Technische Universität Dortmund
Fakultät Statistik
Wintersemester 2022/23

Fallstudien I: Projekt 1

Deskriptive Analyse der Demografie einer klinischen Studie

Dozenten:
Prof. Dr. Guido Knapp
Yassine Talleb, M. Sc.

Verfasserin:
Julia Keiter

Gruppe 1:
Caroline Baer
Julia Keiter
Louisa Poggel
Daniel Sipek

27.10.2022

Inhaltsverzeichnis

1 Einleitung

Das Krankheitsbild der chronischen kongestiven Herzinsuffizienz geht mit einer peripheren Stauung der herznahen Venen einher. Diese führt zu einem Rückstau von Blut und so zu einem unzureichenden Blutvolumen pro Pumpvorgang im Herzen, sodass die Stoffwechsel- und Energiebedürfnisse der Organe und Körpergewebe nicht befriedigt werden können (Van Aken (2007), Seite 965).
In der vorliegenden klinischen Studie wird im Rahmen der Phase III ein Medikament auf Wirksamkeit untersucht, das als Begleittherapie zur Standardbehandlung einer chronischen kongestiven Herzinsuffizenz verabreicht werden soll. Nach einem demographischen Screening wird ein Großteil der 200 Studienprobanden *doppelblind* in zwei Medikationsgruppen randomisiert aufgeteilt. Eine doppelblinde Durchführung der Studie bedeutet, dass weder der Patient selbst, noch die verabreichende Person wissen, ob es sich bei der Verabreichung um ein Placebo im Sinne einer Kontrolltherapie oder um das aktive Medikament handelt.
Die deskriptive Untersuchung wird zur Untersuchung der Randomisierung des Gesamtdatensatzes in die Medikationsgruppen durchgeführt. Es wird der Frage nachgegangen, ob die Randomisierung insofern erfolgreich war, dass sich die Verteilungen der interessierenden demographischen Variablen zwischen den Medikationsgruppen nicht stark unterscheiden.
Der vorliegenden Datensatz wird in Kapitel 2 beschrieben und auf die Datenqualität bewertet. Nachdem das Ziel des Projektes festgelegt wird, werden in Kapitel 3 sowohl numerische wie auch graphische statistische Methoden vorgestellt, die in Kapitel 4 zur Deskription des Gesamtdatensatzes (Kapitel 4.1) und anschließend zur Auswertung des randomisierten Datensatzes (Kapitel 4.2) dienen. Dies ermöglicht schließlich die Zusammenfassung in Kapitel 5 und eine kurze Diskussion der Ergebnisse.

2 Problemstellung

2.1 Datenmaterial

Die Daten stammen aus einer multinationalen, multizentrischen, doppelblinden, placebokontrollierten Phase III Studie zum Beweis der Wirksamkeit eines Medikaments in der Behandlung von älteren Patienten mit chronischer kongestiver Herzinsuffizienz (NYHA functional class IIIV) als Begleittherapie zur Standardbehandlung. Im vorliegenden Datensatz *KHK_Studie_Demographie* wird eine Stichprobe aus 200 Patienten aus 26 verschiedenen Zentren ausschließlich in Deutschland im Hinblick auf insgesamt 15 Variablen x_{ij} mit $i = 1, ..., 200$ und $j = 1, ..., 15$ betrachtet. Es handelt sich um eine Teilerhebung einer Beobachtungsstudie.
Für dieses Projekt sind sieben Variablen von Interesse: Die kategorialen Variablen *Ge-*

schlecht mit den Ausprägungen 1 für männlich und 2 für weiblich und *erlittener Herzinfarkt*, im Folgenden als *Infarkt* bezeichnet, mit den Ausprägungen 1 für ja und 2 für nein sind dichotom erhoben. Die metrischen Variablen *Größe* in Zentimetern und *Gewicht* in Kilogramm sind verhältnisskaliert und diskret erhoben. Die metrischen Variablen *Alter* in Jahren, *Body-Mass-Index* in kg/m^2, im Folgenden als *BMI* bezeichnet und *Dauer der bestehenden Herzinsuffizienz* in Monaten, im Folgenden als *Dauer* bezeichnet, sind verhältnisskaliert und stetig erhoben. Das *Alter* bzw. die *Dauer* ergeben sich als Differenz aus dem Datum des Screenings und dem Datum der Geburt bzw. der Erstdiagnose. Der *BMI* wird mit

