Technische Universität Dortmund
Fakultät Statistik
Wintersemester 22/23

Fallstudien I

Projekt 1: Deskription eines Datensatzes

Prof. Dr. Guido Knapp
M. Sc. Yassine Talleb

Bericht von: Louisa Poggel

Mitglieder der Gruppe 1:
Caroline Baer
Daniel Sipek
Julia Keiter
Louisa Poggel

27.10.2022

Inhaltsverzeichnis

1 Einleitung

Im Rahmen einer multinationalen, multizentrischen, doppelblinden, placebo-kontrollierten Phase III Studie wird die Wriksakeit eines Medikamentes, welches als add-on Therapie zur Standardbehandlung von chronisch kongestiver Herzinsuffizenz (NYHA functional class II-IV) gedacht ist, untersucht. Dabei erfolgt nach einem Screeningphase eine Auswahl von Patienten die in eine Placebo Gruppe und eine aktive Gruppe aufgerteilt werden. Das Ziel dieses Projektes ist es zunächst die Verteilung der demografischen Variablen sowohl im Gesamtdatensatz, als auch in den beiden Medikationsgruppen mithilfe von univariaten Kennzahlen (Lage-, Streuungs-, Schiefe- und Wölbungsmaße) und geeigneten deskriptiven Grafiken (Boxplots, Histogramme, Säulendiagramme und Kerndichteschätzer) zu charakterisieren. Dabei erfolgt zwischen den beiden Medikationsgruppen ein Vergleich der Verteilungen um ergründen zu können, wie erfogreich die randomisierte Aufteilung der Patienten war.
Dazu wird in Kapitel 2 zunächst die Problemstellung näher erläutert, worauf in Kapitel 3 eine detaillierte Beschreibung der verwendeten Methoden folgt. Die statistiche Auswertung, aufgeteilt nach den gescreenten und randomisierten Teilnehmern der Studie, erfolgt in Kapitel 4. Zuletzt erfolgt in Kapitel 5 eine Zusammenfassung aller wichtigen Ergebnisse, sowie das Literaturverzeichnis und der Anhang.
Bei der Analyse der Verteilung der demografischen Variablen wird sich zeigen, dass ausschließlich unimodale Verteilungen vorliegen. Dabei lassen sich die fünf kardinal skalierten Variablen in die Kategorien „nahezu symmetrisch“, „leicht rechtsschief“ und „stark rechtsschief“ einteilen. Auch in Bezug auf die Wölbung gibt es sowohl spitze als auch flache Verteilungen. Beim Vergleich der Variablen zwischen den Medikationsgruppen wird sich herrausstellen, dass sich Unterschiede im Rahmen halten und alleinig mithilfe von deskriptiven Methoden kaum nachweisbar sind. Zu den leichten Unterschieden gehören unter anderem, dass die Variablen in der aktiven Medikationsgruppe tendenziell mehr streuuen und es einen Unterschied in den Häufigkeiten des Geschlechtes von etwa 10% gibt.

2 Problemstellung

Im Folgenden werden die größtenteils demografischen Daten einer multinationalen, multizentrischen, doppelblinden, placebo-kontollierten Phase III Studie zur Prüfung der Wirksamkeit eines Medikamentes untersucht. Das Medikament ist als eine add-on Therapie zur Standardbehandlung bei der Behandlung von älteren Patienten mit chronisch kongestiver Herzinsuffizienz (NYHA functional class II-IV) gedacht.
Dazu ist Datenmatrial in Form des Datensatzes KHK_*Studie*_*Dempgraphie*, bestehend aus 200 Beobachtungen und 15 Variablen, vefügbar. Dieser beinhaltet zunächst Variablen, die aus der Durchführung und Organisation der Studie resultieren. Dazu gehört das *Land*, das *Zentrum*, die *Screeningnummer*, die *Patientennummer* und die *Medikationsgruppe*. Die *Screeningnummer* gibt dabei eine Durchnummerierung aller Patienten an, die an der Screeningphase der Studie teilgenommen haben. Außschließlich geeignete Patienten, die an der Studie teilnehmen sollen, erhalten zudem eine *Patientennummer*. Darauf erfolgt eine Unterteilung der Patienten in zwei *Medikationsgruppen*, welche entweder das Medikament (abgekürzt als „aktiv“) oder ein Placebo (abgekürzt als „placebo“) erhalten. Außerdem wird im folgenden die *Medikationsgruppe* mit MG abgekürzt. Die Variablen *Safety-Analysis Population*, *Intention-To-Treat* und *Per-Protocol-Analysis Population* geben weitere klinisch relavante Informationen.
Im Fokus stehen in diesem Projekt jedoch die demografischen Variablen. Dazu gehören das *Geschlecht*, die *Größe*, das *Gewicht*, das *Alter*, der *Body-Mass-Index*, die *Dauer der bestehenden Herzinsuffizienz* und der *Herzinfarkt*. Die Variablen *Geschlecht* mit den Ausprägungen „männlich“ und „weiblich“ und die Variable *Herzinfarkt* mit den Ausprägungen „ja“ und „nein“ liegen auf einer Nominalskala vor und sind zusätzlich dichotom, da sie nur zwei Ausprägungen haben. Die Bezeichnungen „ja“ und „nein“ der Variable *Herzinfarkt* geben dabei an, ob ein Herzinfarkt vorliegt oder nicht.
Die restlichen demografischen Variablen sind stetig und liegen auf einer Kardinalskala vor. Dabei wird das die Köpergröße bezeinende Merkmal *Größe*in cm und das das Körpergewicht bezeichnende *Gewicht* in kg gemessen. Die zeitlichen Angaben erfolgen bei dem *Alter* in Jahren und bei der *Dauer der bestehenden Herzinsuffizienz* in Monaten. Bei dem *Body-Mass-Index* (BMI) handelt es sich um das Verhältniss aus Körpergröße und Körpergewicht, welcher folgendermaßen definiert ist:

