

Applied Mathematical Modeling

Statistical Learning & Data Systems

Theory-driven Machine Learning,
High-Dimensional Statistics,
and Computational Analysis

Abdallah Abouabdallah

Applied Mathematics • Statistical Learning • Machine Learning Engineering

Core Focus Areas

High-Dimensional Data Analysis
Probabilistic Modeling & Inference
Optimization & Numerical Methods
Statistical Learning Theory
Large-Scale Data Systems

Mathematical Foundations for Real-World Systems

August 2024

Inhaltsverzeichnis

1	Einleitung	2
2	Datenverarbeitung	2
2.1	Datenbereinigung	2
2.2	Datenumformung	2
2.3	Datenvalidierung	2
3	Methode	3
3.1	Umgang mit Fehlenden Werten	3
3.2	Statistische Analyse der Hypothesen	3
4	Ergebnisse	4
4.1	Ruhepuls und BMI	4
4.2	Dominante Hand und Armstärke	5
4.3	Wasserverbrauch nach Geschlecht	6
4.4	Blinzelfrequenz während der Videoaufgabe	7
4.5	Blinzelfrequenz zwischen Männern und Frauen	8
4.6	Körpergröße und Luftanhaltezeit	9
4.7	Linkshänder und Luftanhaltezeit	10
5	Schlussfolgerung und Diskussion	11
6	Fazit	12

Zusammenfassung

Um die Gesundheitsdaten zu analysieren, hat diese Studie robuste statistische Methoden verwendet. Deutliche Befunde haben gezeigt, dass der dominante Arm stärker ist als der nicht-dominante und größere Menschen können ihre Luft länger halten. Der Wasserkonsum war zwischen Männern und Frauen nicht signifikant unterschiedlich; ebenso war die Blinzelfrequenz während eines Videoaufgaben nicht signifikant unterschiedlich. Daher waren die Ergebnisse für Linkshänder wegen einer niedrigen Stichprobe wenig repräsentativ. Insgesamt betonen diese Ergebnisse die Notwendigkeit ausreichend großer Stichproben und sorgfältiger Dateneingabe.

1 Einleitung

Die Initiative „FH Gesundheitscheck“ hat Daten sowohl von Studierenden als auch von einer externen Partei gesammelt. Ziel dieses Projekts ist es, diese Daten zu verarbeiten und zu analysieren, um Fragen mit statistischen Beweisen zu beantworten. Die Rohdaten enthalten mehrere Fehler, die zunächst durch die Anwendung strenger Datenvorverarbeitungsverfahren behoben werden. Anschließend wird eine tiefgehende statistische Analyse durchgeführt, um die verschiedenen von der Initiative gestellten Fragen zu beantworten.

2 Datenverarbeitung

Wie bereits erwähnt, sind die Rohdaten, die von der Initiative „FH Gesundheitscheck“ gesammelt wurden, mit verschiedenen Fehlern und Inkonsistenzen behaftet. Um eine genaue und zuverlässige statistische Analyse zu gewährleisten, ist die Datenverarbeitungsphase von entscheidender Bedeutung. Die folgenden Schritte werden zur Vorverarbeitung der Daten unternommen:

2.1 Datenbereinigung

- Umgang mit fehlenden und falschen Werten.
- Korrektur von Tippfehlern.
- Spalten ausrichten

2.2 Datenumformung

- Standardisierung von Einheiten und Formaten zur Konsistenzsicherung.
- Sicherstellung des numerischen Formats der numerischen Daten zur Analyse.
- Neuberechnung verschiedener Parameter.
- Normalisierung der Daten, um Vergleichbarkeit zu gewährleisten.

2.3 Datenvalidierung

- Kreuzüberprüfung von Dateneinträgen mit bekannten Standards.
- Identifizierung und Handhabung von Ausreißern, die die Analyseergebnisse verzerren könnten.

