



FernUniversität in Hagen
Fakultät für Mathematik und Informatik

Statistisches Praktikum
Sommersemester 2025

Aufgabe 20
Überlebenszeitanalyse

von
Alexander Greif
Matrikelnummer: 4214242

Datum der Abgabe: 04.05.2025

Inhaltsverzeichnis

1. Einleitung	3
2. Grundlagen	5
2.1. Überlebenszeitanalyse	5
2.2. Der Datensatz	6
3. Hauptteil	9
3.1. (i) Kaplan-Meier-Schätzer und -Kurven	9
3.1.1. Theoretischer Hintergrund	9
3.1.2. Umsetzung	11
3.2. (ii) Log-Rank-Test	13
3.2.1. Theoretischer Hintergrund	13
3.2.2. Umsetzung	16
3.2.3. Interpretation	17
3.3. (iii) Störfaktoren	18
3.3.1. Theoretischer Hintergrund	20
3.3.2. Umsetzung	21
Literaturverzeichnis	
A. Anhang	III

1. Einleitung

Zur Untersuchung der Wirksamkeit eines Medikaments wurde eine Längsschnittstudie mit 30 Patienten durchgeführt. Anhand der erhobenen Daten soll untersucht werden, ob das Medikament die Überlebenszeit der Patienten verlängern kann. Dazu führen wir eine Überlebenszeitanalyse durch.

Im ersten Schritt schätzen wir die Überlebensfunktion. Sie gibt an, mit welcher Wahrscheinlichkeit ein Patient einen gegebenen Zeitpunkt überlebt. Dafür verwenden wir den Kaplan-Meier-Schätzer. Wir unterscheiden dabei zwischen Patienten, die das Medikament erhalten haben, und denen, die ein Placebo bekommen haben. Die Ergebnisse stellen wir grafisch durch Kaplan-Meier-Kurven dar. Anschließend vergleichen wir die Überlebensfunktionen der beiden Gruppen mithilfe des Log-Rank-Tests. Zeigt sich ein statistisch signifikanter Unterschied, können wir davon ausgehen, dass das Medikament einen Einfluss auf die Überlebenszeit hat. Schließlich prüfen wir die Aussagekraft der Ergebnisse, indem wir die Kovariate Alter betrachten. Dazu testen wir, ob die Altersverteilungen in beiden Gruppen vergleichbar sind. Ist dies nicht der Fall, könnte das Alter als Störfaktor die Ergebnisse verzerren.

2. Grundlagen

Bevor wir uns den eigentlichen Fragestellungen widmen, besprechen wir die Grundlagen der Überlebenszeitanalyse und verschaffen uns einen Überblick über den vorliegenden Datensatz.

2.1. Überlebenszeitanalyse

Bei der Überlebenszeitanalyse - im Deutschen je nach Anwendungsbereich auch unter anderen Namen, wie Ereigniszeit- oder Zuverlässigkeitsanalyse bekannt - wird eine nicht-negative Zufallsvariable T untersucht, die die Zeit bis zum Eintritt eines bestimmten Ereignisses beschreibt. Beispiele hierfür sind die Laufzeit einer Glühbirne bis zum Defekt, die Zeit zwischen Entzug und Rückfall bei Suchterkrankten oder - wie in unserem Fall, der zugleich namensgebend für den Begriff Überlebenszeitanalyse ist - die Zeit zwischen Krankheitsdiagnose und durch die Krankheit verursachtem Tod. Eine zentrale Aufgabe der Überlebenszeitanalyse besteht darin, die Überlebensfunktion

$$S : \mathbb{R}_0^+ \rightarrow [0, 1], t \mapsto P(T > t)$$

zu schätzen. Die Überlebensfunktion beschreibt die Wahrscheinlichkeit, dass das Ereignis zum Zeitpunkt t noch nicht eingetreten ist. Sind F die Verteilungsfunktion und S die Überlebensfunktion einer Zufallsvariablen T , dann gilt $S(t) = 1 - F(t)$ für alle $t \in \mathbb{R}_0^+$. Zur Schätzung der Überlebensfunktion existieren verschiedene Verfahren. Der Kaplan-Meier-Schätzer, den wir in dieser Arbeit verwenden, ist der nichtparametrische Maximum-Likelihood-Schätzer der Überlebensfunktion[1].

Ein weiterer Bestandteil der Überlebenszeitanalyse ist die Untersuchung des Einflusses weiterer Variablen auf die Zeit bis zum Ereignis. In unserem Fall interessiert uns die Einnahme des Medikaments bzw. Placebos, also

2. Grundlagen

die Gruppenzugehörigkeit zur Interventions- oder Kontrollgruppe. Hierbei handelt es sich um eine kategoriale Variable, und wir werden ihren Einfluss durch einen Log-Rank-Test[2] untersuchen. Abschließend überprüfen wir, ob eine weitere Variable – das Alter der Patient:innen – möglicherweise unsere Ergebnisse zum Einfluss des Medikaments verfälschen konnte. Dazu vergleichen wir die Altersverteilungen der beiden Gruppen.

2.2. Der Datensatz

Der Datensatz umfasst Daten zu 30 Patienten. Von diesen haben 15 das Medikament (Interventionsgruppe) und 15 ein Placebo erhalten (Kontrollgruppe). Zu jedem der Patienten haben wir die folgenden Variablen:

- Zeit - Der Zeitpunkt des Ausscheidens aus der Studie. Die Einheit und der genaue Bezugspunkt sind nicht dokumentiert. Wir gehen im Weiteren davon aus, dass die Variable metrisch skaliert ist.
- status - Der Status des Patienten zum Zeitpunkt seines Ausscheidens (tot oder lebendig).
- Gruppe - Die Gruppenzugehörigkeit des Patienten (Interventions- oder Kontrollgruppe). Patienten in der Interventionsgruppe haben das Medikament, Patienten in der Kontrollgruppe das Placebo erhalten.
- Alter - Das Alter des Patienten zu einem unbekannten Zeitpunkt. Wir gehen im Folgenden davon aus, dass bei allen Patienten das Alter, relativ zur Variable Zeit, zum gleichen Zeitpunkt erhoben wurde. Hier fehlen ebenfalls Informationen zur Einheit. Wir nehmen an, dass das Alter in einer üblichen Zeiteinheit - etwa Tage, Wochen, Monate oder Jahre - gegeben ist. Da für jede dieser Einheiten das Alter metrisch skaliert ist, sind die Analysemethoden und -ergebnisse unabhängig von der genauen Einheit. Lediglich die Interpretation würde sich unterscheiden. Aufgrund der vorliegenden Werte gehen wir davon aus, dass das Alter in Jahren gegeben ist.