$$BMI := \frac{\text{Gewicht in kg}}{(\text{Körpergröße in m})^2} \tag{1}$$

Werte des BMI werden in verschiedene Gewichtskategorien eingeteilt. Ein BMI zwischen 18.5 und 24.9 steht für ein Normalgewicht. Sollte der BMI kleiner oder größer als dieser Bereich sein, spricht man von Untergewicht bzw. Übergewicht/Adipositas.
Zur Datenqualität lässt sich sagen, dass die Daten des Patienten mit der *Screeningnummer* 2 besonders auffällig waren, da bis auf die *Screeningnummer*, das *Land* und das *Zentrum* nur Nullen oder fehlende Werte eingetragen waren. Dieser vermutete Abbrecher wird aus dem Datensatz entfernt, sodass dieser noch aus 199 Individuen besteht. Ansonsten liegen 35 fehlende Werte in der Variable *Patientennummer* und 8 fehlende Werte beim Merkmal *Dauer der Herzinsuffizienz* vor. Die fehlenden Werte bezüglich der *Patientennummer*, lässt sich durch die nicht in die Studie aufegnommene Personen erklären und sind somit nicht negativ auf die Datenqualität auszuwerten.
Ziel des Projektes wird zunächst sein die Verteilung der demografischen Variablen in der Gesamtheit aller gescreenten Personen zu charakterisieren. Darauf folgt eine Betrachtung aller randomisierten Personen, die in die Studie aufgenommen wurden, um den Erfolg bzw. Misserfolg der Randomisierung zu bewerten. Dazu werden die Verteilungen der demografischen Variablen getrennt nach Medikationsgruppe betrachtet. Zur Charakterisierung der Verteilungen werden jeweils univariate Kenngrößen, wie Lage-, Streuungs-, Schiefe- und Wölbungsmaße, als auch deskriptive grafische Verfahren (Boxplots, Histogramme, Kerndichteschätzer und Säulendiagramme) verwendet.

3 Statistische Methoden

3.1 Deskriptive univariate Kennzahlen

Zur Analyse des Datensatzes werden ausschließlich deskrpitive Methoden in Form von univariaten Kennzahlen für Lage, Streuung, Schiefe und Wölbung und grafischen Verfahren zur Darstellung der Verteilung der Variablen verwendet. Dabei werden im Folgenden die Beobachtungen einer Variable mit $x_1, \ldots, x_n$ bezeichnet. Hierbei bezeichnet n die Anzahl der Beobachtungen einer Variable und es gilt für alle x_i für $i = 1, \ldots, n$, dass $x_i \in \mathbb{R}$. Mit den eingeführten Bezeichnungen lässt sich das arithmetische Mittel definieren, als $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$ (Fahrmeir et al. (2011), S.54). Neben diesem klassischem Lagemaß werden Quantile (Hartung et al. (2009), S.34) verwendet, die in Abhängigkeit des Parameters $p \in (0, 1)$, folgendermaßen definiert sind:

$$Q_p := \begin{cases} x_{(\lceil n \cdot p \rceil)} & n \cdot p \text{ nicht ganzzahlig} \\ \frac{1}{2} \cdot (x_{(n \cdot p)} + x_{((n \cdot p)+1)}) & n \cdot p \text{ ganzzahlig} \end{cases} \tag{2}$$

Dabei bezeichnet der Index in runden Klammern von $x_{(i)}$ den i-ten Wert der aufsteigend georneten Beobachtungen, für die $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ für $i = 1, \ldots, n$ gilt. Somit gilt für das Quantil Q_p, dass ein Anteil p der Daten kleiner oder gleich Q_p ist und ein Anteil von $1 - p$ größer oder gleich Q_p ist. Diese Methode entspricht der in der Software R implementierten `quantile` - Funktion unter Angabe des Arguments `type = 2`. Wichtige Spezialfälle der Quantilsfunktion sind dabei $p = 0.25$ und $p = 0.75$, welche als unteres und oberes Quartil bezeichnet werden. Für den Parameter $p = 0.5$ erhält man den Median, der m folgenden auch mit der Schreibweise $med(x_1, \ldots, x_n) = Q_{0.5}$ bezeichnet wird. Dieser kann als eine robuste Alternative zum arithmetischen Mittel verwendet werden. Robust meint in diesem Fall eine Robustheit gegenüber Aureißern, also einzelnen sehr kleinen oder sehr großen Werten. Der Begriff Ausreißer wird im Kapitel deskriptive grafische Verfahren näher spezifiziert und meint in diesem Bericht Datenpunkte, die im Boxplot als Ausreißer klassifiziert werden.
Vorallem für nominale Variablen ist der Modus (bzw. Modalwert) (Fahrmeir et al. (2011), S.57) ein wichtiges Lagemaß. Für dessen Definition bezeichne zunächst die m verschiedenen Ausprägungen der Beobachtungen $x_1, \ldots, x_n$ mit b_j für $j = 1, \ldots, m$, wobei $m, j \in \mathbb{N}$ ist. Nun werden die absolute und relative Häufigkeit der Ausprägung b_j folgendermaßen

definiert (Fahrmeir et al. (2011), S.32):

$$H_{i,j} := \text{Anzahl der Werte } x_i \text{ mit der Ausprägung } b_j \text{ (absolute Häufigkeit)} \tag{3}$$

$$h_{i,j} := \frac{H_{i,j}}{n} \text{ (relative Häufigkeit)} \tag{4}$$

Der Modus wird nun als die Ausprägung b_j bezeichnet, die die größte absolute Häufigkeit ($H_{i,j}$) und somit auch die größte relative Häufigkeit ($h_{i,j}$) hat.
Um mehr Kenntniss über die Verteilung einer Variable zu erlangen ist auch die Streuung von Interesse. Dazu werden zunächst die empirische Varianz (s^2) und Standardabweichung (s) als klasischen Streuungsmaße, unter Verwendung des Vorfaktors $\frac{1}{n-1}$, verwendet (Hartung et al. (2009), S.44):

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad s := \sqrt{s^2} \tag{5}$$

Um einen ersten Überblick um die Streuung zu gewinnen, werden die Spannweite und der Interquartilsabstand genutzt. Die Spannweite $r := \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n)$ bezeichnet die Spanne des Wertebereiches (Fahrmeir et al. (2011), S.67) der beobachteten Werte. Hingegen gibt der Interquartilsabstand $IQA := Q_{0.75} - Q_{0.25}$ einen Bereich an, in dem 50% der Beobachtungen liegen (Hartung et al. (2009) S.41). Auch bei den Streuungsmaßen wird ein gegen Ausreißer robustes Maß eingestzt, welches auf den robsuten Eigenschaften des Medians beruht. Diese Maß wird als (korrigierte) Mittlere absolute Abweichung vom Median (MAD) bezeichnet. Bei der Definition wird auf die in R implememtierte Version mit dem Vorfaktor 1.4826 zurückgegriffen (Hartung et al. (2009) S.865):

$$mad := 1.4826 \cdot med(|x_i - med(x_1, \ldots, x_n)|) \tag{6}$$

Weitere interessante Merkmale einer Verteilung sind Schiefe und Wölbung. Kennzahlen die diese Merkmale charakterisieren verwenden häufig k-te Momente, welche als $m_k := \sum_{i=1}^{n} (x_i - \bar{x})^k$ definiert werden. Unter Verwedung des dritten Momentes lässt sich der Momentenkoeffizient der Schiefe (Hartung et al. (2009) S.47) $g_1 := \frac{m_3}{s^3}$ definieren. Falls dieser den Wert Null annehmen sollte, spricht man von einer symmetrischen Verteilung. Negative Werte sprechen für eine linksschiefe und positive Werte für eine rechtsschiefe Verteilung. Das Maß für die Wölbung wird als $g_2 := \frac{m_4}{(s^2)^2}$ unter Verwendung des vierten Momentes definiert. Verglichen wird die Wölbung mit (Hartung et al. (2009) S.49 und

`moments` - Paket Komsta und Novomestky (2022)) der einer Normalverteilung, die bei dem Wert 3 vorliegt. Werte die größer als 3 sind sprechen für eine spitzerer Verteilung und Werte kleiner als 3 für eine flachere Verteilung.