Diese Vorverarbeitungsschritte stellen sicher, dass der Datensatz robust ist und für die anschließende detaillierte statistische Analyse bereit ist, die zur Beantwortung der von der Initiative „FH Gesundheitscheck“ gestellten Forschungsfragen verwendet wird.

3 Methode

In diesem Abschnitt beschreiben wir die statistischen Methoden und Ansätze, die zur Analyse des 'gereinigten' Datensatzes verwendet wurden, der nach der Anwendung der Datenverarbeitungsverfahren auf die Rohdaten erhalten wurde.

3.1 Umgang mit Fehlenden Werten

Ansatz: Anstatt fehlende Werte durch Imputation zu ersetzen oder die Daten zu löschen, wurden die fehlenden Werte in ihrem ursprünglichen Zustand belassen. Um Verzerrungen zu vermeiden, wurden für die jeweilige Analyse nur vollständige Datensätze herangezogen. Durch diesen Ansatz wurde sichergestellt, dass die Ergebnisse auf realen und unveränderten Daten basieren, was die Aussagekraft und Genauigkeit der Analysen erhöht.

3.2 Statistische Analyse der Hypothesen

Für jede Hypothese wurden die folgenden Schritte durchgeführt:

1. Deskriptive Statistiken: Zusammenfassung der Daten mittels Mittelwert, Standardabweichung, Minimum, Maximum und Perzentilen.
2. Datenvisualisierung: Verwendung von Histogrammen, Boxplots, Scatterplots und Stabdiagrammen.
3. Normalitätsprüfung: Überprüfung der Normalverteilung der Daten mit dem Shapiro-Wilk-Test.
 - Wenn die p-Werte des Shapiro-Wilk-Tests größer als 0,05 waren, wurden die Daten als normalverteilt betrachtet.
 - Wenn die p-Werte kleiner als 0,05 waren, wurden die Daten als nicht normalverteilt betrachtet.
4. Entscheidung für statistische Tests: Auswahl geeigneter Tests basierend auf den Ergebnissen der Normalitätsprüfungen:
 - Parametrische Tests (z.B. t-Test, ANOVA) bei normalverteilten Daten.
 - Nichtparametrische Tests (z.B. Mann-Whitney-U-Test, Wilcoxon-Rangsummentest) bei nicht normalverteilten Daten.
 - Pearson-Korrelation bei normalverteilten Daten, Spearman-Korrelation bei nicht normalverteilten Daten.

Für die statistischen Tests wurde die Hypothese bei einem p-Wert kleiner als 0,05 abgelehnt, was auf einen signifikanten Unterschied hinweist.

4 Ergebnisse

4.1 Ruhepuls und BMI

Die Hypothese lautet, dass Menschen mit niedrigem BMI generell einen niedrigeren Ruhepuls haben (Gruppe 1 und 2). Die Teilnehmer wurden in zwei Gruppen eingeteilt: Teilnehmer mit niedrigem BMI (<23.00) und Teilnehmer mit hohem BMI (≥ 23.00). Der Schwellenwert basiert auf dem Median-BMI der gesamten Gruppe.

Tabelle 1: Deskriptive Statistiken für niedrigen BMI

	Mittelwert	Std	Min	25. Perzentil	50. Perzentil	75. Perzentil	Max
BMI	20.83	1.55	17.50	19.52	21.00	22.18	22.90
Ruhepuls	72.26	12.50	52.00	63.50	72.00	79.75	98.00

Tabelle 2: Deskriptive Statistiken für hohen BMI

	Mittelwert	Std	Min	25. Perzentil	50. Perzentil	75. Perzentil	Max
BMI	27.81	6.46	23.10	23.92	25.10	27.92	49.50
Ruhepuls	75.98	14.33	51.00	65.75	73.50	84.25	105.00

Die deskriptiven Statistiken zeigen einen durchschnittlichen BMI von 20.83 und einen durchschnittlichen Ruhepuls von 72.26 in der Gruppe mit niedrigem BMI. In der Gruppe mit hohem BMI beträgt der durchschnittliche BMI 27.81 und der durchschnittliche Ruhepuls 75.98.