Der Datensatz enthält keine fehlenden Einträge. In beiden Gruppen waren zum Zeitpunkt des Ausscheidens jeweils 10 Patienten tot und 5 lebendig. Scheiden Patienten mit Status „lebendig“ aus, sprechen wir von Zensurierung. Die Todes- und Zensurierungszeitpunkte der beiden Gruppen sind

in Abbildung 2.1 als Boxplots inklusive der individuellen Datenpunkte dargestellt.¹

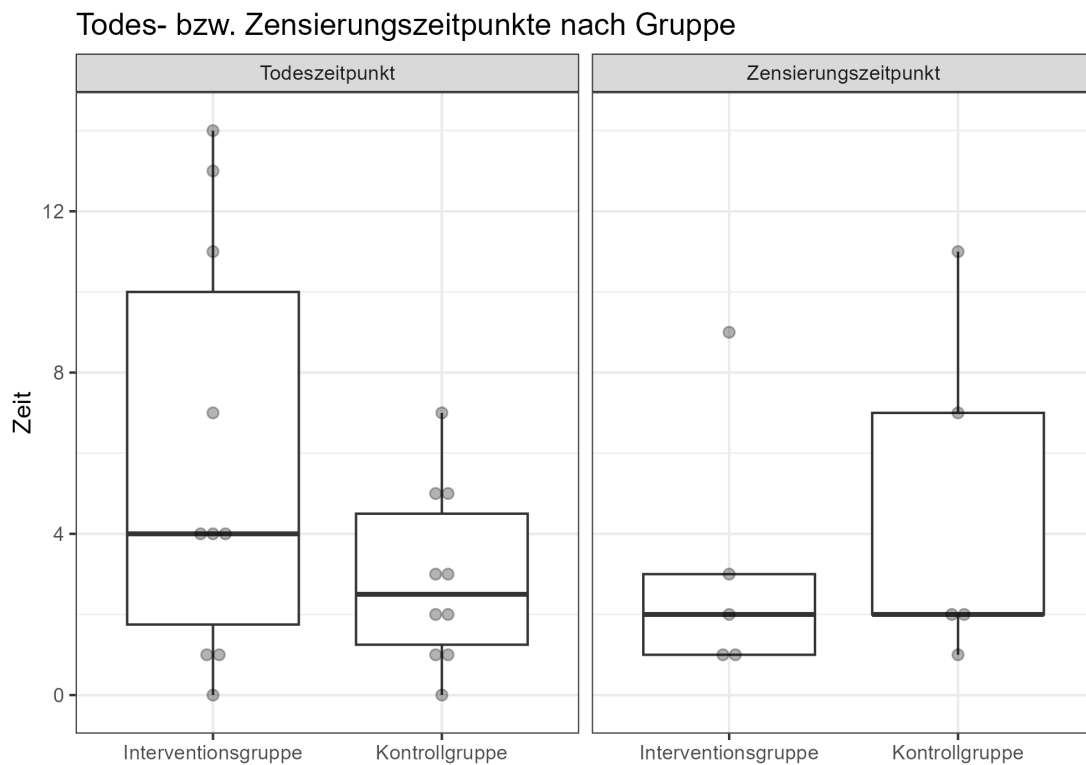


Abbildung 2.1.: Boxplots für Todes- und Zensierungszeitpunkte nach Gruppe

Tabelle 2.1 enthält neben Mittelwert und Standardabweichung der Zeitpunkte auch Mittelwert und Standardabweichung des Alters der beiden Gruppen.

	Todeszeitpunkt (n=10)	Zensierungszeitpunkt (n=5)	Alter (n=15)
Interventionsgruppe	5.9 (5.13)	3.2 (3.35)	41.7 (12.84)
Kontrollgruppe	2.9 (2.18)	4.6 (4.28)	76.0 (8.78)

Tabelle 2.1.: Mittelwert (Standardabweichung); Teilnehmerzahlen n pro Gruppe

Die deskriptiven Darstellungen deuten auf mögliche Gruppenunterschiede bezüglich Todeszeitpunkt und Alter hin. In den folgenden Abschnitten werden wir diese Tendenzen systematisch untersuchen.

¹Boxplots zum Alter befinden sich in Abschnitt 3.3, wo wir die Altersverteilungen genauer untersuchen.

3. Hauptteil

3.1. (i) Kaplan-Meier-Schätzer und -Kurven

Das Ziel dieses Abschnitts ist die Berechnung der Kaplan-Meier-Schätzer beider Gruppen sowie ihre grafische Darstellung in Form der Kaplan-Meier-Kurven. Zuvor klären wir noch, was der Kaplan-Meier-Schätzer genau ist, wie er berechnet wird und warum er sich in unserem Fall zur Schätzung der Überlebensfunktion eignet.

3.1.1. Theoretischer Hintergrund

Wir haben bereits in Abschnitt 2.1 besprochen, dass es sich beim Kaplan-Meier-Schätzer um einen nichtparametrischen Schätzer der Überlebensfunktion handelt. In diesem Abschnitt werden wir den Kaplan-Meier-Schätzer heuristisch herleiten und seine Voraussetzungen sowie einige Eigenschaften besprechen.

Die Überlebensfunktion gibt die Wahrscheinlichkeit an, dass zu einem bestimmten Zeitpunkt noch kein Ereignis eingetreten ist. Das Ereignis sowie das beobachtete Objekt, auf dem das Ereignis eintritt, können dabei verschiedene Formen annehmen (Abschnitt 2.1). Um unsere Formulierungen einfach zu halten, gehen wir im Folgenden von unserem konkreten Fall aus, dass wir Personen untersuchen und das uns interessierende Ereignis deren Tod ist. Die folgenden Ausführungen gelten aber unabhängig vom konkreten untersuchten Objekt oder Ereignis allgemein für die Überlebensfunktion und den Kaplan-Meier-Schätzer.