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

berechnet und gibt den Ernährungsstatus einer Person an. BMI Werte zwischen 18.5 $\frac{kg}{m^2}$ und 24.9 $\frac{kg}{m^2}$ lassen auf ein Normalgewicht schließen, kleinere Werte auf Untergewicht, größere Werte auf Übergewicht (World Health Organization (2010)).
Die Variablen *Land* , *Zentrum* , *Screeningnummer* , *Patientennummer* , *Medikationsgruppe*, im Folgenden als *Gruppe* bezeichnet, *Safety-Analysis Population* , *Intention-To-Treat Population* und *Per-Protocol-Analysis Population* werden lediglich erwähnt, im Folgenden aber nicht weiter betrachtet.
Um die Qualität des Datensatzes einschätzen zu können, werden die Daten auf fehlende Werte mit R Core Team (2022) untersucht. Im Gesamtdatensatz mit $200 \cdot 15 = 3000$ Einträgen befinden sich 86 fehlende Werte (2.87%), die sich auf 42 gescreente Patienten verteilen, was auf eine allgemein gute Datenqualität schließen lässt. 72 der 86 fehlenden Werte lassen sich damit begründen, dass für 36 Patienten im Screeningverfahren die interessierenden Variablen *Geschlecht*, *Größe*, *Gewicht*, *Alter*, *BMI*, *Dauer* und *Infarkt* erhoben werden, diese Patienten im Anschluss aber nicht randomisiert in die Gruppen 1 für aktives und 2 für placebo Medikament eingeteilt werden, sodass bei ihnen begründet in zwei der insgesamt 15 Variablen fehlende Werte generiert werden. Anders sieht es bei dem Patienten mit der *Zentrums-Screeningnummer* Kombination `44 - 2` aus, bei demKo in neun von 15 Variablen fehlende Werte generiert werden. Diese Beobachtung wird als Messfehler betrachtet und aus dem Datensatz entfernt. Die Gesamt-Stichprobe ($N_g = 199$) wird im ersten Teil des vierten Kapitels (4.1) betrachtet und die Verteilung der neun interessierenden Variablen beschrieben. Der Fragestellung einer erfolgreichen Randomisierung der Patienten in die beiden Gruppen wird im zweiten Teil des vierten Kapitels (4.2)nachgegangen. Dafür werden nur die 164 randomisierten Patienten mit einem Eintrag in der Variable *Gruppe* betrachtet. Auch in dieser Teilmenge des Datensatz gibt es fünf Beobachtungen mit fehlenden Werten in der Variable *Dauer* (0.2% der $164 \cdot 15 = 2460$ Einträge). Auch in der randomisierten Stichprobe ($N_r = 199$) lässt sich von einer guten Datenqualität sprechen. Um die Verteilungen der Variable *Dauer* in der aktiv bzw. placebo *Gruppe* sinnvoll beurteilen zu können, werden die fehlenden Werte in der Untersuchung dieser

Variable entfernt.

2.2 Ziele des Projekts

Ziel des Projektes ist es, die Randomisierung der Patienten in *Gruppe* aktiv und placebo durch den Vergleich der interessierenden Variablen in diesen beiden Gruppen zu beurteilen. Sollten deutliche Unterschiede erkennbar sein, könnte dies Anlass sein, einen Wirksamkeitsunterschied zwischen den Gruppen zu erwarten, der nicht auf die Wirksamkeit des Medikaments selbst, sondern auf sogenannte Confounder Variablen (Störgrößen) zurückzuführen ist.

3 Statistische Methoden

3.1 Univariate Kenngrößen

Für die Beschreibung nominaler Variablen ist der **Modalwert** (kurz Modus) ein wichtiges Lagemaß. Der Modus ist die häufigste Ausprägung eines Merkmals, also diejenige, die die größte **absolute Häufigkeit** bzw. **relative Häufigkeit** gegeben durch

$$\text{Absolute Häufigkeit } H_{i,j} = \text{Anzahl der Werte } x_i \text{ mit der Ausprägung } \nu_j \tag{2}$$

$$\text{Relative Häufigkeit } f(K_1) = \frac{H_{i,j}}{n} \tag{3}$$

besitzt (Burkschat (2012), Seite 63).

Die folgenden univariaten Kenngrößen sind für die Anwendung an metrisch skalierten Variablen definiert.
Betrachtet man die geordneten Werte $x_{1j}, ..., x_{nj}$ der Variablen $x_1, ..., x_n$ mit Stichprobenumfang N, dann ist das **p - Quantil** $\tilde{x}_p$ für $p \in (0, 1)$ gegeben durch

$$\tilde{x}_p = \begin{cases} x_{(k)}, & np < k < np + 1, \text{np} \notin \mathbb{N} \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), & k = np, \text{np} \in \mathbb{N} \end{cases} \tag{4}$$

Hierbei sind mindestens $p \cdot 100\%$ aller Beobachtungswerte kleiner oder gleich $\tilde{x}_p$ und mindestens $(1-p) \cdot 100\%$ größer oder gleich $\tilde{x}_p$. Zur Berechnung der Quantile im nächsten Kapitel wird die `R` Funktion `quantile()` verwendet (Burkschat (2012), Seite 122 f.). Diese kennt neun verschiedene Definitionen von Quantilen. Die Definition für `type=2` entspricht der aus Formel 1.
Das 0.5-Quantil $\tilde{x}_{0.5}$ entspricht dem Wert in der Mitte aller geordneten Beobachtungswerte und wird **Median** genannt (Burkschat (2012), Seite 159).
Der **Quartilsabstand** (IQR) ist als die Differenz zwischen dem 0.75-Quantil und dem