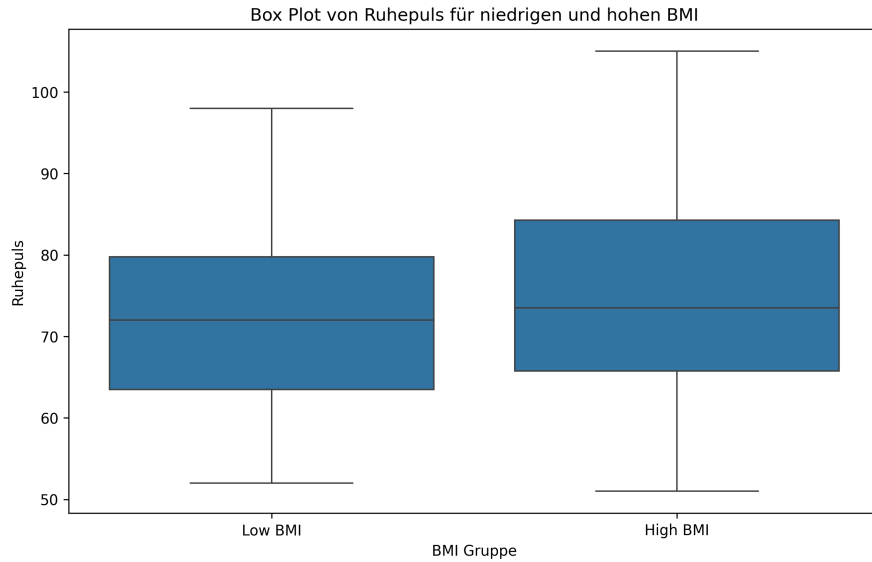


Abbildung 1: Boxplot des Ruhepulses für niedrigen und hohen BMI

Um die Normalverteilung der Daten zu prüfen, wurden Shapiro-Wilk-Tests durchgeführt. Die Ergebnisse zeigten, dass die Daten nicht normalverteilt sind. Daher wurde der Mann-Whitney-U-Test angewendet.

- Low BMI Gruppe: Ruhepuls $p = 0.166$
- High BMI Gruppe: Ruhepuls $p = 0.054$

Der Mann-Whitney-U-Test ergab einen p-Wert von 0.276, was darauf hinweist, dass der Unterschied im Ruhepuls zwischen den Gruppen nicht signifikant ist.

Zusammengefasst zeigen die Ergebnisse des Mann-Whitney-U-Tests keinen signifikanten Unterschied im Ruhepuls zwischen Teilnehmern mit niedrigem und hohem BMI in dieser Stichprobe. Daher stützen diese Ergebnisse nicht die Hypothese, dass Menschen in dieser Stichprobe mit niedrigem BMI generell einen niedrigeren Ruhepuls haben.

4.2 Dominante Hand und Armstärke

Um die Hypothese zu testen, dass die dominante Hand größer ist und der entsprechende Arm stärker ist, wurden die Teilnehmer basierend auf ihrer dominanten Hand kategorisiert: Rechtshänder und Linkshänder.

Die deskriptiven Statistiken sind in der folgenden Tabelle zusammengefasst:

Tabelle 3: Deskriptive Statistiken für Handlänge und Stamina

	Mittelwert	Std	Min	25.	50.	75.	Max
Handlänge Rechts (cm)	17.82	1.74	16.00	17.00	18.00	18.18	28.00
Handlänge Links (cm)	17.82	1.78	15.00	17.00	18.00	18.48	28.00
Stamina Rechts (s)	209.22	94.34	85.00	146.75	180.00	290.00	430.00
Stamina Links (s)	193.00	95.08	62.00	127.00	150.00	250.50	426.00