Wir betrachten einen Datensatz von $N \in \mathbb{N}$ Personen mit $n \in \mathbb{N}$ verschiedenen Todeszeitpunkten. Für $i \in \{1, \dots, n\}$ bezeichne

- t_i den entsprechenden Zeitpunkt,

3. Hauptteil

- n_i die Anzahl der zum Zeitpunkt t_i beobachteten Personen, also diejenigen die nicht vor dem Zeitpunkt t_i ausgeschieden sind,
- d_i die Anzahl der zum Zeitpunkt t_i verstorbenen (unter den beobachteten) Personen.

Dabei wählen wir die Indizierung so, dass die $t_i, i \in \{1, \dots, n\}$ monoton (wegen der Eindeutigkeit der Todeszeitpunkte sogar streng monoton) steigen. Wir nehmen an, dass die Todeszeitpunkte der Personen durch stochastisch unabhängige und identisch verteilte Zufallsvariablen beschrieben werden.

Gehen wir zunächst davon aus, dass Personen nur dann aus der Studie ausscheiden, wenn sie (aus dem uns interessierenden Grund) versterben. Wir sprechen in diesem Fall von unzensierten Daten. Es gilt dann

$$n_i = N - \sum_{j=1}^{i-1} d_j = n_{i-1} - d_{i-1}$$

und wir können $S(t_i)$ intuitiv sinnvoll schätzen durch

$$\hat{S}_{naiv}(t_i) = \frac{(n_i - d_i)}{N} = \frac{n_{i+1}}{N}.$$

Tatsächlich ist dieser Schätzer im unzensierten Fall genau der Kaplan-Meier-Schätzer und maximiert somit die empirische Likelihood-Funktion. Wir haben bereits gesehen, dass unser Datensatz nicht unzensiert ist. So scheiden einige Teilnehmende mit dem Status „lebend“ aus unbekannten Gründen aus. Dies ist typisch für Langzeitstudien, da Studienteilnehmende aus verschiedensten Gründen ausscheiden können. Wir sprechen von rechtszensierten Daten. Der zuvor eingeführte Schätzer verliert in diesem Fall an Bedeutung, denn er geht davon aus, dass zum Zeitpunkt t_i **genau** $n_i - d_i$ Personen noch leben und **genau** $N - (n_i - d_i)$ tot sind. Bei rechtszensierten Daten wissen wir aber nur, dass zum Zeitpunkt t_i **mindestens** $n_i - d_i$ Personen noch leben und **höchstens** $N - (n_i - d_i)$ tot sind, da zuvor ausgeschiedene Personen nicht notwendigerweise tot sind. Der Kaplan-Meier-Schätzer berücksichtigt diesen Umstand unter der Annahme, dass die Zufallsvariable des Todeszeitpunkts unabhängig von der des Zensierungszeitpunkts ist. Dann wird der Kaplan-Meier-Schätzer $\hat{S}(t_i)$ für den Zeitpunkt t_i als das Produkt der Schätzungen der bedingten Wahrscheinlichkeiten $P(\text{lebendig in } t_j | \text{lebendig bis zu } t_j)$ für $j \in \{1, \dots, i\}$ berechnet.

Da zu einem bestimmten Zeitpunkt lebendig zu sein komplementär dazu ist, zum gleichen Zeitpunkt tot zu sein, gilt

$$P(\text{lebendig in } t_j | \text{lebendig bis zu } t_j) = 1 - P(\text{tot in } t_j | \text{lebendig bis zu } t_j).$$

Wir schätzen $P(\text{tot in } t_j | \text{lebendig bis zu } t_j)$ durch $\frac{d_j}{n_j}$ und erhalten

$$\hat{S}(t_i) = \prod_{j=1}^i \left(1 - \frac{d_j}{n_j}\right)$$

und

$$\hat{S}(t) = \hat{S}(\max\{t_i | i \in \{1, \dots, n\}, t_i \leq t\}) = \prod_{\substack{i \in \{1, \dots, n\} \\ t_i \leq t}} \left(1 - \frac{d_i}{n_i}\right).$$

Für die Herleitung haben wir angenommen, dass Todeszeitpunkte durch stochastisch unabhängige und identisch verteilte Zufallsvariablen beschrieben werden, die zudem vom Zensierungszeitpunkt unabhängig sind. Unter diesen Annahmen ist der Kaplan-Meier-Schätzer der Maximum-Likelihood-Schätzer der Überlebensfunktion[1].

3.1.2. Umsetzung

Zur Berechnung der Kaplan-Meier-Schätzer für die vorliegenden Daten folgen wir dem in Abschnitt 3.1.1 beschriebenen Vorgehen. Dabei schätzen wir zunächst

$$P(\text{lebendig in } t_i | \text{lebendig bis zu } t_i)$$

durch $1 - \frac{d_i}{n_i}$ für alle $i \in \{1, \dots, n\}$ und anschließend die Kaplan-Meier-Schätzer als Produkte dieser Schätzungen. Dies machen wir getrennt für die Interventions- und die Kontrollgruppe.

Wir führen die Berechnung mithilfe von Tabellen durch. Dafür sortieren wir zunächst die Daten jeder Gruppe nach der Variable Zeit (Tabelle A.1 im Anhang). Aus dieser sortierten Tabelle lassen sich die Todeszeitpunkte t_i ablesen. Damit können wir die Anzahl der verschiedenen Zeitpunkte n bestimmen. Anschließend legen wir eine neue Tabelle mit n Zeilen und den Spalten i , t_i , d_i , n_i , $1 - \frac{d_i}{n_i}$ und $\hat{S}(t_i)$ an. Die erste Spalte enthält die Zahlen 1 bis n , die zweite bis vierte Spalte können wir mithilfe der sortierten

3. Hauptteil

Tabelle der Rohdaten ausfüllen. Die restlichen Spalten werden schrittweise berechnet – zunächst der Ausdruck $1 - \frac{d_i}{n_i}$, dann das kumulative Produkt $\hat{S}(t_i)$. Die vollständigen Tabellen für beide Gruppen sind in Tabelle 3.1 und Tabelle 3.2 dargestellt.

i	t _i	n _i	d _i	d _i /n _i	1 – d _i /n _i	$\hat{S}(t_i)$
1	0	15	1	0.067	0.933	0.933
2	1	14	2	0.143	0.857	0.800
3	2	10	0	0.000	1.000	0.800
4	3	9	0	0.000	1.000	0.800
5	4	8	3	0.375	0.625	0.500
6	7	5	1	0.200	0.800	0.400
7	9	4	0	0.000	1.000	0.400
8	11	3	1	0.333	0.667	0.267
9	13	2	1	0.500	0.500	0.133
10	14	1	1	1.000	0.000	0.000