0.25-Quantil $Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$ (IQR) definiert(Burkschat (2012), Seite 161). In diesem Intervall liegen also 50 % der Beobachtungen einer Variable.
Ein Lagemaß, das in seiner Ausprägung empfindlich auf **Ausreißer**, das heißt auf (für die Verteilung) ungewöhnlich große oder kleine Werte, und auf Schiefe der Verteilung reagiert, ist das **arithmetischen Mittel**, das gegeben ist durch

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5}$$

Es ist sachlogisch definiert als der Durchschnittswert aller Beobachtungen (Assenmacher (2013), Seite 70).
Die angesprochene **Schiefe** einer Verteilung lässt sich mit dem empirischen Schiefekoeffizient

$$s_k = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3} \tag{6}$$

berechnen (Becker (2022)). Eine Verteilung heißt rechtsschief fallst $s_k > 0$, linksschief wenn $s_k < 0$ und symmetrisch falls $s_k = 0$ ist.
In Formel 6 steht s für die Standardabweichung. Die Standardabweichung ergibt sich als Wurzel der **Varianz**, die gegeben ist durch

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{7}$$

Die Varianz ist die durchschnittliche quadratische Abweichung der Beobachtungen von $\bar{x}$. Wegen des Quadrierens besitzt sie jedoch eine andere Dimension als das betrachtete Merkmal. Um diesen Umstand zu beseitigen, wird die positive Wurzel aus der Varianz gezogen und so die **Standardabweichung** $s = \sqrt{s^2}$ gewonnen, die die selbe Dimension wie das betrachtete Merkmal besitzt ((Hartung (2009), Seite 44 ff.).
Da für die Berechnung der Standardabweichung das arithmetischen Mittel verwendet wird, reagiert diese Kenngröße ebenfalls empfindlich auf Ausreißer und Asymmetrie. Als unempfindlichere (oder robustere) Methode wird der **Median der absoluten Abweichungen vom Median (MAD)** betrachtet, der gegeben ist durch

$$mad := med(\mid x_i - \tilde{x}_{0.5} \mid) \tag{8}$$

Der *korrigierte* MAD ist eine Skalierung der in 5 angegebenen Definition mit 1.4826 (Hartung (2009), Seite 865). So berechnet auch `R` mit der Funktion `mad()` in den MAD.
Eine weitere Verteilungscharakterisierung ist die **Wölbung** der Verteilung. Die Wölbung gibt an, wie gleichmäßig die Beobachtungen um den Median liegen. Der empirische Wölbungkoeffizient gegeben durch

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} \tag{9}$$

ist = 3 falls die Beobachtungen im Sinne der Normalverteilung (ausgewogen oder mesokurtisch) um den Median gelegen sind, $\in [1, 3)$ falls die Beobachtungen eher gleichmäßig (flach, platykurtischen) zwischen dem kleinsten Wert, dem **Minimum**, und dem größten Wert, dem **Maximum** der Stichprobe, verteilt sind und > 3 falls die Beobachtungen dicht (spitz, leptokurtisch) um den Median verteilt sind (Becker (2022)).
Schiefe und Wölbung von Verteilungen lassen sich in `R` mit dem Paket `moments` und den Funktionen `skewness()` für die Schiefe und `kurtosis()` für die Wölbung (Komsta und Novomestky (2022)) nach den in Formel 6 und 9 gegebenen Definitionen berechnen. Die **Spannweite** ist die Differenz aus Maximum und Minimum (Hartung (2009), Seite 194).

3.2 Deskriptive Graphiken

Ein **Säulendiagramm** (englisch Barplot) ist eine einfache grafische Methode, um die Häufigkeiten von nominal skalierten Beobachtungswerte in einem Datensatz darzustellen. Hierzu werden auf der x-Achse des Koordinatensystems die verschiedenen Merkmalsausprägungen im Datensatz abgetragen und auf der y-Achse werden die absoluten bzw. die relativen Häufigkeiten angegeben. Über jeder Merkmalsausprägung auf der horizontalen Achse werden die entsprechenden Häufigkeiten in Form von Säulen dargestellt. Die Breite aller Säulen wird gleich gewählt, daher sind die einzelnen Häufigkeiten zusätzlich proportional zu den Flächen der zugehörigen Säulen. Werden die absoluten Häufigkeiten abgetragen, so ergeben die Höhen der einzelnen Säulen die absolute Häufigkeit der Merkmalsausprägung. Werden die relativen Häufigkeiten abgetragen, so ist zu beachten, dass diese sich auf die jeweilige Merkmalsausprägung und nicht auf den gesamten Datensatz bezieht. Hierbei handelt es sich um bedingte relative Häufigkeiten (Burkschat (2012), Seite 37 ff.).
Metrische Daten mit sehr vielen Beobachtungswerten lassen sich übersichtlich darstellen, indem sie (unter Inkaufnahme eines gewissen Informationsverlustes) in Klassen zusammengefasst werden.
Für die Klassifikation der Merkmalsausprägungen mindestens ordinal skalierter Daten in Intervalle, kann ein **Histogramm** als grafisches Hilfsmittel hinzugezogen werden. Auf der x-Achse eines Histogramms befinden sich die Klassengrenzen $\nu_0, ..., \nu_M$ der Intervalle, die y-Achse zeigt die absolute bzw. relative Häufigkeit der Beobachtungen in der jeweiligen Klasse an. Dabei ist zu beachten, dass das erste und das letzte Intervall keine offenen Klassen sein dürfen. Die Klassenintervalle für m=1,...,M Klassen werden definiert als