Die Ergebnisse der Shapiro-Wilk-Tests zeigten, dass die Daten nicht normalverteilt sind:

- Handlänge Rechts : $p = 7.60 \times 10^{-10}$
- Handlänge Links : $p = 4.42 \times 10^{-9}$
- Stamina Rechts : $p = 5.41 \times 10^{-5}$
- Stamina Links : $p = 3.54 \times 10^{-6}$

Der Wilcoxon Signed-Rank Test für die Handlänge ergab einen p-Wert von 0.838 (> 0.05), was darauf hinweist, dass es keinen signifikanten Größenunterschied zwischen der dominanten und nicht-dominanten Hand gibt. Der Wilcoxon Signed-Rank Test für die Stamina zeigte jedoch einen p-Wert von 9.30×10^{-6} (< 0.05), was auf einen signifikanten Unterschied in der Stärke zwischen dem dominanten und nicht-dominanten Arm hinweist.

Daher wird die Hypothese, dass die dominante Hand größer ist, nicht unterstützt, aber die Hypothese, dass der dominante Arm stärker ist, wird durch die Daten gestützt.

4.3 Wasserverbrauch nach Geschlecht

Um die Hypothese zu testen, dass männliche Teilnehmer mehr als weibliche Teilnehmer trinken, wurde eine statistische Analyse des täglichen Wasserkonsums durchgeführt.

Die deskriptiven Statistiken für den Wasserkonsum bei männlichen und weiblichen Teilnehmern sind wie folgt:

Tabelle 4: Deskriptive Statistiken für Wasserkonsum nach Geschlecht

	Mittelwert	Std	Min	25.	50.	75.	Max
Männlich	2031.25	733.72	500.00	1500.00	2000.00	2500.00	3600.00
Weiblich	1717.50	684.96	500.00	1450.00	1500.00	2000.00	4000.00

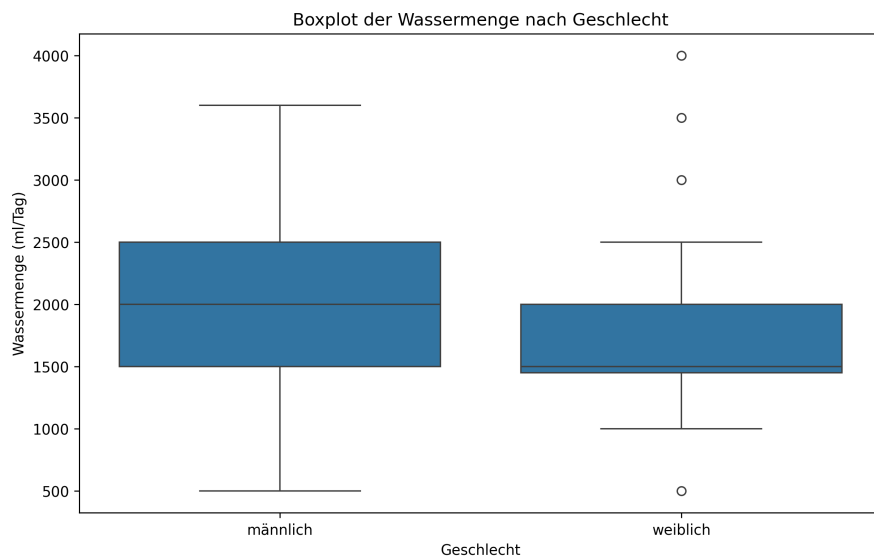


Abbildung 2: Boxplot des Wasserkonsums nach Geschlecht

Die Ergebnisse des Shapiro-Wilk-Tests zeigen, dass die Daten für beide Gruppen nicht normalverteilt sind.