Tabelle 3.1.: Kaplan-Meier-Schätzer Interventionsgruppe

i	t _i	n _i	d _i	d _i /n _i	1 – d _i /n _i	$\hat{S}(t_i)$
1	0	15	1	0.067	0.933	0.933
2	1	14	2	0.143	0.857	0.800
3	2	11	2	0.182	0.818	0.655
4	3	7	2	0.286	0.714	0.468
5	5	5	2	0.400	0.600	0.281
6	7	3	1	0.333	0.667	0.187
7	11	1	0	0.000	1.000	0.187

Tabelle 3.2.: Kaplan-Meier-Schätzer Kontrollgruppe

Die grafische Darstellung des Funktionsgraphen der geschätzten Überlebensfunktion $\hat{S}$, die Kaplan-Meier-Kurve, befindet sich in Abbildung 3.1. Sie wurde mithilfe der R-Packages `survival` (zur Berechnung) und `ggplot2` (zur grafischen Darstellung) erstellt.

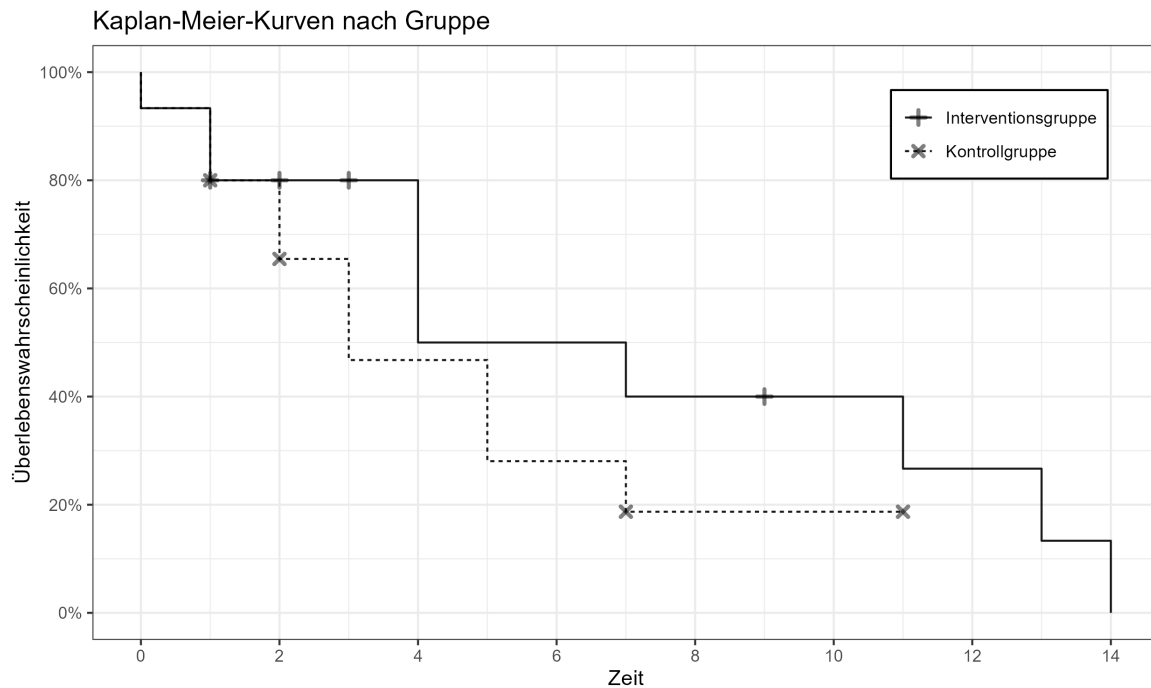


Abbildung 3.1.: Kaplan-Meier-Kurven nach Gruppe inklusive Zensierungszeitpunkte

3.2. (ii) Log-Rank-Test

Die Kaplan-Meier-Kurven deuten für jeden Zeitpunkt auf eine höhere Überlebenswahrscheinlichkeit in der Interventionsgruppe hin. Dies legt einen positiven Effekt der Therapie nahe. Um zu prüfen, ob dieser Effekt statistisch signifikant ist, nutzen wir den Log-Rank-Test.

3.2.1. Theoretischer Hintergrund

Der Log-Rank-Test ist ein Hypothesentest zum Vergleich der Überlebensfunktionen zweier Stichproben. Die Hypothesen lauten:

H_0 : Die beiden Überlebensfunktionen sind identisch.

H_1 : Die beiden Überlebensfunktionen sind verschieden.

Dabei gelten wie beim Kaplan-Meier-Schätzer die Annahmen, dass die Todeszeitpunkte durch stochastisch unabhängige und identisch verteilte Zufallsvariablen beschrieben werden und Todes- und Zensierungszeitpunkt voneinander unabhängig sind[3].

3. Hauptteil

Zur heuristischen Herleitung betrachten wir die Überlebenszeiten zweier Gruppen, Gruppe 1 und Gruppe 2, mit insgesamt $k \in \mathbb{N}$ verschiedenen Todeszeitpunkten. Wir gehen davon aus, dass alle Annahmen erfüllt sind. Für $i \in \{1, 2\}$ und $j \in \{1, \dots, k\}$ bezeichne

- t_j den entsprechenden Zeitpunkt,
- n_{ij} die Anzahl der zum Zeitpunkt t_j beobachteten Personen aus Gruppe i , also diejenigen die nicht vor dem Zeitpunkt t_j ausgeschieden sind,
- d_{ij} die Anzahl der zum Zeitpunkt t_j verstorbenen (unter den beobachteten) Personen aus Gruppe i ,
- D_{ij} die entsprechenden Zufallsvariablen.

Wir wollen nun eine Teststatistik herleiten, deren Verteilung unter der Nullhypothese (zumindest näherungsweise) bekannt ist. Dabei betrachten wir ohne Beschränkung der Allgemeinheit Gruppe 1. Zunächst untersuchen wir die Verteilungen der D_{1j} .

