachige Wohnbevölkerung in Privathaushalten ab 18 Jahren in Deutschland. Dies entspricht etwa 70 Millionen Personen.

5.1.2 Stichprobe

Die **Stichprobe** besteht aus den tatsächlich befragten [TODO: N] Personen, die nach einem Zufallsverfahren ausgewählt wurden. Diese Stichprobe soll repräsentativ für die Grundgesamtheit sein.

5.1.3 Inferenzstatistischer Ansatz

Die Inferenzstatistik erlaubt es, von den Stichprobendaten auf die Grundgesamtheit zu schließen:

- **Punktschätzung:** Der Stichprobenmittelwert $\bar{x}$ schätzt den Populationsmittelwert μ
- **Intervallschätzung:** Konfidenzintervalle geben einen Bereich an, in dem μ mit bestimmter Wahrscheinlichkeit liegt
- **Hypothesentests:** Prüfen von Vermutungen über die Population

5.2 Aufgabe 3b: Konfidenzintervall für `lm02`

Variable `lm02` erfasst die tägliche Fernsehdauer in Minuten.

5.2.1 Deskriptive Statistik

- **Stichprobenmittelwert:** $\bar{x} =$ [TODO] Minuten
- **Standardabweichung:** $s =$ [TODO] Minuten
- **Stichprobengröße:** $n =$ [TODO]

5.2.2 95%-Konfidenzintervall

Das Konfidenzintervall wurde mit der t-Verteilung berechnet (da σ unbekannt):

$$\text{KI}_{95\%} = \bar{x} \pm t_{n-1;0,975} \cdot \frac{s}{\sqrt{n}} \tag{1}$$

Ergebnis: [TODO: untere Grenze] $< \mu <$ [TODO: obere Grenze] Minuten

5.2.3 Interpretation

Das 95%-Konfidenzintervall bedeutet: Wenn wir das Stichprobenverfahren unendlich oft wiederholen würden, lägen in 95% der Fälle die so konstruierten Intervalle den wahren Populationsmittelwert μ enthalten.

Wichtig: Es bedeutet *nicht*, dass der wahre Wert mit 95% Wahrscheinlichkeit in diesem spezifischen Intervall liegt (der wahre Wert ist fix, nur unbekannt).

5.3 Aufgabe 3c: t-Test für zwei unabhängige Gruppen

5.3.1 Fragestellung

Unterscheidet sich die durchschnittliche Fernsehdauer (`lm02`) zwischen Männern und Frauen (`sex`)?

5.3.2 Hypothesen

- H_0: $\mu_{\text{männlich}} = \mu_{\text{weiblich}}$ (kein Unterschied in der mittleren Fernsehdauer)
- H_1: $\mu_{\text{männlich}} \neq \mu_{\text{weiblich}}$ (Unterschied besteht, zweiseitig)

Signifikanzniveau: $\alpha = 0{,}05$

5.3.3 Deskriptive Statistik nach Gruppen

5.3.4 Testergebnis

- **t-Statistik:** $t =$ [TODO]
- **Freiheitsgrade:** $df =$ [TODO]
- **p-Wert:** $p =$ [TODO]

5.3.5 Interpretation

[TODO: Entscheidung über H□]

Bei $p < 0{,}05$: Wir verwerfen H_0 und schließen, dass ein statistisch signifikanter Unterschied in der mittleren Fernsehdauer zwischen den Geschlechtern besteht.

Bei $p \geq 0{,}05$: Wir können H_0 nicht verwerfen. Die Daten liefern keine ausreichende Evidenz für einen Unterschied.

Praktische Relevanz: [TODO: Größe des Unterschieds bewerten]

6 Korrelation und Regression: `age` und `hhinc`

6.1 Aufgabe 4a: Pearson-Korrelation und Signifikanztest

6.1.1 Fragestellung

Besteht ein linearer Zusammenhang zwischen Alter (`age`) und Haushaltsnettoeinkommen (`hhinc`)?

6.1.2 Korrelationskoeffizient

- **Pearson-Korrelation:** $r =$ [TODO]
- **p-Wert:** $p =$ [TODO]
- **Stichprobengröße:** $n =$ [TODO]

6.1.3 Interpretation

Stärke: [TODO: schwach/moderat/stark basierend auf $|r|$]

Richtung: [TODO: positiv/negativ]

Signifikanz: Bei $\alpha = 0{,}05$ ist die Korrelation [TODO: signifikant/nicht signifikant].