- Männlich: $p = 0.3073$
- Weiblich: $p = 0.0005$

Da die Daten nicht normalverteilt sind, wurde der Mann-Whitney-U-Test verwendet, um den Wasserkonsum zwischen den Geschlechtern zu vergleichen. Der Test ergab einen p -Wert von 0.0523, was statistisch nicht signifikant ist ($p > 0.05$).

Somit zeigen die Ergebnisse keinen signifikanten Unterschied im Wasserkonsum zwischen männlichen und weiblichen Teilnehmern in dieser Stichprobe, wodurch die Hypothese nicht gestützt wird.

4.4 Blinzelfrequenz während der Videoaufgabe

Die Hypothese lautet, dass die Teilnehmer während der Aufgabe „Video“ mehr blinzeln als der Normalwert (Literaturwert 10-15 mal pro Minute) angibt. Die Blinkraten der Teilnehmer während der Aufgabe wurden analysiert.

Tabelle 5: Deskriptive Statistiken für Häufigkeit Blinzeln (/min)

	Mittelwert	Std	Min	25.	50.	75.	Max
Häufigkeit Blinzeln (/min)	13.27	9.78	0.52	7.76	9.83	16.81	45.52

Die deskriptiven Statistiken zeigen die Verteilung der Blinkraten unter den Teilnehmern.

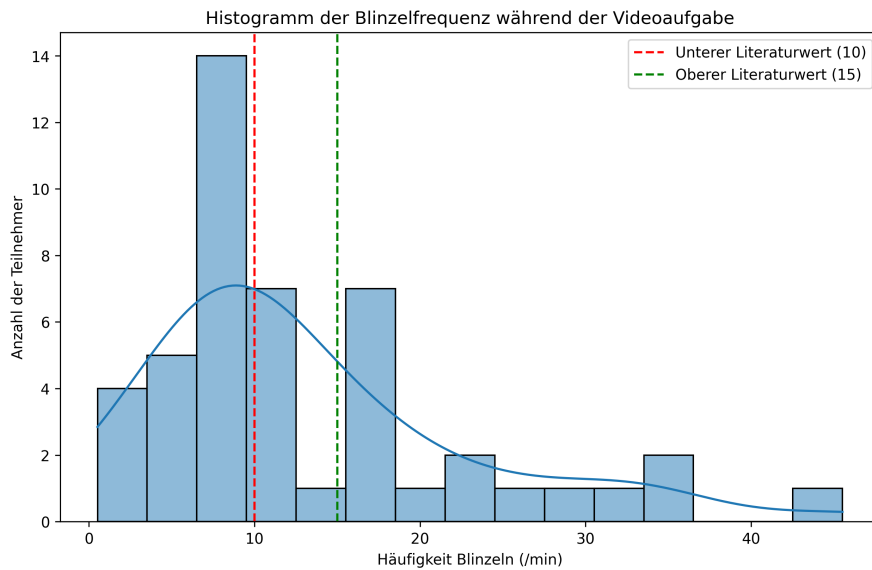


Abbildung 3: Histogramm der Blinzelfrequenz

Der Shapiro-Wilk-Test Ergebnis: p -Wert = 0.0001 zeigen, dass die Daten nicht normal verteilt sind. Dies wird auch durch das Histogramm bestätigt, welches eine deutliche Abweichung von der Normalverteilung zeigt. Folglich wurde der Wilcoxon Signed-Rank Test angewendet.

Der Wilcoxon Signed-Rank Test ergab einen p -Wert von 0.5255, was darauf hinweist, dass kein signifikanter Unterschied zur normalen Blinkrate besteht ($p > 0.05$).

Daher wird die Hypothese, dass die Teilnehmer während der Aufgabe „Video“ mehr blinzeln als der Normalwert, durch die Daten in dieser Stichprobe nicht gestützt.

4.5 Blinzelfrequenz zwischen Männern und Frauen

Um die Hypothese zu testen, dass Frauen häufiger als Männer blinzeln, wurden die Blinzelfrequenzdaten für beide Geschlechter analysiert.