Gilt H_0 , sind die Überlebensfunktionen beider Gruppen identisch. Es ist also jeder der $d_{1j} + d_{2j}$ Todesfälle zum Zeitpunkt t_j gleichwahrscheinlich auf eine der $n_{1j} + n_{2j}$ beobachteten Personen verteilt. Dies lässt sich durch ein Urnenmodell beschreiben, bei dem aus $n_{1j} + n_{2j}$ Kugeln d_j ohne Zurücklegen entnommen werden. Dabei befinden sich n_{1j} Kugeln mit der gewünschten Eigenschaft (Zugehörigkeit zu Gruppe 1) in der Urne. Wir interessieren uns für die Verteilung von D_{1j} , also der Anzahl der gezogenen Kugeln (gestorbenen Personen), die die gewünschte Eigenschaft besitzen (aus Gruppe 1 sind). Diese wird genau durch die Hypergeometrische Verteilung mit den Parametern

- Größe der Grundgesamtheit $n_{1j} + n_{2j} =: N$,
- Anzahl der „Erfolge“ (in unserem Fall Tode) in der Grundgesamtheit $d_{1j} + d_{2j} =: K$ und
- Größe der Stichprobe $n_{1j} =: n$

beschrieben. Mit den Formeln für Erwartungswert und Varianz der hypergeometrischen Verteilung[4] folgt

$$E(D_{1j}) = n * \frac{K}{N} = n_{1j} * \frac{(d_{1j} + d_{2j})}{(n_{1j} + n_{2j})}$$

und

$$V(D_{1j}) = n * \frac{K}{N} * (1 - \frac{K}{N}) * \frac{N - n}{N - 1} = E(D_{1j}) * (1 - \frac{d_{1j} + d_{2j}}{n_{1j} + n_{2j}}) * \frac{n_{2j}}{n_{1j} + n_{2j} - 1}.$$

Um einen Gesamteffekt über alle Zeitpunkte hinweg analysieren zu können betrachten wir die Zufallsvariable $\sum_{j=1}^k D_{1j}$. Nach dem zentralen Grenzwertsatz konvergiert die Verteilung von $\sum_{j=1}^k D_{1j}$ für $k \rightarrow \infty$ gegen die Normalverteilung $N(E(D_1), V(D_1))$. Da Erwartungswert und Varianz mit der Addition vertauschen, gilt

$$E(\sum_{j=1}^k D_{1j}) = \sum_{j=1}^k E(D_{1j})$$

und

$$V(\sum_{j=1}^k D_{1j}) = \sum_{j=1}^k V(D_{1j})$$

und wir erhalten

$$Z := \frac{\sum_{j=1}^k D_{1j} - \sum_{j=1}^k E(D_{1j})}{\sqrt{\sum_{j=1}^k V(D_{1j})}} \stackrel{\text{apx}}{\sim} N(0, 1)$$

und

$$Z^2 \stackrel{\text{apx}}{\sim} \chi_1^2.$$

Der kritische Wert zu unserer Teststatistik lässt sich also mithilfe der Dichtefunktion der χ_1^2 -Verteilung[4]

$$f_{\chi_1^2}(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}$$

näherungsweise bestimmen.

Anmerkung: Die Herleitung erfolgte exemplarisch für Gruppe 1. Analog ergibt sich für Gruppe 2 derselbe Wert für Z^2 , da gilt:

$$\sum_{j=1}^k D_{1j} - E(\sum_{j=1}^k D_{1j}) = -(\sum_{j=1}^k D_{2j} - E(\sum_{j=1}^k D_{2j}))$$

und

$$V(\sum_{j=1}^k D_{1j}) = V(\sum_{j=1}^k D_{2j}).$$

3.2.2. Umsetzung

Wir berechnen die Teststatistik für unsere Stichprobe durch

$$z = \frac{\sum_{j=1}^k d_{1j} - \sum_{j=1}^k E(D_{1j})}{\sqrt{\sum_{j=1}^k V(D_{1j})}}.$$

Anschließend vergleichen wir, da wir einen ungerichteten Hypothesentest durchführen, z^2 mit dem kritischen Wert der χ_1^2 -Verteilung.

Wir müssen also zunächst die Erwartungswerte und Varianzen für die verschiedenen Todeszeitpunkte berechnen. Als Hilfstabelle stellen wir Tabelle 3.3 auf. Die dafür benötigten Informationen entnehmen wir den Tabellen 3.1 und 3.2 für die Kaplan-Meier-Schätzer.

j	t _j	n _{1j}	n _{2j}	d _{1j}	d _{2j}	$\frac{d_{1j}+d_{2j}}{n_{1j}+n_{2j}}$	E(D _{1j})	V(D _{1j})
1	0	15	15	1	1	0.067	1.000	0.483
2	1	14	14	2	2	0.143	2.000	0.889
3	2	10	11	0	2	0.095	0.952	0.474
4	3	9	7	0	2	0.125	1.125	0.459
5	4	8	5	3	0	0.231	1.846	0.592
6	5	5	5	0	2	0.200	1.000	0.444
7	7	5	3	1	1	0.250	1.250	0.402
8	11	3	1	1	0	0.250	0.750	0.188
9	13	2	0	1	0	0.500	1.000	0.000
10	14	1	0	1	0	1.000	1.000	
Summe	-	-	-	10	10	-	11.924	3.930

Tabelle 3.3.: Hilfstabelle zur Berechnung der Log-Rank-Teststatistik.

Wir erhalten

$$z \approx \frac{10 - 11.924}{\sqrt{3.930}} \approx -0.970$$

und

$$z^2 \approx 0.941.$$

Den kritischen Wert der χ^2_1 -Verteilung zum Signifikanzniveau 5% können wir einer Tabelle entnehmen oder mithilfe von Software berechnen. Wir finden $\chi^2_{krit} = 3.841$. Es ist

$$z^2 \approx 0.941 < 3.841 \approx \chi^2_{krit},$$

der Gruppenunterschied ist also nicht statistisch signifikant.

In R können wir den Log-Rank-Test mit dem `survival`-Package durchführen, oder den genauen p-Wert zu $z^2 \approx 0.941$ mit `pchisq(q=0.941, df=1, lower.tail=F)` bestimmen. Wir erhalten $p \approx 0.332 > 0.05$.