6.2 Aufgabe 4b: Streudiagramm

Abb. 5: Streudiagramm: Zusammenhang zwischen Alter (`age`) und Haushaltsnettoeinkommen (`hhinc`).

Eigene Darstellung auf Basis ALLBUS 2021.

Visuelle Beurteilung: [TODO: Linearität, Ausreißer, Streuung beschreiben]

6.3 Aufgabe 4c: Einfache lineare Regression

6.3.1 Modellspezifikation

Wir schätzen das lineare Modell:

$$\text{hhinc}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \varepsilon_i \tag{2}$$

wobei:

- β_0: Achsenabschnitt (erwartetes Einkommen bei Alter = 0)
- β_1: Steigung (Änderung des Einkommens pro Jahr Alter)
- ε_i: Fehlerterm

6.3.2 Schätzergebnisse

- **Intercept** ($\hat{\beta}_0$): [TODO] Euro
- **Steigung** ($\hat{\beta}_1$): [TODO] Euro pro Jahr
- **R²:** [TODO]
- **Adjustiertes R²:** [TODO]
- **F-Statistik:** $F =$ [TODO], $p =$ [TODO]

6.3.3 Interpretation der Koeffizienten

Intercept: [TODO: Interpretation – meist nicht sinnvoll, da Alter = 0 außerhalb des Datenbereichs]

Steigung: Pro zusätzlichem Lebensjahr ändert sich das erwartete Haushaltsnettoeinkommen um [TODO] Euro. [TODO: Vorzeichen und praktische Bedeutung diskutieren]

R²: Das Modell erklärt [TODO]% der Varianz im Haushaltsnettoeinkommen durch das Alter.

6.3.4 Grafische Darstellung mit Regressionsgerade

Abb. 6: Streudiagramm mit geschätzter Regressionsgerade.

Eigene Darstellung auf Basis ALLBUS 2021.

6.3.5 Beispielprognose

Für eine Person im Alter von 45 Jahren lautet die Prognose:

$$\widehat{\text{hhinc}}_{45} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 45 = \text{[TODO] Euro} \tag{3}$$

Einschränkung: Diese Prognose ist nur innerhalb des beobachteten Altersbereichs sinnvoll (keine Extrapolation).

6.3.6 Modellannahmen und Limitationen

Das einfache lineare Modell unterstellt:

- Lineare Beziehung zwischen Alter und Einkommen
- Homoskedastizität (konstante Fehlervarianz)
- Normalverteilte Residuen
- Keine Ausreißer mit starkem Einfluss

[TODO: Kurze Diskussion, ob diese Annahmen erfüllt sind oder Einschränkungen bestehen]

7 Einfaktorielle ANOVA: gd02 nach hs01

7.1 Aufgabe 5a: Gruppenbildung und deskriptive Statistik

7.1.1 Gruppierungsvariable

Variable hs01 erfasst den subjektiven Gesundheitszustand mit [TODO: Anzahl] Kategorien (z. B. „sehr gut", „gut", „zufriedenstellend", „weniger gut", „schlecht").

7.1.2 Zielvariable

Variable gd02 misst die Wohndauer im aktuellen Wohnort in Jahren. Personen mit Wohndauer „unter 1 Jahr"wurden als 0 Jahre kodiert.

7.1.3 Deskriptive Statistik nach Gruppen

Die Tabelle zeigt für jede Gesundheitsgruppe:

- Anzahl der Beobachtungen (n)
- Mittelwert der Wohndauer ($\bar{x}$)
- Standardabweichung (s)

Erste Eindrücke: [TODO: Unterschiede zwischen Gruppen beschreiben]

7.2 Aufgabe 5b: Unabhängigkeit der Gruppen

7.2.1 Voraussetzung für ANOVA

Die einfaktorielle ANOVA setzt voraus, dass die Beobachtungen in den Gruppen unabhängig sind (d. h. keine Person gehört zu mehreren Gruppen).