Die deskriptiven Statistiken für die Blinzelfrequenz sind wie folgt:

Tabelle 6: Deskriptive Statistiken für die Blinzelfrequenz nach Geschlecht

	Mittelwert	Std	Min	25. Perzentil	50. Perzentil	75. Perzentil	Max
Männlich	14.97	12.57	2.59	5.43	9.31	20.43	45.52
Weiblich	12.12	7.37	0.52	8.28	10.09	15.65	34.14

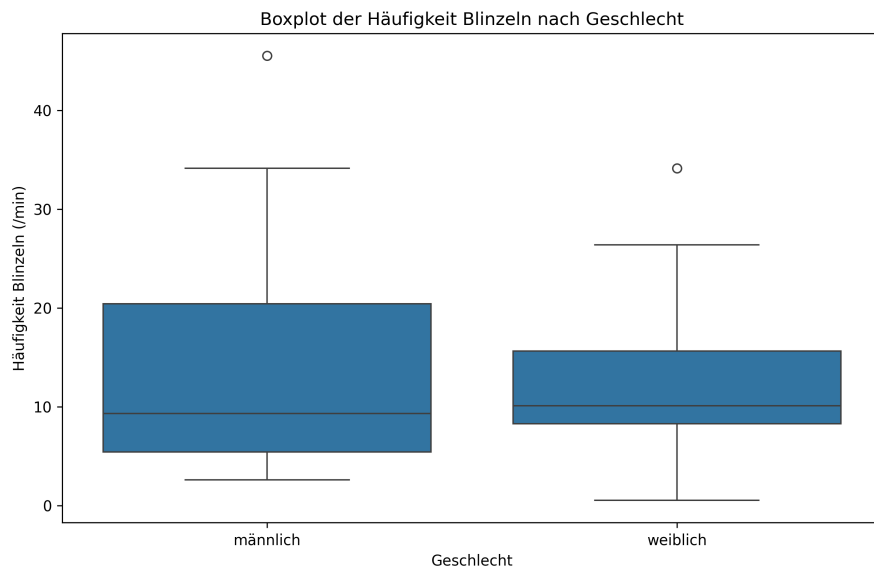


Abbildung 4: Boxplot der Blinzelfrequenz nach Geschlecht

Der Shapiro-Wilk-Test Ergebnisse: p -Wert = 0.0075 für Männer und p -Wert = 0.0087 für Frauen zeigen, dass die Daten nicht normalverteilt sind. Folglich wurde der Wilcoxon Signed-Rank Test angewendet.

Der Wilcoxon Signed-Rank Test ergab einen p -Wert von 0.948, was darauf hinweist, dass es keinen signifikanten Unterschied in der Blinzelfrequenz zwischen Männern und Frauen in dieser Stichprobe gibt. Daher wird die Hypothese, dass Frauen häufiger blinzeln als Männer, durch die Daten nicht gestützt.

4.6 Körpergröße und Luftanhaltezeit

Um die Hypothese zu testen, dass größere Menschen länger die Luft anhalten können, wurden die Daten analysiert, indem die Körpergröße und die Dauer des Luftanhaltens verglichen wurden.

Die deskriptiven Statistiken für die Körpergröße und die Dauer des Luftanhaltens sind in der folgenden Tabelle dargestellt:

Tabelle 7: Deskriptive Statistiken für Körpergröße und Dauer des Luftanhaltens

	Mittelwert	Std	Min	25. Perzentil	50. Perzentil	75. Perzentil	Max
Körpergröße, cm	172.50	10.59	155.00	165.00	170.50	178.00	204.00
Luftanhalten (s)	63.32	22.99	24.25	45.50	59.25	73.75	143.00