3.2.3. Interpretation

Aufgrund des geringen Stichprobenumfangs ist das Testergebnis nur bedingt aussagekräftig. Die Sensitivität des Tests – also die Wahrscheinlichkeit, bei tatsächlichem Effekt ein signifikantes Ergebnis zu erhalten – hängt bei gegebener Effektstärke stark von der Stichprobengröße ab. Unter einigen zusätzlichen Annahmen können wir die Sensitivität des durchgeführten Log-Rank-Tests berechnen. Für die Berechnung vernachlässigen wir die Tatsache, dass unsere Daten zensiert sind und nehmen an, dass das Verhältnis zwischen den Steigungen der Überlebenskurven der beiden Gruppen, die sogenannte Hazard Ratio, konstant ist. Diesen können wir mithilfe unserer Daten schätzen. Dafür nutzen wir die Funktion `coxph` aus dem R-Package `Survival`. Wir erhalten einen Wert von 0.640. Gehen wir davon aus, dass die Hazard Ratio zwischen den beiden untersuchten Populationen konstant ist und unserem geschätzten Wert von 0.640 entspricht, können wir mithilfe der Methoden von Schoenfeld[5] die Sensitivität berechnen. Dazu nutzen wir die Funktion `nEvents` aus dem R-Package `gsDesign`, welche diese Methoden implementiert. Bei insgesamt 30 Todesfällen in zwei gleichgroßen Gruppen erhalten wir eine Sensitivität von 0.230. Wäre der Gruppenunterschied zwischen den wahren Überlebensfunktionen in etwa so groß, wie der zwischen unseren geschätzten Überlebensfunktionen, hätten wir mit unserer Stichprobengröße, ohne Zensierung, nicht mal eine 25%-ige Wahrscheinlichkeit, mit dem Log-Rank-Test ein signifikantes Ergebnis zu erhalten. Genauso können wir ausrechnen, dass wir mit den gleichen Annahmen ungefähr 159 Ereignisse, also Todesfälle in unseren Stichproben bräuchten, um eine Sensitivität von 80% zu erreichen. Eine Hazard Ratio

3. Hauptteil

von 0.640 kann aber durchaus einen klinisch relevanten Effekt bedeuten. Für eine fundierte Beurteilung des Effekts ist daher eine größere Stichprobe erforderlich.

Darüber hinaus kann eine kleine Anzahl beobachteter Ereignisse zu einem ungenauen Testergebnis führen, da die Standardnormalverteilung bei einer geringen Anzahl verschiedener Ereigniszeiten nicht unbedingt eine gute Näherung für die Verteilung der Teststatistik (bei Gültigkeit der Nullhypothese) darstellt. Des Weiteren reicht unser Wissen zu Ursprung und Generierung der Daten nicht aus, um beurteilen zu können, ob die Annahmen des Kaplan-Meier-Schätzers und damit des Log-Rank-Tests erfüllt sind. Ist dies nicht der Fall, wird die Aussagekraft der Teststatistik weiter reduziert.

3.3. (iii) Störfaktoren

Der vorliegende Datensatz enthält, neben der abhängigen Variable Zeit und der unabhängigen Variable Gruppe, lediglich das Alter als weitere potenzielle erklärende Variable. Wir wissen, dass die untersuchte Krankheit bei älteren Menschen schneller zum Tod führt. Ist das Alter zwischen unseren beiden Gruppen ungleich verteilt, ist also nicht davon auszugehen, dass die Überlebenszeiten in beiden Gruppen ähnlich verteilt sind.

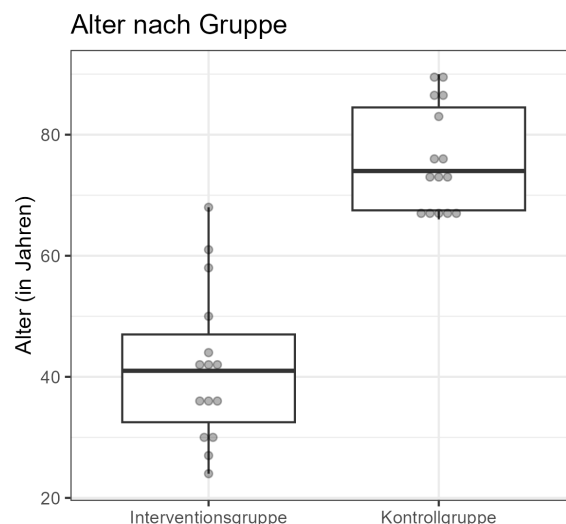


Abbildung 3.2.: Boxplots Alter nach Gruppe mit individuellen Datenpunkten

Abbildung 3.2 stellt die empirischen Altersverteilungen der beiden Gruppen grafisch durch Boxplots dar. Aufgrund des geringen Stichprobenumfangs können wir zusätzlich die individuellen Datenpunkten ergänzen. Es scheint tatsächlich ein Unterschied zwischen den Verteilungen zu bestehen. Die Wertebereiche überschneiden sich kaum. Das Alter kommt also als Störfaktor in Frage. Wir wollen unsere Vermutung wieder mit einem Hypothesentest überprüfen. Sind die Daten beider Gruppen annähernd normalverteilt, bietet sich ein T-Test an. Am Boxplot erkennen wir bereits, dass das Alter in der Kontrollgruppe nicht normalverteilt zu sein scheint, da sich viele Datenpunkte an den Enden der Verteilung sammeln. Zur genaueren Betrachtung erstellen wir Quantil-Quantil-Diagramme (Abbildung 3.3). Dabei wird der rohe Wert jedes Datenpunkts x_i auf der y-Achse gegen $\Phi^{-1}(F(x_i))$ auf der x-Achse abgebildet, wobei F die empirische Verteilungsfunktion der x_i und Φ die Verteilungsfunktion der Standardnormalverteilung ist. Sind die x_i Realisierungen unabhängiger, identisch normalverteilter Zufallsvariablen, liegen die Punkte in etwa auf einer Geraden.