3.2 Deskriptive grafische Verfahren

McGill et al. (1978) Tukey (1977) Eine nützliche Darstellung von der Verteilung von mindestens ordinal skalierten Variablen ist der verfeinerte Boxplot. Dort werden auf der y-Achse die Werte der Variablen abgetragen. Die Grafik besteht dann aus einem Kasten, dessen untere Linie das untere Quartil ($Q_{0.25}$) und dessen obere Linie das obere Quartil ($Q_{0,75}$) repräsentieren. Im inneren des Kasten wird eine fette Linie für den Median eingetragen. Zusätzlich gehen vom Kasten Verbindungslinien, parallel zur y-Achse, bis zum „inneren Zaun“ aus. Der „innere Zaun“ besteht aus einem unteren Grenzpunkt $g_u := Q_{0.25} - 1.5 \cdot IQR$ und einem oberen Grenzpunkt $g_o := Q_{0.25} - 1.5 \cdot IQR$. Die Verbindungslinien werden auch Whisker genannat und alle Datenpunkte die außerhalb dieses inneren Zaunes liegen werden als Ausreißer klassifiziert. Da für die Erstellung der Boxplots die in R implementierte Funktion `boxplot` verwendet wird, muss beachtet werden, dass das obere und untere Quartil anders als im obigen Teil definiert sind. Außschließlich bei der Verwendung von Boxplots sind die Quartile also folgendermaßen definiert: blablub

Eine klassische Darstellung der Verteilung von kardinal skalierten, stetigen Merkmalen ist das Histogramm (Fahrmeir et al. (2011), S.40-43). Dazu werden die Beobachtungen $x_1, \ldots, x_n$ einer Variable in s verschiedene Klassen $K_1, \ldots, K_s$ mit $s \in \mathbb{N}$ eingeteilt. Jede Klasse wird durch ein linksoffenes Intervall mit $(k_{a-1}, k_a]$ mit $a = 1, \ldots, s$ begrenzt. Dabei ist die Klassenbreite definiert als $d_a = k_a - k_{a-1}$. Pro Klasse wird im Histogramm ein Balken gezeichnet, dessen Breite der Klassenbreite d_a entspricht. Die Höhe des Balkens berechnet sich aus $\frac{h_{a,j}}{d_a}$, wobei $h_{a,j}$ der realtiven Häufigkeit aller Ausprägungen b_j die in Klasse a liegen entspricht. Dementsprechend wird auf der x-Achse das Merkmal und auf der y-Achse $h_{a,j}$ abgetragen.

Zur Darstellung der Dichte $\hat{f}$ von kardinal skalierten Variablen eignet sich ein Kerndichteschätzer (7) welcher unter der Bandbreite $b \in \mathbb{R}$ definiert ist (Fahrmeir et al. (2011),

S.100-102).

$$\hat{f}(x) := \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big) \tag{7}$$

Dabei wird die Bandbreite b passend gewählt und als Kernfunktion $K(u)$ mit $u \in \mathbb{R}$ wird der Gaußkern (8) verwendet.

$$K(u) := \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}u^2\Big) \tag{8}$$

Zur Darstellung von nominal skalierten Merkmalen wird ein Säulendiagramm genutzt, welches die relativen Häufigkeiten $h_{i,j}$ einer Ausprägung b_j an der Stelle x_i in Form eines horizontalen Rechteckes darstellt (Fahrmeir et al. (2011), S.35). Somit wird das Merkmal auf der x-Achse und die relative Häufigkeit $h_{i,j}$ auf der y-Achse abgetragen.

4 Statistische Auswertung

4.1 Charakterisieung der Verteilung der interessierenden Variablen in der Gesamtheit aller gescreenten Patienten

Zunächst werden die beiden binäeren Variablen *Geschlecht* und *Herzinfarkt* in Form von Häufigkeitstabelle betrachtet. Bei der Variable *Geschlecht* sind deutliche Disbalancen in der Geschlechterverteilung zu erkennen, da etwa 66% der gescreenten Patienten Männer und nur 34% Frauen sind. Somit ist der Modalwert in diesem Fall „männlich“. Eine ähn-

	$H_{i,j}$	$h_{i,j}$
männlich	131	0.66
weiblich	68	0.34

Tabelle 1: Häufigkeitstabelle - *Geschlecht* (n = 199)

	$H_{i,j}$	$h_{i,j}$
ja	72	0.36
nein	127	0.64

Tabelle 2: Häufigkeitstabelle - *Herzinfarkt* (n = 199)

liche Verteilung ist bei der Variable *Herzinfarkt* vorzufinden. Hier gibt es ebenfalls einen eindeutigen Modalwert, welcher „nein“ ist. Denn es geben etwa 64% der Probanden an keinen Herzinfarkt gehabt zu haben. Bei etwa 36%, was einer absoluten Häufigkeit von

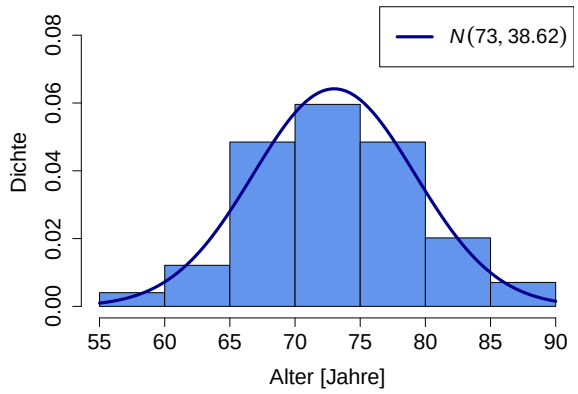

Abbildung 1: Histogramm von *Alter* mit Dichte einer Normalverteilung

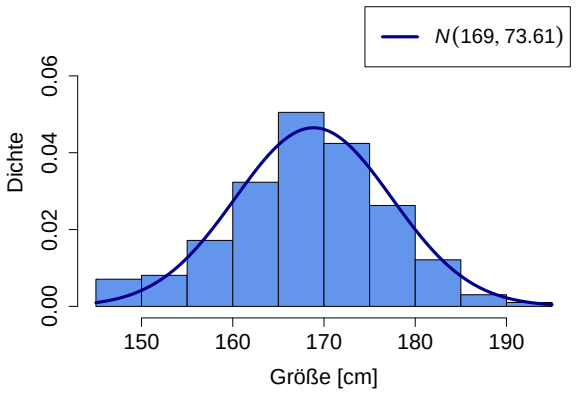

Abbildung 2: Histogramm von *Größe* mit Dichte einer Normalverteilung

72 Probanden entspricht, lag jedoch bereits ein Herzinfarkt vor.

Deutlich umfangreicher ist die Betrachtung der kardinal skalierten Variablen, welche sich auf mehrer univariate Kennzahlen und grafische Methoden stützt. Im Folgenden werden die Variablen in der Reihenfolge von symmetrisch verteilten, über Verteilungen mit leichter bis hin zu deutlich ausgeprägter Schiefe vorgestellt.