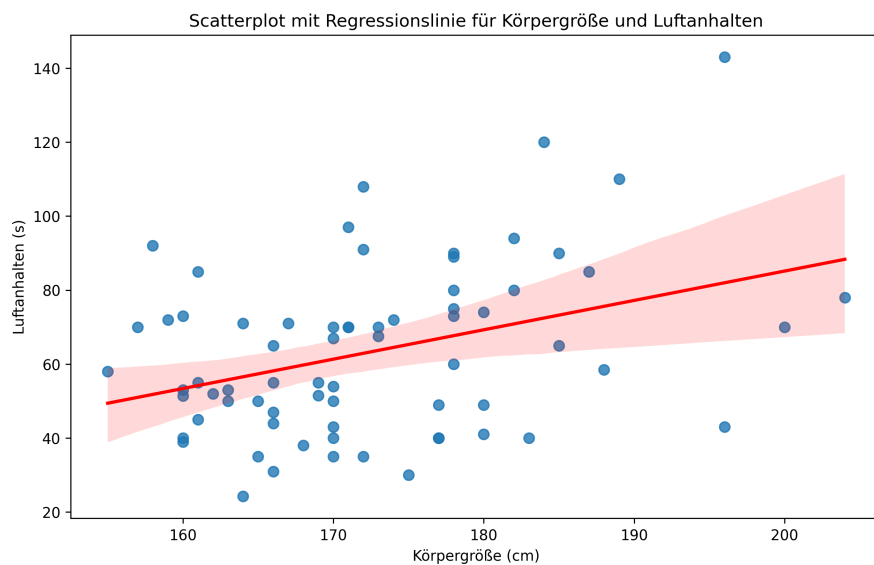


Abbildung 5: Scatterplot der Körpergröße und der Dauer des Luftanhaltens

Die Verteilung der Daten wurde mit dem Shapiro-Wilk-Test überprüft, der zeigte, dass sowohl die Körpergröße ($p = 0.0076$) als auch die Dauer des Luftanhaltens ($p = 0.0055$) nicht normal verteilt sind.

Ein Spearman-Korrelationstest wurde durchgeführt, um die Beziehung zwischen Körpergröße und Dauer des Luftanhaltens zu bewerten. Die Ergebnisse zeigten einen Spearman-Korrelationskoeffizienten von 0.315 mit einem p -Wert von 0.0079, was auf eine signifikante positive Korrelation zwischen Körpergröße und Dauer des Luftanhaltens hinweist.

Zusammenfassend unterstützen die Ergebnisse die Hypothese, dass größere Menschen länger die Luft anhalten können, wie durch die signifikante positive Korrelation zwischen Körpergröße und Dauer des Luftanhaltens in dieser Stichprobe belegt.

4.7 Linkshänder und Luftanhaltezeit

Um die Hypothese zu testen, dass Linkshänder länger die Luft anhalten können als Rechtshänder, wurden die Teilnehmer basierend auf ihrer priorisierten Hand kategorisiert: Rechtshänder und Linkshänder.

Die deskriptiven Statistiken für die Dauer des Luftanhaltens (in Sekunden) für beide Gruppen sind wie folgt:

Tabelle 8: Descriptive Statistics for Luftanhalten (s) grouped by Priorisierte Hand

	Stichprobengröße	Mittelwert	Std	Min	25.	50.	75.	Max
Linkshänder	4.00	56.56	22.59	24.25	49.56	64.00	71.00	74.00
Rechsthänder	65.00	63.21	22.92	30.00	45.00	58.50	73.00	143.00

Es ist in diesem Fall nicht sinnvoll, die statistischen Tests fortzusetzen, da die geringe Stichprobengröße der Linkshänder keine aussagekräftigen Ergebnisse ermöglicht.

5 Schlussfolgerung und Diskussion

Ziel der Studie war es, fundierte statistische Techniken anzuwenden, um eine Reihe von Hypothesen in Bezug auf Gesundheitsparameter zu testen. Die Ergebnisse sind in der folgenden Tabelle zusammengefasst.