$$K_1 = [\nu_0, \nu_1], K_2 = [\nu_1, \nu_2], ..., K_M = [\nu_{M-1}, \nu_M],$$

wobei $b_1 = \nu_1 - \nu_0, ..., b_M = \nu_M - \nu_{M-1}$ die jeweilige Klassenbreite und $f(K_1), ..., f(K_M)$ die relative Häufigkeit der Klasse darstellt.
Zwischen den Klassengrenzen, also in jedem Intervall K_m, wird ein Kasten gezeichnet. Die Breite des Kastens entspricht der Länge des Intervalls, also der Klassenbreite b_m, die Höhe h_m wird als Quotient $f(K_m)/b_m$ der relativen Klassenhäufigkeit und der Klassenbreite berechnet (Burkschat (2012), Seite 138 ff.). Daraus ergibt sich, dass die Fläche des Kastens die relative Häufigkeit der Klasse ist (Hartung (2009), Seite 22). In diesem Bericht werden Histogramme mit einer passenden Normalverteilungskurve dargestellt. Die Dichte dieser angepassten Normalverteilung erhält dabei als Parameter μ das arithmetische Mittel der Beobachtungen und als σ^2 die empirische Varianz der Beobachtungen einer Variable. Dabei wird auf zwei Nachkommastellen gerundet.
Boxplots eignen sich besonders zum visuellen Vergleich von Lage und Streuung zweier Variablen. In Abhängigkeit einer Achse, welche die Skala der Daten angibt, werden ein Kasten („box“) und zwei Linien („whiskers“) dargestellt. Dabei wird der Kasten unten durch das 0.25-Quantil und oben durch das 0.75-Quantil begrenzt. Innerhalb des Kastens wird zusätzlich der Median des Datensatzes durch einen Querstrich markiert. Die zwei Linien werden jeweils durch den kleinsten und größten Beobachtungswert des Intervalls $[\tilde{x}_{0.25} - 1.5 \cdot Q, \tilde{x}_{0.75} + 1.5 \cdot Q]$ begrenzt, wobei Q dem Quartilsabstand entspricht. Alle Punkte, die außerhalb dieses Intervalls liegen heißen Außenpunkte und werden mit $\circ$ gekennzeichnet(Burkschat (2012), Seite 105 ff.). In diesem Bericht stimmt die Definition von Ausreißern mit der gerade genannten überein.
In der `R` Funktion `boxplot()` wird mit der von Tukey (1978) vorgestellten Definition von Quantilen gearbeitet.
Die **empirische Verteilungsfunktion** gibt für mindestens ordinale skalierte Daten die Folge der Summenhäufigkeiten S_m mit $m = 1, ..., M$ Merkmalsklassen an. Die Summenhäufigkeiten sind definiert als

$$S_m = \begin{cases} 0 & \forall x < \nu_1 \\ \sum_{m=1}^{M} h_m & x \in [\nu_m, \nu_{m+1}] \\ 1 & \forall x \geq \nu_M \end{cases} \tag{10}$$

In der grafischen Darstellung der empirischen Verteilungsfunktion wird die Summenhäufigkeitsfunktion zwischen den Klassengrenzen konstant gesetzt, sodass sich eine stufenförmige Funktion ergibt. Auf der y-Achse des Grafen der empirischen Verteilungsfunktion wird der Wert der Summenhäufigkeitsfunktion abgetragen, auf der x-Achse die verschiedenen Merkmalsausprägungen.
Bei der „unklassierten“ empirischen Verteilungsfunktion gibt es so viele Klassen wie Merkmalsausprägungen, in der grafischen Darstellung werden also die geordneten Beobachtungen kumuliert dargestellt (Hartung (2009), Seite 23).