7.2.2 Argumentation

Im ALLBUS-Datensatz gilt:

- Jede Person wurde nur einmal befragt
- Der Gesundheitszustand (hs01) ist ein Personenmerkmal – jede Person fällt in genau eine Kategorie
- Es gibt keine offensichtlichen Cluster (z. B. Haushalte mit mehreren Befragten)
- Die Stichprobenziehung erfolgte unabhängig

Fazit: Die Unabhängigkeitsannahme ist plausibel erfüllt.

7.3 Aufgabe 5c: Einfaktorielle ANOVA bei $\alpha = 0{,}20$

7.3.1 Hypothesen

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (alle Gruppen haben dieselbe mittlere Wohndauer)
- H_1: Mindestens zwei Gruppenmittelwerte unterscheiden sich

Signifikanzniveau: $\alpha = 0{,}20$ (ungewöhnlich hoch, aber laut Aufgabenstellung vorgegeben)

7.3.2 ANOVA-Tabelle

- **F-Statistik:** $F =$ [TODO]
- **Freiheitsgrade:** $df_{\text{between}} =$ [TODO], $df_{\text{within}} =$ [TODO]
- **p-Wert:** $p =$ [TODO]

7.3.3 Interpretation

Bei $p < 0{,}20$**:** Wir verwerfen H_0 und schließen, dass sich die mittlere Wohndauer zwischen mindestens zwei Gesundheitsgruppen signifikant unterscheidet.

Bei $p \geq 0{,}20$**:** Die Daten liefern keine ausreichende Evidenz für Unterschiede.

7.3.4 Post-Hoc-Tests

Da die ANOVA nur feststellt, *dass* Unterschiede bestehen (nicht *welche*), werden Post-Hoc-Tests (z. B. Tukey HSD) durchgeführt:

Signifikante Paarvergleiche: [TODO: Welche Gruppen unterscheiden sich?]

7.4 Diskussion des Signifikanzniveaus

7.4.1 Einfluss von α auf die Ergebnisse

Das gewählte $\alpha = 0{,}20$ ist deutlich höher als das übliche 5%-Niveau:

- **Bei** $\alpha = 0{,}05$**:** [TODO: Wären die Ergebnisse noch signifikant?]
- **Bei** $\alpha = 0{,}20$**:** Höhere Wahrscheinlichkeit, H_0 zu verwerfen (auch bei kleinen Effekten)

7.4.2 Typ-I- vs. Typ-II-Fehler

- **Typ-I-Fehler** (α): Fälschliche Ablehnung von H_0 (falsch positiv)
- **Typ-II-Fehler** (β): Fälschliches Beibehalten von H_0 (falsch negativ)

Ein höheres α erhöht die Wahrscheinlichkeit eines Typ-I-Fehlers, reduziert aber gleichzeitig die Wahrscheinlichkeit eines Typ-II-Fehlers (höhere Power).

7.4.3 Praktische Implikationen

In explorativen Studien kann ein höheres α gerechtfertigt sein, um keine potenziell wichtigen Effekte zu übersehen. In konfirmatorischen Studien (z. B. klinische Tests) ist hingegen ein strengeres Niveau erforderlich.

Fazit: Die Wahl von $\alpha = 0{,}20$ führt dazu, dass [TODO: mehr/weniger] Gruppen- unterschiede als signifikant identifiziert werden. Dies sollte bei der Interpretation berücksichtigt werden.

8 Diskussion

8.1 Zusammenfassung der Hauptergebnisse

Dieses Workbook wendete deskriptive und inferenzstatistische Methoden auf den ALLBUS 2021 Teildatensatz an. Die wichtigsten Ergebnisse:

- **Block 1:** [TODO: Kernaussagen zur univariaten Analyse]
- **Block 2:** [TODO: Kernaussagen zu bivariaten Zusammenhängen]
- **Block 3:** [TODO: Kernaussagen zu Konfidenzintervall und t-Test]
- **Block 4:** [TODO: Kernaussagen zur Regression]
- **Block 5:** [TODO: Kernaussagen zur ANOVA]

8.2 Methodische Überlegungen

8.2.1 Stärken des Vorgehens

- Systematische Anwendung skalenniveaugerechter Methoden
- Verwendung etablierter statistischer Tests mit klaren Voraussetzungen
- Transparente Dokumentation aller Analyseschritte
- Reproduzierbarkeit durch Python-Skripte