Bei den Variablen *Größe* und *Alter* liegt eine nahezu symmetrische, unimodale Verteilung vor, die Anlass dazu gibt eine Normalverteilung zu vermuten. In Abbildung 1 und 2 sind durch diese Vermutung motiviert die Histogramme der beiden Varieblen mit einer passenden Normalverteilungsdichte dargestellt. Diese Dichte erhält dabei als Parameter μ das arithmetische Mittel der Beobachtungen und als σ^2 die empirische Varianz der Beobachtungen. Dabei wird auf zwei Nachkommastellen gerundet. Zur Prüfung des grafischen Eindrucks werden nun die univariaten Kennzahlen der beiden Variablen betrachtet, die den Tabellen 3 und 4 zu entnehmen sind. Auffällig ist dabei, dass bei bei den Variablen die klassichen und robuste Maße kaum voneinander abweichen. Das ungefähre Übereinstimmen des arithmetischen Mittels und des Medians ist dabei ein Anzeichen für Symmetrie (Fahrmeir et al. (2011), S. 60). Bei der Variable *Größe* beträgt der Momentenkoeffizient der Schiefe -0.24, was gegen eine exakt symmetrische Verteilung und für eine leichte linkschiefe Tendez spricht. Hingegen liegt das Wölbungsmaß mit 3.01 sehr nah bei der Wölbung einer Normalverteilung. Bei der Variable *Alter* ist es genau anders herrum. Der Momentenkoeffizient der Schiefe spricht mit 0.05 für eine nahezu exakt symmetrische Verteilung. Hingegen ist die Wölbung mit 2.82 leicht flacher als bei der Normalverteilung.

Bei Betrachtung der Variablen *Gewicht* und *Body-Mass-Index* ist jeweils eine etwas

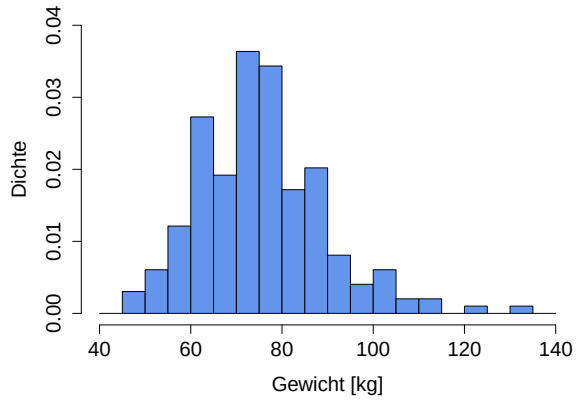

Abbildung 3: Histogramm - *Gewicht*

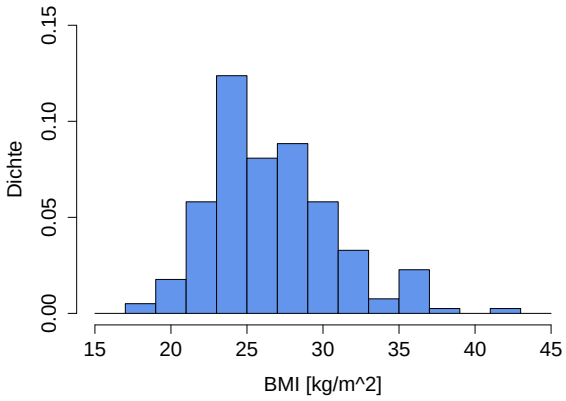

Abbildung 4: Histogramm - *BMI*

größere Abweichung von einer symmetrischen Verteilung zu erkennen. Dies ist zunächst grafisch in Abbildung 3 und 4 zu erkennen. Weiterhin handelt es sich um eine unimodale Verteilung, jedoch liegen mehr Beobachtungen im kleineren Werteberich der angenommen Werte der jeweiligen Variable. Dies deutet auf eine rechtsschiefe Tendenz hin. Um die Verteilung der Variablen weiter zu charakterisieren, werden die univariaten Kennzahlen aus Tabelle 5 und 6 betrachtet. Auch hier liegt kein nennenswerter Unterschied der klassichen und robusten Lage- und Streuungsmaße vor, das deutet drauf hin, dass die rechtsschiefe nur leicht ausgeprägt ist (Fahrmeir et al. (2011), S. 60). Bei der Variable *Gewicht* liegt das arithmetische Mittel bei etwa 76.27 und der Median etwas kleiner bei 75. Bei der Variable *Body-Mass-Index* ist das arithmetische Mittel mit ungefähr 26.69 ebenfalls minimal größer als der Median mit ungefähr 25.95. Der Momentenkoeffizient der Schiefe spricht für Rechtsschiefe und beträgt für die Variable *Gewicht* etwa 0.74 und für die Variable *Body-Mass-Index* etwa 0.71. In Bezug auf die Kurtosis sind beide Verteilungen mit einem Wölbungsmaß von 4.28 (*Gewicht*) und 3.46 (*Body-Mass-Index*) etwas spitzer als die Normalverteilung.