4 Statistische Auswertung

4.1 Gesamtdatensatz

Um die Verteilungen der Variablen im randomisierten Datensatz besser einordnen zu können, wird zunächst der Gesamtdatensatz mit $N_g = 199$ Patienten betrachtet.
Bei der Betrachtung von Tabelle 1 (siehe Anhang, S. 14) fällt der große Unterschied der relativen Anteile von männlichen und weiblichen Patienten (0.66 vs. 0.34) auf.
Auch die Verteilung der Variable *Infarkt* (Tabelle 2) (siehe Anhang, Seite 14) entspricht fast einer 2:1 Aufteilung: Gut ²/₃ aller Patienten, also 127 Patienten, haben noch keinen Herzinfarkt erlitten, ¹/₃ aller Patienten, also 72, schon.
Die Aufteilung des Gesamtdatensatzes in die *Gruppen* aktiv und placebo erfolgte relativ ausgeglichen, wie in Tabelle 3 (siehe Anhang, Seite 14) zu erkennen ist: 84 der 164 randomisierten Patienten befinden sich in der aktiv, 80 in der placebo *Gruppe*.
Tabelle 4 (siehe Anhang, Seite 15) zeigt eine Auswahl von 12 univariaten Kenngrößen zur Beschreibung der metrischen Variablen im Gesamtdatensatz.
Die Patienten sind zwischen 146cm und 191cm groß. Die Verteilung der *Größe* ist leicht linksschief, was durch den Schiefekoeffizient von -0.24 gezeigt wird. Mit einem Koeffizienten von 3.01 entspricht die Wölbung der Größenverteilung nahezu der einer Normalverteilung. Diese Ausprägung lässt auf eine ausreißerarme Verteilung der *Größe* schließen
Ähnlich ausgeglichen ist die Variable *Gewicht* verteilt. 50% der Patient sind zwischen 66.5 kg (0.25-Quantil) und 84.5 kg (0.75-Quantil) schwer. Dass es in dieser Variable wenig Verzerrung durch extreme Werte gibt, wird durch die Ähnlichkeit des arithmetischen Mittels (76.27 kg) und des Medians (75 kg) bzw. der Standardabweichung (13.75 kg) und des MADs (13.34 kg) deutlich (Fahrmeir (2016), Seite 60).
Da sich der *BMI* wie in Formel 1 beschrieben aus der *Größe* und dem *Gewicht* einer Person ergibt, ist folgerichtig auch in dieser Variable eine ausgewogene Verteilung zu beobachten. 50 % der Patienten haben einen BMI zwischen 23.88 $^{kg}/_{m^2}$ und 28.95 $^{kg}/_{m^2}$, die Verteilung des BMIs ist leicht rechtsschief (Schiefe 0.71) und spitz (Wölbung 3.46).
Wie in Abbildung 1 ersichtlich, ist die Verteilung der Variable *Alter* ausgewogen.
Obwohl das Minimum von 56.58 Jahren vom Intervall $[68.21, 77.46$in Jahren$]$, in dem sich 50 % der Patient befinden, abweicht, ist dies wohl eine Ausnahme, es lässt sich nämlich keine große Auswirkung auf die ausreißerempfindlichen Kenngrößen arithmetisches Mittel (73 Jahre) und Standardabweichung (6.2 Jahre) erkennen, die sehr ähnlich zu den entsprechend unempfindlicheren Vergleichsgrößen Median (72.86 Jahre) und MAD (6.92 Jahre) erkennen. Der Schiefekoeffizient von 0.05 und die Wölbung von 2.82 zeigen, dass es sich um eine symmetrische und ausgeglichene Verteilung handelt.

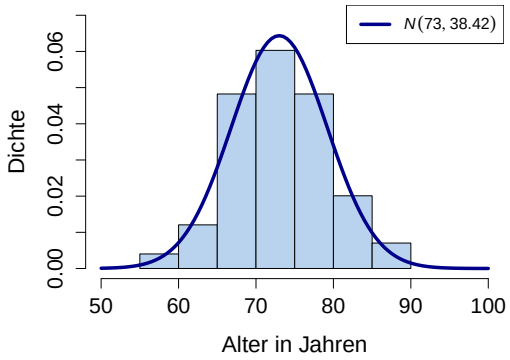

Abbildung 1: Histogramm für Variable *Alter* im Gesamtdatensatz

Dagegen ist die Verteilung der Variable *Dauer* deutlich durch das Vorhandensein von extremen Werten geprägt, was an den deulichen Unterschieden zwischen dem arithmetischen Mittel (48.67 Monate) und dem Median (25.23 Monate) bzw. der Standardabweichung (57.45 Monate) und dem MAD (31.21 Monate) numerisch zu erkennen ist. Während 50 % der Patienten eine *Dauer* zwischen 9.03 Monaten und 69.6 Monaten haben, gibt es andere Patienten, deren *Dauer* nur bei 0.57 Monaten (Minimum) oder bei 315.47 Monaten (Maximum) liegen. Dass die Verteilung der *Dauer* unter den Patienten flach und deutlich rechtsschief ist, zeigen nicht nur die Kenngrößen für Schiefe (2.3) und Wölbung (9.61), sondern auch ein Blick auf Abbildung 2, in der die Dauerverteilung in einem Histogramm abgebildet ist.