8.2.2 Limitationen

- **Querschnittsdaten:** Keine kausalen Aussagen möglich (nur Korrelationen)
- **Fehlende Werte:** Bereinigung könnte systematische Verzerrungen einführen
- **Einfache Modelle:** Regression mit nur einem Prädiktor ignoriert konfundierende Variablen
- **Annahmenverletzungen:** Nicht alle Voraussetzungen (z. B. Normalverteilung) wurden formal getestet

8.3 Interpretation im Kontext

8.3.1 Plausibilität der Ergebnisse

[TODO: Sind die gefundenen Zusammenhänge inhaltlich plausibel? Decken sie sich mit Erwartungen oder bisheriger Forschung?]

8.3.2 Statistische vs. praktische Signifikanz

Ein statistisch signifikantes Ergebnis bedeutet nicht automatisch praktische Relevanz. Bei großen Stichproben (wie im ALLBUS) werden auch sehr kleine Effekte signifikant.

[TODO: Beispiel: Ist der gefundene Unterschied/Zusammenhang groß genug, um praktisch bedeutsam zu sein?]

8.4 Weiterführende Analysen

Aufbauend auf diesen Ergebnissen könnten folgende Analysen sinnvoll sein:

- **Multiple Regression:** Einbezug mehrerer Prädiktoren zur Kontrolle von Drittvariablen
- **Interaktionseffekte:** Prüfen, ob Zusammenhänge in verschiedenen Subgruppen unterschiedlich stark sind
- **Robustheitschecks:** Nicht-parametrische Tests bei Verletzung von Annahmen
- **Vergleich mit anderen ALLBUS-Wellen:** Zeitliche Entwicklung untersuchen

9 Fazit

9.1 Zentrale Erkenntnisse

Diese Arbeit demonstrierte die praktische Anwendung grundlegender statistischer Methoden auf reale Umfragedaten. Die wichtigsten Erkenntnisse:

1. **Skalenniveaus entscheiden über Methoden:** Die korrekte Identifikation von nominal, ordinal und metrisch ist essentiell für die Wahl geeigneter Analysen.
2. **Deskription vor Inferenz:** Eine gründliche deskriptive Analyse (Häufigkeiten, Verteilungen, Grafiken) ist die Grundlage für alle weiterführenden Tests.
3. **Korrelation ≠ Kausalität:** Beobachtete Zusammenhänge in Querschnittsdaten erlauben keine kausalen Schlüsse ohne zusätzliche theoretische und methodische Absicherung.
4. **Signifikanz vs. Relevanz:** Statistische Signifikanz und praktische Bedeutung sind zu unterscheiden – besonders bei großen Stichproben.
5. **Reproduzierbarkeit durch Code:** Die Verwendung von Python-Skripten ermöglicht transparente und nachvollziehbare Analysen.

9.2 Methodische Reflexion

Die Bearbeitung dieses Workbooks verdeutlichte:

- Die Bedeutung sorgfältiger Datenbereinigung (Umgang mit fehlenden Werten, Sondercodes)
- Die Notwendigkeit, Voraussetzungen statistischer Tests zu prüfen
- Den Wert grafischer Darstellungen für das Verständnis von Daten
- Die Grenzen einfacher statistischer Modelle bei komplexen sozialen Phänomenen

9.3 Ausblick

Weiterführende Analysen könnten:

- Multiple Prädiktoren in Regressionsmodellen berücksichtigen
- Interaktionseffekte zwischen Variablen untersuchen
- Zeitliche Trends durch Vergleich mit früheren ALLBUS-Wellen analysieren
- Fortgeschrittene Methoden (z. B. logistische Regression, Strukturgleichungsmodelle) anwenden

9.4 Abschließende Bemerkung

Der ALLBUS-Datensatz bietet eine wertvolle Ressource für empirische sozialwissenschaftliche Forschung. Die hier durchgeführten Analysen kratzen nur an der Oberfläche dessen, was mit diesen Daten möglich ist. Sie demonstrieren jedoch grundlegende statistische Kompetenzen, die für weiterführende Analysen unerlässlich sind.