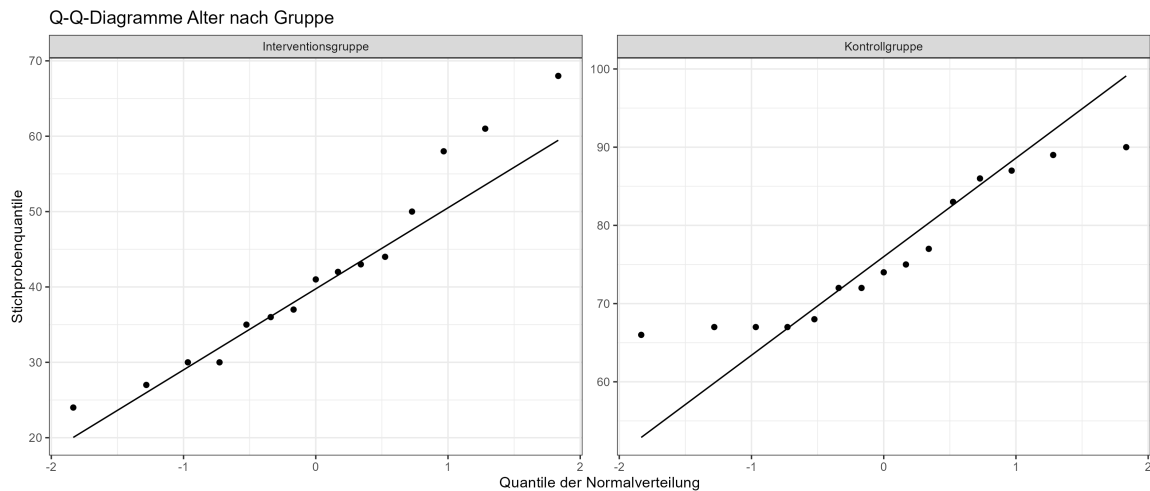


Abbildung 3.3.: Quantil-Quantil-Diagramme Alter nach Gruppe

Insbesondere in der Kontrollgruppe weicht die Anordnung der Punkte systematisch von einer Geraden ab. Bei einer Normalverteilung wäre zu erwarten, dass die Extrempunkte an beiden Enden weiter vom Mittelwert abweichen. Die Diagramme verstärken also den Eindruck, dass in der Kontrollgruppe keine Normalverteilung vorliegt. Wir entscheiden uns daher für den nichtparametrischen Wilcoxon-Rangsummen-Test[6].

3.3.1. Theoretischer Hintergrund

Der Wilcoxon-Rangsummen-Test ist ein nichtparametrischer Test zum Vergleich der Verteilungsfunktionen zweier unabhängiger Stichproben. Seien F_x und F_y Verteilungsfunktionen und $X_i \sim F_x, i \in \{1, \dots, m\}$, sowie $Y_j \sim F_y, j \in \{1, \dots, n\}$ Zufallsvariablen mit Realisierungen $x_1, \dots, x_m$ und $y_1, \dots, y_n$, unseren beiden Stichproben. Der Wilcoxon-Rangsummen-Test prüft die Hypothesen

$$H_0 : F_x = F_y$$

und

$$H_1 : F_x \neq F_y$$

oder äquivalent dazu

$$H_0 : P(X < Y) = P(X > Y)$$

und

$$H_1 : P(X < Y) \neq P(X > Y)$$

mit $X \sim F_x$ und $Y \sim F_y$.

In unserem Fall bedeutet die Nullhypothese, dass die Wahrscheinlichkeit, dass eine zufällig gewählte Person aus der Population, der die Interventionsgruppe entstammt, jünger ist als eine aus der Population der Kontrollgruppe, gleich groß ist wie die umgekehrte Wahrscheinlichkeit.

Wir bezeichnen die Teststatistik mit W und ihre Realisierung mit w . Um w zu berechnen, kombinieren wir zunächst beide Stichproben, ordnen die Werte und berechnen die Ränge. Anschließend berechnen wir

$$w = \sum_{i=1}^m R(x_i),$$

wobei $R(x_i), i \in \{1, \dots, m\}$ den Rang von x_i in der geordneten kombinierten Stichprobe bezeichnet. Wir betrachten nun die Verteilung von W unter der Nullhypothese. Es gibt $\binom{m+n}{m}$ Möglichkeiten, welche Ränge die x_i annehmen. Ist die Nullhypothese wahr, ist jede dieser Möglichkeiten gleichwahrscheinlich. Es gibt genau eine Möglichkeit, die die kleinstmögliche Rangsumme $\sum_{i=1}^m i = \frac{m*(m+1)}{2}$ und eine, die die größtmögliche Rangsumme $\sum_{i=n+1}^{m+n} i = \frac{(m+n)*(m+n+1)}{2} - \frac{m*(m+1)}{2}$ erzeugt. Allgemein ist die Anzahl der möglichen Anordnungen, die eine Rangsumme s erzeugen, die Anzahl der

Partitionen von s der Länge m , deren Elemente in $\{1, \dots, m+n\}$ liegen und paarweise verschieden sind. Die Verteilung von W unter der Nullhypothese ist also berechenbar, für große Stichproben allerdings nur unter großem Aufwand. Deshalb wird die Verteilung von W in diesem Fall häufig durch eine Normalverteilung approximiert[7]. Für unsere Stichproben mit jeweils 15 Teilnehmern können wir die genauen Werte einer Tabelle entnehmen oder mithilfe von Programmen berechnen. Da die Summe der Rangsummen der beiden Gruppen immer identisch ist, ist die Verteilung der Teststatistik symmetrisch und man kommt unabhängig von der betrachteten Gruppe zum selben Ergebnis.

3.3.2. Umsetzung

Zur Berechnung der Rangsumme ordnen wir die kombinierten Stichprobenwerte nach Alter und markieren die jeweilige Gruppenzugehörigkeit (Tabelle 3.4). Nun können wir die Ränge der Interventionsgruppe einfach ablesen und zur Rangsumme aufsummieren. Wir erhalten¹

$$w = 19 + \sum_{i=1}^{14} i = 19 + \frac{14 * 15}{2} = 124.$$

Als unteren kritischen Wert für einen zweiseitigen Test zum Signifikanzniveau 5% mit zwei Stichproben der Größe 15 finden wir 184. Auch hier lässt sich der p-Wert in R berechnen – entweder über `wilcox.test` oder mit `pwilcox(q=4, m=15, n=15)`². Wir erhalten $p \approx 1.547 * 10^{-7} < 0.025$. Die Verteilungen des Alters beider Gruppen unterscheiden sich also statistisch signifikant. Die Randomisierung war in Bezug auf die bekannte Störgröße Alter offenbar nicht erfolgreich. Die Patienten der Kontrollgruppe sind fast alle älter als die Patienten der Interventionsgruppe. Die ohnehin statistisch nicht signifikante, höhere Überlebenszeit der Interventionsgruppe lässt sich also nicht nur auf das Medikament zurückführen. Der Störfaktor

¹Genau genommen könnten wir auch 125 als Rangsumme erhalten, denn zwischen Rang 19 und 20 liegt ein Tie vor, den wir auch anders auflösen könnten. In beiden Fällen liegt unsere Rangsumme aber deutlich unter dem kritischen Wert.