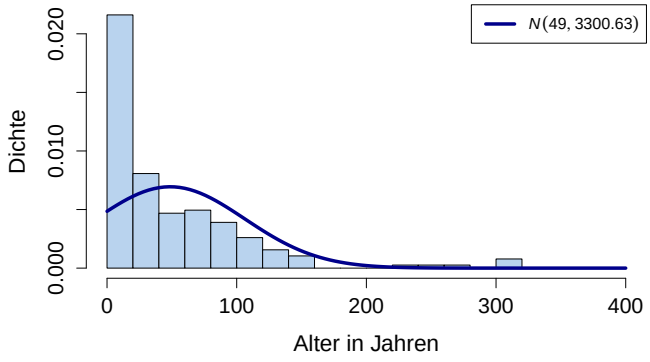

Abbildung 2: Histogramm für Variable *Dauer* im Gesamtdatensatz

4.2 Randomisierter Datensatz

Die Fragestellung, inwieweit die Randomisierung der in die Studie aufgenommenen Patienten erfolgreich verlaufen ist, lässt sich mit einem Vergleich der 164 randomisierten Probanden zwischen den *Gruppen* aktiv und placebo für jede der interessierenden Variablen beantworten.

In Tabelle 1 und Tabelle 2 sind Kenngrößen zu den beiden nominal skalierten interessierenden Variablen *Geschlecht* und *Infarkt* dargestellt. Dass mehr Männer als Frauen an

der Studie teilnehmen, ist in beiden Gruppen festzustellen und ist auf die Geschlechterverteilung im Gesamtdatensatz zurückzuführen. Jedoch unterscheiden sich die relativen Anteile der Geschlechter zwischen den Gruppen relativ stark um neun Prozentpunkte. In der aktiv *Gruppe* befinden sich 51 Männer und 33 Frauen, in der placebo *Gruppe* 56 Männer und 24 Frauen. Das Überwiegen von männlichen Patienten gegenüber weiblichen Patientinnen wurde im Vergleich zum Gesamtdatensatz mit 66 % in der placebo *Gruppe* noch verstärkt (70%).
Hingegen ist das Verhältnis von Patienten, die bereits einen *Infarkt* erlitten haben in beiden Gruppen vergleichbar mit der Stichprobe im Gesamtdatensatz (vgl. auch Abbildung 3). Die relativen Anteile der verschiedenen Merkmalsausprägungen unterscheiden sich in dieser Variable um vier Prozentpunkte, die meisten Patienten haben wie auch im Gesamtdatensatz noch keinen Herzinfarkt erlitten (Modalwert nein).

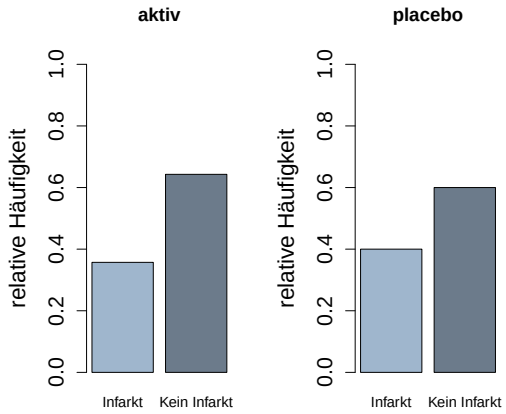

Abbildung 3: Säulendiagramm für Variable *Infarkt* im randomisierten Datensatz getrennt nach *Gruppe*

Tabelle 5 (siehe Anhang, Seite 15) zeigt die gleiche Auswahl an univariaten Kenngrößen für die metrischen interessierenden Variablen jeweils in der *Gruppe* aktiv und placebo wie für den Gesamtdatensatz.
Die *Größe* ist in beiden Gruppen relativ ähnlich ausgeprägt. Mit Blick auf den Interquartilsabstand (12.25 cm (aktiv) vs. 9 cm (placebo)) ist zu sehen, dass die *Größe* von 50 % der Patienten in beiden Gruppen in ähnlichen Intervallen liegt. Die Größenverteilungen der Patienten sind in beiden Gruppen leicht linksschief (-0.14 (aktiv) vs. -0.26(placebo)), was sich mit der Linksschiefe der Größenverteilung im Gesamtdatensatz deckt.
Wie in Abbildung 4 zu sehen ist, unterscheidet sich die Wölbung der Größenverteilungen in den beiden Gruppen sichtlich, numerisch jedoch nur um 0.54 Einheiten, beide Wölbungskoeffizienten sind nahe dem Optimum von 3 (2.57 (aktiv) vs. 3.11 (placebo)) (vgl. 3.1).
Wie auch im Gesamtdatensatz, streuen die Beobachtungen in der Variable *Gewicht* in beiden *Gruppen*. Dabei unterscheiden sich die MADs um 5.19 kg. Auch die Spannweiten der jeweiligen *Gruppen*unterscheiden sich um 17 kg. Insgesamt ist jedoch festzuhalten, dass die Lagemaße Median (73 kg (aktiv) vs. 75.5 kg(placebo)), 0.25-Quantil (64 kg (aktiv) vs. 68.75 kg(placebo)) bzw. 0.75-Quantil (85 kg (aktiv) vs. 82kg (placebo)) der jeweiligen

Gruppe nicht weit auseinander liegen.