Tabelle 9: Hypothesen und Ergebnisse der statistischen Tests

Hypothese	P-Wert	Bedeutung
Niedriger BMI, niedriger Puls	0.276	Nicht signifikant
Dominante Hand größer	0.838	Nicht signifikant
Dominanter Arm stärker	9.29×10^{-6}	Sehr hoch signifikant
Männer trinken mehr Wasser	0.052	Nicht signifikant
Mehr Blinzeln als Literaturwert	0.526	Nicht signifikant
Frauen blinzeln öfter	0.948	Nicht signifikant
Größere Menschen halten länger die Luft an	0.008	Hoch signifikant
Linkshänder halten länger die Luft an	-	Zu kleine Stichprobe

Die Hypothese, dass Personen mit niedrigem BMI einen niedrigeren Ruhepuls haben, zeigte keinen signifikanten Unterschied, was darauf hindeutet, dass sie für unseren Datensatz möglicherweise nicht zutrifft. Die Untersuchung, ob die dominante Hand größer und stärker ist, ergab widersprüchliche Ergebnisse: Es gab keinen signifikanten Größenunterschied, aber der dominante Arm war deutlich stärker.

Unsere Analyse des Wasserverbrauchs nach Geschlecht hat keine signifikante Unterschiede gefunden und somit widerlegt das Argument, dass Männer mehr trinken als Frauen. Außerdem fanden wir keine Anzeichen dafür, dass die Teilnehmer während des Videoauftritts häufiger blinzeln als normal. Auch die Untersuchung der Geschlechtsunterschiede in der Blinzelfrequenz zeigte keinen signifikanten Unterschied. Dennoch bestätigte die Untersuchung, dass größere Personen länger die Luft anhalten können, indem sie eine starke positive Beziehung zwischen Größe und Atemanhaltezeit zeigte.

Abschließend war die geringe Anzahl von linkshändigen Probanden ein Einschränkungsfaktor, wodurch valide statistische Tests für diese Gruppe erschwert wurden. Um zuverlässige Schlussfolgerungen durch qualitative Analysen zu ziehen, sind daher mehr Teilnehmer erforderlich.

6 Fazit

Die Erhebung des Datenmaterials für den „FH Gesundheitscheck“ wurde ohne Beratung der Analytiker gemacht. Die Datenverarbeitung hat erstaunlicherweise die Unregelmäßigkeiten in den gesammelten Informationen ausgeglichen. Der Ansatz zum Umgang mit fehlenden Daten war entscheidend, um so viele Fragen wie möglich ohne Verzerrungen zu beantworten. Es gibt die größte Fehlermöglichkeiten hinsichtlich dem Prozess der Datenerhebung selbst und wir konnten letztens die Aufstellung der endgültigen Hypothese wegen unzureichender Stichprobe für Linkshänder nicht beantworten.

Tabellenverzeichnis

1	Deskriptive Statistiken für niedrigen BMI	4
2	Deskriptive Statistiken für hohen BMI	4
3	Deskriptive Statistiken für Handlänge und Stamina	5
4	Deskriptive Statistiken für Wasserkonsum nach Geschlecht	6
5	Deskriptive Statistiken für Häufigkeit Blinzeln (/min)	7
6	Deskriptive Statistiken für die Blinzelfrequenz nach Geschlecht	8
7	Deskriptive Statistiken für Körpergröße und Dauer des Luftanhaltens	9
8	Descriptive Statistics for Luftanhalten (s) grouped by Priorisierte Hand	10
9	Hypothesen und Ergebnisse der statistischen Tests	11

Abbildungsverzeichnis

1	Boxplot des Ruhepulses für niedrigen und hohen BMI	4
2	Boxplot des Wasserkonsums nach Geschlecht	6
3	Histogramm der Blinzelfrequenz	7
4	Boxplot der Blinzelfrequenz nach Geschlecht	8
5	Scatterplot der Körpergröße und der Dauer des Luftanhaltens	9