²Da die Wilcoxon-Verteilung in R so definiert ist, dass ihr kleinstmöglicher Wert bei 0 liegt (statt bei $\frac{m(m+1)}{2}$, wie bei unserer Teststatistik)[8], muss der berechnete Rangsummenwert entsprechend angepasst werden. Es ergibt sich ein Eingabewert von $w - \frac{m*(m+1)}{2} = 124 - 120 = 4$.

3. Hauptteil

Rang	Alter	Gruppe	Rang	Alter	Gruppe
1	24	Interventionsgruppe	16	67	Kontrollgruppe
2	27	Interventionsgruppe	17	67	Kontrollgruppe
3	30	Interventionsgruppe	18	67	Kontrollgruppe
4	30	Interventionsgruppe	19	68	Interventionsgruppe
5	35	Interventionsgruppe	20	68	Kontrollgruppe
6	36	Interventionsgruppe	21	72	Kontrollgruppe
7	37	Interventionsgruppe	22	72	Kontrollgruppe
8	41	Interventionsgruppe	23	74	Kontrollgruppe
9	42	Interventionsgruppe	24	75	Kontrollgruppe
10	43	Interventionsgruppe	25	77	Kontrollgruppe
11	44	Interventionsgruppe	26	83	Kontrollgruppe
12	50	Interventionsgruppe	27	86	Kontrollgruppe
13	58	Interventionsgruppe	28	87	Kontrollgruppe
14	61	Interventionsgruppe	29	89	Kontrollgruppe
15	66	Kontrollgruppe	30	90	Kontrollgruppe

Tabelle 3.4.: Rangliste Alter; Interventionsgruppe farblich hervorgehoben.

Alter erschwert die Interpretation der vorliegenden Daten. Ein möglicher Lösungsansatz ist die Durchführung einer Cox-Regression[9]. Diese untersucht den Einfluss metrischer Variablen, wie z.B. Alter, und kategorialer Variablen, wie z.B. Gruppenzugehörigkeit, auf die Hazardfunktion h mit $h(t) = \lim_{\Delta_t \searrow 0} \frac{P(t < T < t + \Delta_t | T > t)}{\Delta_t} = -\frac{S'(t)}{S(t)}$. Bei der Cox-Regression wird die Hazardfunktion modelliert durch $h(t; x) = h_0(t) \exp(x \cdot \beta)$, wobei x ein Vektor von Ausprägungen der Prädiktorvariablen, h_0 eine unbekannte, sogenannte Basis-Hazardfunktion und β ein Vektor von Koeffizienten ist, der basierend auf den erhobenen Daten aus Prädiktoren und zugehörigen Todes- bzw. Zensierungszeitpunkten durch Maximum-Likelihood-Schätzung ermittelt wird. Die zu einem Prädiktor gehörende geschätzte Koeffiziente ist dann ein Maß des Effekts des Prädiktors auf die Hazardfunktion. Mithilfe einer Cox-Regression ließe sich so der Effekt der Gruppenzugehörigkeit unter Berücksichtigung des Störfaktors Alter untersuchen.

Literatur

- [1] E. L. Kaplan und P. Meier, “Nonparametric Estimation from Incomplete Observations,” *Journal of the American Statistical Association*, Jg. 53, Nr. 282, S. 457–81, 1958. DOI: 10.2307/2281868.
- [2] N. Mantel, “Evaluation of survival data and two new rank order statistics arising in its consideration,” *Cancer chemotherapy reports*, Jg. 50, Nr. 3, S. 163–70, 1966, <https://pubmed.ncbi.nlm.nih.gov/5910392/>.
- [3] J. M. Bland und D. G. Altman, “The logrank test,” *BMJ (Clinical research ed.)*, Jg. 328, Nr. 7447, S. 1073, 2004. DOI: 10.1136/bmj.328.7447.1073.
- [4] G. Casella und R. L. Berger, *Statistical inference*. Duxbury Pacific Grove, CA, 2002, Bd. 2.
- [5] D. Schoenfeld, “The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions,” *Biometrika*, Jg. 68, Nr. 1, S. 316–19, 1981. DOI: 10.2307/2335833.
- [6] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, Jg. 1, Nr. 6, S. 80–83, 1945. DOI: 10.2307/3001968.
- [7] C. A. Bellera, M. Julien und J. A. Hanley, “Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What N? Graphical Insights,” *Journal of Statistics Education*, Jg. 18, Nr. 2, S. 1, 2010. DOI: 10.1080/10691898.2010.11889486. eprint: <https://doi.org/10.1080/10691898.2010.11889486>.
- [8] R Core Team, *R: A Language and Environment for Statistical Computing*, <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Wilcoxon.html>, R Foundation for Statistical Computing, Wien, Österreich, 2024.
- [9] D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, Jg. 34, Nr. 2, S. 187–202, 1972. DOI: <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>. eprint: <https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x>.

A. Anhang

Zeit	status	Gruppe	Alter	Zeit	status	Gruppe	Alter
0	tot	IG	58	0	tot	KG	77
1	lebt	IG	41	1	lebt	KG	72
1	lebt	IG	42	1	tot	KG	72
1	tot	IG	30	1	tot	KG	86
1	tot	IG	44	2	lebt	KG	67
2	lebt	IG	61	2	lebt	KG	87
3	lebt	IG	30	2	tot	KG	67
4	tot	IG	24	2	tot	KG	74
4	tot	IG	27	3	tot	KG	75
4	tot	IG	68	3	tot	KG	89
7	tot	IG	43	5	tot	KG	66
9	lebt	IG	37	5	tot	KG	67
11	tot	IG	35	7	lebt	KG	68
13	tot	IG	36	7	tot	KG	90
14	tot	IG	50	11	lebt	KG	83

Tabelle A.1.: Geordnete Rohdaten für beide Gruppen

IG = Interventionsgruppe, KG = Kontrollgruppe