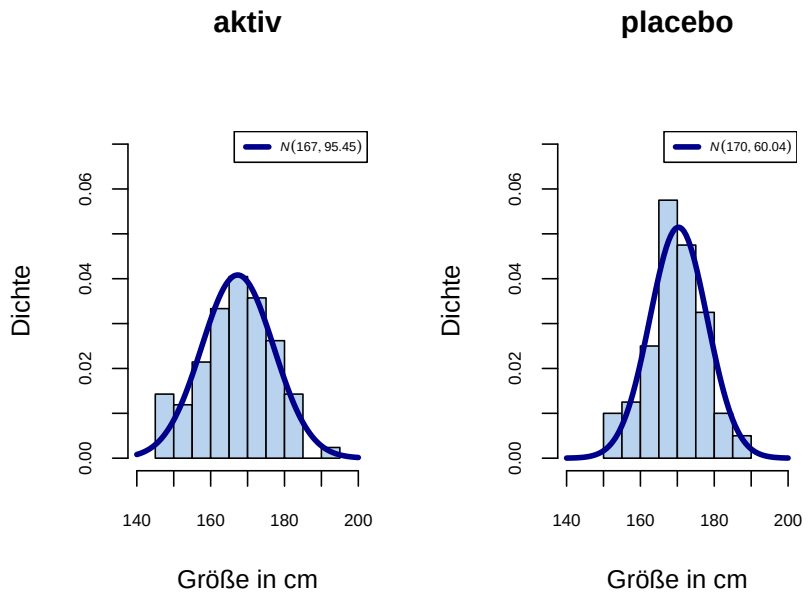

Abbildung 4: Histogramme für Variable *Größe* im randomisierten Datensatz getrennt nach *Gruppe*

Der Vergleich der Kenngrößen für die Variable *BMI* zwischen den *Gruppen* fällt homogen aus. In beiden Gruppen gibt es Patienten mit einem BMI, der deutlich die Mediane (25.79 $^{kg}/_{m^2}$ (aktiv) vs. 25.73$^{kg}/_{m^2}$(placebo)) übersteigt (Maxima: 41.2 $^{kg}/_{m^2}$ (aktiv) vs. 36.93$^{kg}/_{m^2}$(placebo)). Dass diese aber in beiden Gruppen wenig weitreichenderen Einfluss haben, sieht man an der Ähnlichkeit der Standardabweichungen (4.37 $^{kg}/_{m^2}$(aktiv) vs. 3.47$^{kg}/_{m^2}$(placebo)) zu den MADS (3.94 $^{kg}/_{m^2}$(aktiv) vs. 2.75$^{kg}/_{m^2}$(placebo)) und grafisch in Abbildung 5.

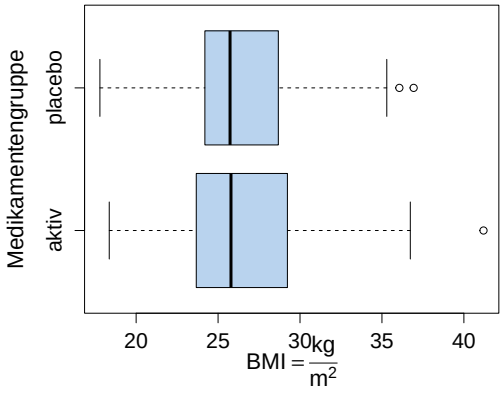

Abbildung 5: Boxplots für BMI aufgeteilt im randomisierten Datensatz getrennt nach *Gruppe*

Man erkennt, dass in beiden Gruppen die Länge der Whisker vergleichbar gleich lang sind und sich die Größen der Boxen nur leicht unterscheiden (IQR: 5.47 $^{kg}/_{m^2}$ (aktiv) vs. 4.47 $^{kg}/_{m^2}$ (placebo)).