Bei der Variable *Dauer der Herzinsuffizienz* kann man in Abbildung 5 und 6 eine unimodale, spitze, deutlich rechtsschiefe Verteilung erkennen. Außerdem sind im Boxplot einige Beobachtungen als Ausreißer klassifiziert. Bei Betrachtung der univariaten Kennzahlen in Tabelle 7 ist erstmals eine deutliche Abweichung zwischen den robusten und klassischen Lage- und Streuungsmaßen erkennbar. Das arithmetische Mittel ist mit etwa 48.67 knapp doppelt so groß wie der Median, welcher etwa 25.57 beträgt. Ähnliches ist beim Verhältniss von Standardabweichung und MAD ($s = 57.45$, $mad = 31.21$) erkennbar. Vorallem die Differenzen in den Lagemaßen sprechen dabei für eine Rechtsschiefe

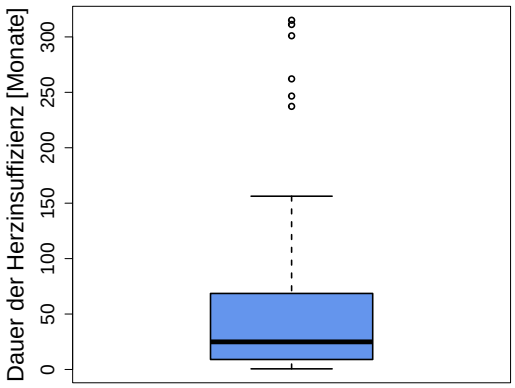

Abbildung 5: Boxplot - *Dauer Herzinsuffizienz*

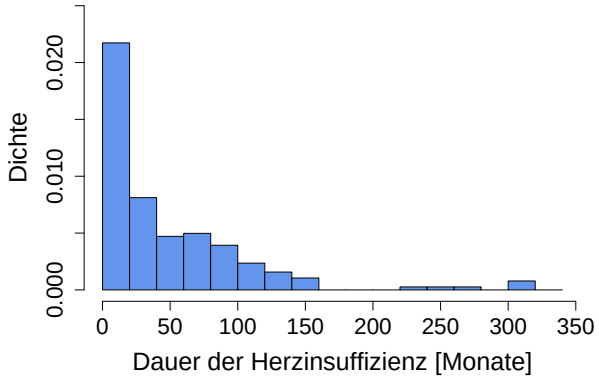

Abbildung 6: Histogramm - *Dauer Herzinsuffizienz*

(Fahrmeir et al. (2011), S. 60). Ein recht großer positiver Wert von etwa 2.30 des Momentenkoeffizienten der Schiefe bekräftigt dies. Das Maß der Kurtosis spricht mit ungefähr 9.61 für eine deutlich spitzere Verteilung als die Normalverteilung.

4.2 Vergleich der Verteilungen der interessierenden Variablen zwischen den Medikationsgruppen

Zur Beurteilung des Erfolges der Randomisierung erfolgt pro Variable zwischen den Medikationsgruppen ein Vergleich der Verteilungen. Falls die Verteilungen größtenteils übereinstimmen, wird die Randomisierung als erfolgreich angesehen.
Wie in Abbildung 7 erkennbar, gibt es in der aktiven MG mit einer relativen Häufigkeit von etwa 0.61 im Vergleich zu 0.7 in der placebo MG etwas weniger Männer. Die Abweichungen der relativen Häufigkeiten bewegen sich also im Rahmen von etwa 10%. Bei der Variable *Herzinfarkt* bewegen sich die Abweichungen im Rahmen von etwa 4%. Denn in der placebo MG hatten, mit etwa 40%, mehr Personen einen Herzinfarkt als in der aktiven MG mit etwa 36%. Die vollständigen Häufigkeitstabellen der nominalen Varaiblen sind im Anahng in den Tabellen 8 bis 11 zu finden.

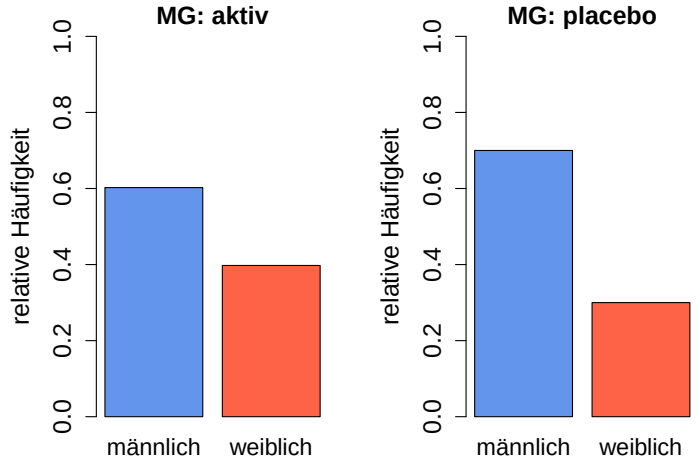

Abbildung 7: Geschlecht - Vergleich in Medikationsgruppen

Bei der Variable *Größe* ist in Abblidung 8 an den längeren Whiskern und der größeren Box zu erkennen, dass die Streuung in der aktiven MG etwas größer ist. In der placebo MG liegt der Großteil der Beobachtungen konzentrierter in der Mitte der Verteilung, sodass einzelne große oder kleine Beobachtungen als Ausreißer klassifiziert werden. Das arithmetische Mittel ist in der aktiven MG mit 167.31 um etwa 3 cm kleiner als in der placebo MG mit 170.32 (vgl. Tabellen 12 und 13).
Auch bei der Variable *Gewicht* ist eine etwas größere Streuung in der aktiven MG erkennbar. Der MAD liegt, wie in den Tabelle 14 und 15 erkennbar, in der aktiven MG

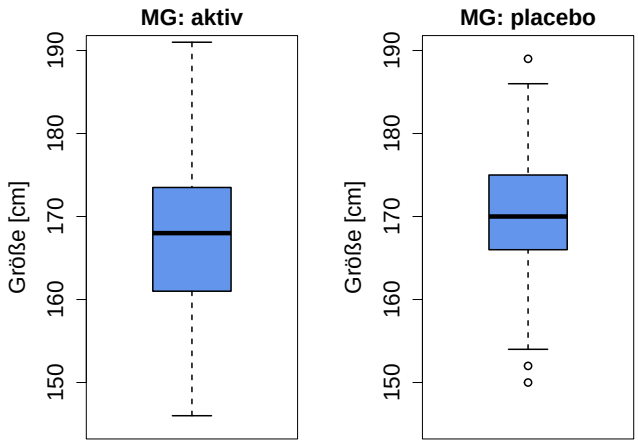

Abbildung 8: Boxplots der Variable Größe getrennt nach MG

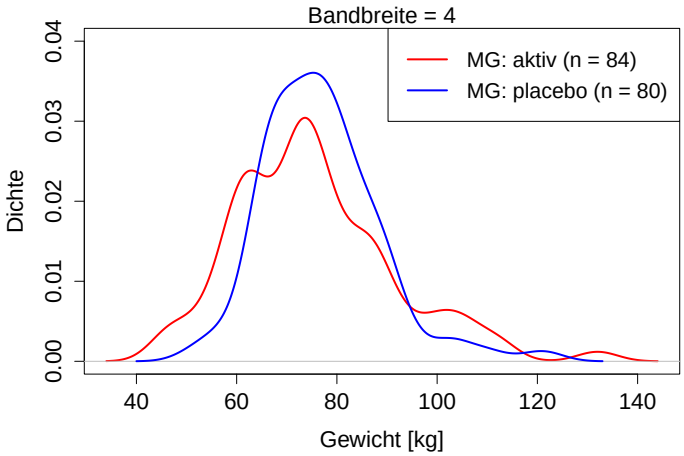

Abbildung 9: Kerndichteschätzer der Variable Gewicht getrennt nach MG

bei etwa 14.83 und in der placebo MG bei etwa 9.64. Zudem ist die Verteilung in der aktiven MG etwas flacher als in der placebo MG. Gemeinsam haben die beide Verteilungen die Rechtsschiefe. Der Momentenkoeffizient der Schiefe beträgt in der aktiven MG etwa 0.82 und in der placebo MG etwa 1.01. Außerdem leigt das arithmetische Mittel der Variable *Gewicht* mit 75.40 in der aktiven MG und mit 76.89 in der placebo MG recht nah beienander (Abweichung von etwa 1.5 kg).

In Abbildung 10 ist zu erkennen, dass bei der Variable *Alter* die Eigenschaft der Symmetrie aus der Grundgesamtheit aller gescreenten Personen nahezu beibehalten wird. Wie in den Tabellen 16 und 17 erkennbar besteht in der placebo MG ist eine leichte Tendenz zur Rechtsschiefe (vgl. $g_1 = 0.27$). Auch das arithmetische Mittel weicht mit

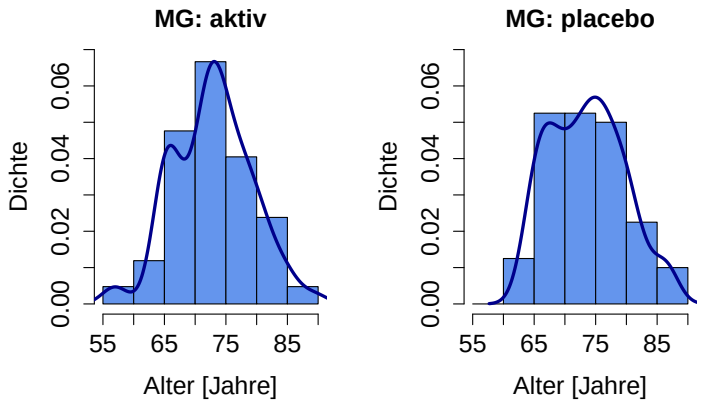

Abbildung 10: Kerndichteschätzer (Bandbreite = 2) und Histogramm der Variable Alter getrennt nach MG

72.85 in der aktiven und 73.55 in der placebo MG kaum voneinander ab (Abweichung von etwa 8 Monaten). Ein Unterschied der Verteilungen ist jedoch in der Kurtosis zu erkennen. Die Verteilung in der placebo MG ist flacher ($g_2 = 2.29$) als in der aktiv MG ($g_2 = 2.98$). Obgleich die Streuung mit robusten oder klassischen Maßen gemessen ist, variiert sie nicht nennenswert.

Bei der Variable *Body-Mass-Index* ist in Abbildung 11 zu erkennen, dass sich die Ver-

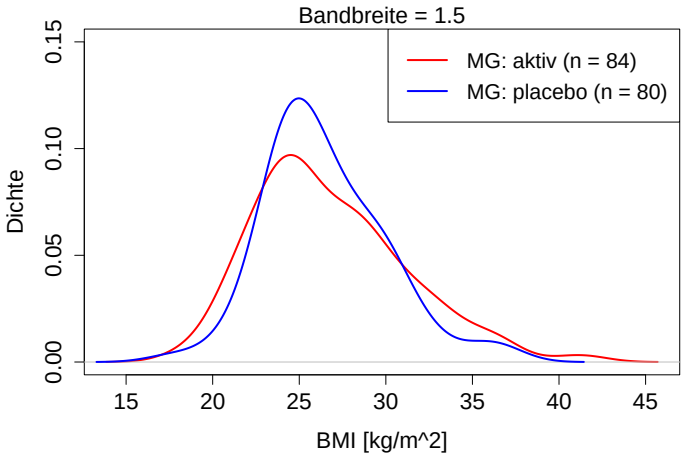

Abbildung 11: Kerndichteschätzer der Variable BMI getrennt nach MG

teilung in den beiden MG nur geringfügig unterscheiden. Den Tabellen 18 bis 19 ist zu entnehmen, dass die Streuung in der aktiven MG (MAD = 3.94, $s = 4.37$) minimal größer als in der placebo MG (MAD = 2.75, $s = 3.47$) ist, was sich vorallem im Bereich eines BMI von 40 bis 45 zeigt. Dabei ist die Verteilung in der aktiven MG ($g_2 = 3.44$) etwas

flacher als in der placebo MG ($g_2 = 3.82$). In beiden Verteilungen ist eine Tendez zur Rechtsschiefe erkennbar, die in der aktiven Gruppe etwas ausgeprägter ist. Außerdem sind beide Verteilungen im Mittel mit einem arithmetischem Mittel von 26.79 (aktiv) und 26.49 (placebo) nahezu gleich (Abweichung von 0.3).

Analog zur Verteilung der Variable *Dauer der Herzinsuffizienz* in der Gesamtheit aller

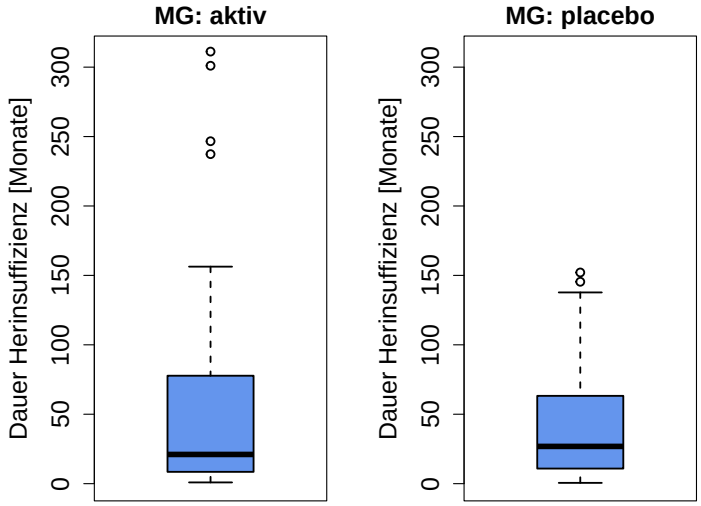

Abbildung 12: Boxplots der Variable Dauer der Herzinsuffizienz getrennt nach MG

gescreenten Patienten, kann man in Abbildung 12 auch in den beiden Medikationsgruppen eine deutlich ausgeprägte Rechtsschiefe erkennen. In Bezug auf die Kurtosis lässt sich sagen, dass die Verteilung in der aktiven MG deutlich spitzer ($g_2 = 7.67$) als in der placebo MG ($g_2 = 3.24$). Zudem ist auffällig, dass in der aktiven MG größere Ausreißer auftreten als in der placebo MG. In Bezug auf die Streuung liefern klassische und robuste Methoden widersprüchliche Ergebnisse. So ist der MAD in der aktiven MG kleiner als ind er placebo MG, bei der Varianz ist dies genau anders herrum (vgl. Tabellen 20 und 21). Der längere obere Whisker und das größere obere Quartil in der aktiven MG, spricht aber dafür, dass die Streuung in der aktiven MG etwas größer ist. Auch weichen die beiden Verteilungen im Mittel, gemessen am Median, um etwa 5 Monate voneinander ab (vgl. aktiv: $Q_{0.5} = 21.03$, placebo: $Q_{0.5} = 26.83$).

5 Zusammenfassung

Mithilfe des Datensatzes KHK_*Studie_Dempgraphie* bestehend aus 199 Beobachtungen und 15 Variablen wird die Frage beantwortet, welche Verteilung die demografischen Variablen in der Gesamtheit aller gescreenten Personen haben. Dabei lassen sich bei den Häufigkeiten von *Geschlecht* (männlich = 66%, weiblich = 34%) und *Herzinfarkt* (ja = 36%, nein = 64%) deutliche Disbalancen erkennen. Durch Verwendung von Boxplots, Histogrammen sowie Lage-, Streuungs-, Schiefe- und Wölbungsmaßen werden die kardinal skalierten Variabeln charakterisiert. Eine der Normalverteilung ähnliche symmetrische, unimodale Verteilung ist bei den Variablen *Alter* und *Gewicht* vorzufinden. Bei den Variablen *Gewicht* und *Body-Mass-Index* stellt sich hingegen herraus, dass eine unimodale, leicht rechtschiefe, spitzere Verteilung vorliegt. Noch ausgeprägter sind die Aspekte der Rechtsschiefe und der spitzeren Wölbung bei der Variable *Dauer der Herzinsuffizienz*.
Um zu beantworten ob die Randomisierung der Patienten in die beiden Medikationsgruppen erfolgreich war, werden die 164 randomisierten Probanden untersucht. Dabei wird die Verteilung der demografischen Variablen zwischen den *Medikationsgruppen* (MG) aktiv (84 Patienten) und placebo (80 Patienten) verglichen. Weitere Methoden sind hier Herndichteschätzer und Balkendiagramme. Die Abweichungen der Häufigkeiten betragen beim *Geschlecht* etwa 10% und be der Variable *Herzinfarkt* etwa 4%. Bei den kardinalen Variablen betragen die Abweichungen des Mittels bei der *Größe* etwa 3cm, beim Gewicht etwa 1.5kg, beim Alter etwa 8 Monate, beim BMI etwa 0.3 $^{kg}/_{m^2}$ und bei der *Dauer der Herzinsuffizienz* etwa 5 Monate. Zur Berechnung des Mittels wird hier, bis auf Ausnahem der *Dauer der Herzinsuffizienz* das arithmetische Mittel verwendet. Ausnahme. Aufrgrund der starken Rechtsschiefe wird beid der *Dauer der Herzinsuffizienz* der, die Mitte besser repräsentierende, Median verwendet. Ingesamt halten sich die Abweichungen im Mittel in einem akzeptablen Rahmen, was für eine ausreichende Randomisierung spricht. Etwas kritischer könnte man die in der aktiven MG tendenziell größere Streuung sehen, welche sich aber auch in Grenzen hält. Auch die Kennzahlen und das Aussehen der Verteilung ist nie exakt gleich, aber es treten auch keine extremen Unterschiede auf. Um das Vorliegen von Unterschieden besser bewerten zu können müsste man statistsiche Tests, beispielsweise den Binomialtest oder Kolmogorow-Smirnow-Test, durchführen. Außerdem wären weitere Hintergrundinformationen nützlich um mögliche Confounder zu finden zu können und beispielsweise die schwere der Auswirkungen von einem Geschlechterunterschied von 10% in den beiden MG einschätzen zu können.

R Core Team (2021) Fahrmeir et al. (2011)

6 Literaturverzeichnis

Literatur

Fahrmeir, L., R. Künstler, I. Pigeot und G. Tutz (2011). *Der Weg zur Datenanalyse. 7. Auflage*. Springer Verlag: München.

Hartung, J., B. Elpelt und K.-H. Klösener (2009). *Statistik Lehr- und Handbuch der angewandten Statistik. 15.Auflage*. Oldenbourg Verlag: München.

Komsta, L. und F. Novomestky (2022). *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.14.1. URL: `https://CRAN.R-project.org/package=moments`.

McGill, R., J. W. Tukey und W. A. Larsen (1978). „Variations of Box Plots“. English. In: *The American Statistician* 32(1), S. 12–16. URL: `http://www.jstor.org/stable/2683468`.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading: MA, USA.

7 Anhang

Lagemaße	
Minimum	146.00
unteres Quartil	164.00
Arithmetische Mittel	168.86
Median	169.00
oberes Quartil	175.00
Maximum	191.00
Streuungsmaße	
Varianz	73.24
Standardabweichung	8.56
Spannweite	45.00
Interquartilsabstand	11.00
MAD	8.90
Schiefe und Wölbung	
Schiefe	-0.24
Kurtosis	3.01

Tabelle 3: Univariate Kennzahlen *Größe*

Lagemaße	
Minimum	56.58
unteres Quartil	68.18
Arithmetische Mittel	73.00
Median	72.86
oberes Quartil	77.53
Maximum	89.60
Streuungsmaße	
Varianz	38.42
Standardabweichung	6.20
Spannweite	33.02
Interquartilsabstand	9.26
MAD	6.92
Schiefe und Wölbung	
Schiefe	0.05
Kurtosis	2.82

Tabelle 4: Univariate Kennzahlen *Alter*

Lagemaße	
Minimum	46.00
unteres Quartil	66.00
Arithmetische Mittel	76.27
Median	75.00
oberes Quartil	85.00
Maximum	132.00
Streuungsmaße	
Varianz	189.16
Standardabweichung	13.75
Spannweite	86.00
Interquartilsabstand	18.00
MAD	13.34
Schiefe und Wölbung	
Schiefe	0.74
Kurtosis	4.28

Tabelle 5: Univariate Kennzahlen *Gewicht*

Lagemaße	
Minimum	17.79
unteres Quartil	23.88
Arithmetische Mittel	26.69
Median	25.95
oberes Quartil	29.05
Maximum	41.20
Streuungsmaße	
Varianz	16.44
Standardabweichung	4.05
Spannweite	23.41
Interquartilsabstand	5.07
MAD	3.63
Schiefe und Wölbung	
Schiefe	0.71
Kurtosis	3.46

Tabelle 6: Univariate Kennzahlen *Body-Mass-Index*

Lagemaße	
Minimum	0.57
unteres Quartil	9.00
Arithmetische Mittel	48.67
Median	25.23
oberes Quartil	69.87
Maximum	315.03
Streuungsmaße	
Varianz	3300.63
Standardabweichung	57.45
Spannweite	314.47
Interquartilsabstand	60.57
MAD	31.21
Schiefe und Wölbung	
Schiefe	2.30
Kurtosis	9.61

Tabelle 7: Univariate Kennzahlen *Dauer der Herzinsuffizienz*

	$H_{i,j}$	$h_{i,j}$
maennlich	51.00	0.61
weiblich	33.00	0.39

Tabelle 8: Häufigkeitstabelle - *Geschlecht* in MG aktiv

	$H_{i,j}$	$h_{i,j}$
maennlich	56.00	0.70
weiblich	24.00	0.30

Tabelle 9: Häufigkeitstabelle - *Geschlecht* in MG placebo

	$H_{i,j}$	$h_{i,j}$
ja	30.00	0.36
nein	54.00	0.64

Tabelle 10: Häufigkeitstabelle - *Herzinfarkt* in MG aktiv

	$H_{i,j}$	$h_{i,j}$
ja	32.00	0.40
nein	48.00	0.60

Tabelle 11: Häufigkeitstabelle - *Herzinfarkt* in MG placebo

Lagemaße	
Minimum	146.00
unteres Quartil	161.00
Arithmetische Mittel	167.31
Median	168.00
oberes Quartil	173.50
Maximum	191.00
Streuungsmaße	
Varianz	95.45
Standardabweichung	9.77
Spannweite	45.00
Interquartilsabstand	12.25
MAD	10.38
Schiefe und Wölbung	
Schiefe	-0.14
Kurtosis	2.57

Tabelle 12: Univariate Kennzahlen - *Größe* in MG aktiv

Lagemaße	
Minimum	150.00
unteres Quartil	166.00
Arithmetische Mittel	170.32
Median	170.00
oberes Quartil	175.00
Maximum	189.00
Streuungsmaße	
Varianz	60.04
Standardabweichung	7.75
Spannweite	39.00
Interquartilsabstand	9.00
MAD	7.41
Schiefe und Wölbung	
Schiefe	-0.26
Kurtosis	3.11

Tabelle 13: Univariate Kennzahlen - *Größe* in MG placebo

Lagemaße	
Minimum	46.00
unteres Quartil	64.00
Arithmetische Mittel	75.40
Median	73.00
oberes Quartil	85.00
Maximum	132.00
Streuungsmaße	
Varianz	257.55
Standardabweichung	16.05
Spannweite	86.00
Interquartilsabstand	21.00
MAD	14.83
Schiefe und Wölbung	
Schiefe	0.82
Kurtosis	3.98

Tabelle 14: Univariate Kennzahlen - *Gewicht* in MG aktiv

Lagemaße	
Minimum	52.00
unteres Quartil	68.50
Arithmetische Mittel	76.89
Median	75.50
oberes Quartil	82.00
Maximum	121.00
Streuungsmaße	
Varianz	134.25
Standardabweichung	11.59
Spannweite	69.00
Interquartilsabstand	13.25
MAD	9.64
Schiefe und Wölbung	
Schiefe	1.01
Kurtosis	5.12

Tabelle 15: Univariate Kennzahlen - *Gewicht* in MG placebo

Lagemaße	
Minimum	56.58
unteres Quartil	67.77
Arithmetische Mittel	72.85
Median	72.64
oberes Quartil	77.08
Maximum	89.60
Streuungsmaße	
Varianz	40.11
Standardabweichung	6.33
Spannweite	33.02
Interquartilsabstand	9.07
MAD	7.02
Schiefe und Wölbung	
Schiefe	0.04
Kurtosis	2.98

Tabelle 16: Univariate Kennzahlen - *Alter* in MG aktiv

Lagemaße	
Minimum	63.67
unteres Quartil	68.21
Arithmetische Mittel	73.55
Median	73.31
oberes Quartil	77.94
Maximum	86.71
Streuungsmaße	
Varianz	36.48
Standardabweichung	6.04
Spannweite	23.04
Interquartilsabstand	9.72
MAD	7.22
Schiefe und Wölbung	
Schiefe	0.27
Kurtosis	2.29

Tabelle 17: Univariate Kennzahlen - *Alter* in MG placebo

Lagemaße	
Minimum	18.36
unteres Quartil	23.67
Arithmetische Mittel	26.79
Median	25.79
oberes Quartil	29.23
Maximum	41.20
Streuungsmaße	
Varianz	19.14
Standardabweichung	4.37
Spannweite	22.84
Interquartilsabstand	5.47
MAD	3.94
Schiefe und Wölbung	
Schiefe	0.75
Kurtosis	3.44

Tabelle 18: Univariate Kennzahlen - *Body-Mass-Index* in MG aktiv

Lagemaße	
Minimum	17.79
unteres Quartil	24.19
Arithmetische Mittel	26.49
Median	25.73
oberes Quartil	28.68
Maximum	36.93
Streuungsmaße	
Varianz	12.01
Standardabweichung	3.47
Spannweite	19.14
Interquartilsabstand	4.47
MAD	2.75
Schiefe und Wölbung	
Schiefe	0.67
Kurtosis	3.82

Tabelle 19: Univariate Kennzahlen - *Body-Mass-Index* in MG placebo

Lagemaße	
Minimum	0.90
unteres Quartil	8.50
Arithmetische Mittel	52.29
Median	21.03
oberes Quartil	77.73
Maximum	311.20
Streuungsmaße	
Varianz	4324.20
Standardabweichung	65.76
Spannweite	310.30
Interquartilsabstand	68.61
MAD	26.32
Schiefe und Wölbung	
Schiefe	2.09
Kurtosis	7.67

Tabelle 20: Univariate Kennzahlen - *Dauer der Herinsuffizienz* in MG aktiv

Lagemaße	
Minimum	0.57
unteres Quartil	10.87
Arithmetische Mittel	43.33
Median	26.83
oberes Quartil	63.23
Maximum	152.10
Streuungsmaße	
Varianz	1577.06
Standardabweichung	39.71
Spannweite	151.53
Interquartilsabstand	52.37
MAD	30.59
Schiefe und Wölbung	
Schiefe	1.06
Kurtosis	3.24

Tabelle 21: Univariate Kennzahlen - *Dauer der Herinsuffizienz* in MG placebo