Das *Alter* war im Gesamtdatensatz die am meisten ausgeglichene Variable und ist auch in den beiden *Gruppen* vergleichbar homogen verteilt. Das *Alter* der jüngsten Patienten unterscheidet sich zwar um 7.09 Jahre, die Interquartilsabstände beider Gruppen liegen jedoch dicht beieinander (9.07 (aktiv) vs. 9.72 (placebo) in Jahre), was auf eine ähnliche Verteilung von 50% der Patienten schließen lässt. Im Mittel sind die Patienten in der aktiv *Gruppe* 72.85 Jahre und in der placebo *Gruppe* 73.55 Jahre alt.
Wieder ist die Variable *Dauer* die Variable mit den gravierendsten Unterschieden in den Beobachtungen, auch im Vergleich zwischen den *Gruppen*. Während sich die Minima der Gruppen (0.9 (aktiv) vs. 0.57 (placebo) in Monate) nicht allzu stark unterscheiden, ist das Maximum mit 311.2 Monaten *Dauer* in der aktiv *Gruppe* im Vergleich zur placebo *Gruppe* (152.1 Monate) Grund, warum die übrigen Kenngrößen teils sehr unterschiedlich ausfallen. Die ausreißerempfindlichen Kenngrößen arithmetisches Mittel (52.29 (aktiv) vs. 43.33 (placebo) in Monate) und Standardabweichung (65.76 (aktiv) vs. 39.71 (placebo) in Monate) zeigen, dass sich die Wölbung der Verteilungen stark unterscheiden. Während die Dauerverteilung in der placebo *Gruppe* nahezu mesokurtisch ausfällt (Wölbung 3.24), ist die Wölbung in der aktiv *Gruppe* mehr als doppelt so hoch (7.67), was auf eine deutlich flache Dauerverteilung in dieser Gruppe schließen lässt.
Abgesehen von dem Ausreißer und dessen numerischen Auswirkungen sieht man jedoch im Boxplot in Abbildung 6, dass sich die Mediane, die Interquartilsabstände und sogar das Symmetrieverhalten nicht wesentlich zwischen den Gruppen unterscheiden.

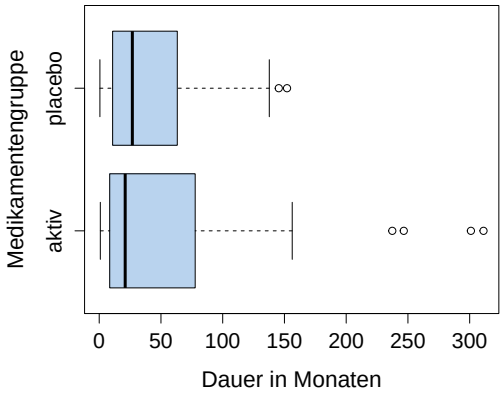

Abbildung 6: Boxplots für *BMI* im randomisierten Datensatz getrennt nach *Gruppe*

Hier ist also trotz der großen Unterschiede in fünf univariaten Kenngrößen in Tabelle 5 nicht von einer eindeutig gescheiterten Randomisierung auszugehen. Um diesen Beobachtungen noch weiter nachzugehen, sind in Abbildung 7 die empirischen Verteilungsfunktionen der beiden Gruppen für die Variable *Dauer* abgezeichnet. Man erkennt, dass die Verteilungen der Beobachtungen bis ca. 75 Monate relativ gleich ansteigen. Danach sinkt der Steigung der empirischen Verteilungsfunktion der *Dauer* in der aktiv *Gruppe* ab, während in der placebo *Gruppe* die Verteilungsfunktion weiter stark ansteigt und so früher die 1 erreicht. Eine vollends erfolgreiche Randomisierung in die beiden Gruppen

aktiv und placebo wurde bei der Variable *Dauer* also nicht erreicht.

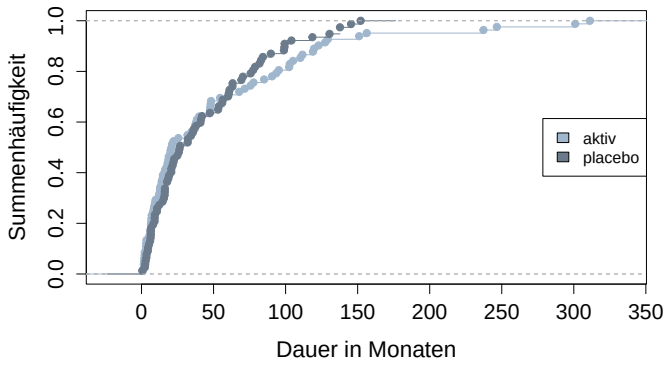

Abbildung 7: empirische Verteilungsfunktionen der *Dauer* im randomisierten Datensatz getrennt nach *Gruppe*

5 Zusammenfassung und Diskussion

Die deskriptive Auswertung des Datensatzes *KHK_Studie_Demographie* der interessierenden Variabeln, die im Rahmen einer Phase III Studie zum Beweis der Wirksamkeit eines Medikaments in der Behandlung mit chronischer kongestiver Herzinsuffizienz an 200 älteren Patienten erhoben wurden, hat gezeigt, dass die Verteilungen der Variablen in den randomisierten Medikationsgruppen aktiv und placebo größtenteils homogen erfolgt ist. Die Variablen *Größe*, *Gewicht*, *Body-Maß-Index* und *Alter* unterscheiden sich zwischen den Gruppen nur leicht hinsichtlich der Streuung. Die Variable *erlittener Herzinfarkt* hebt sich in der placebo *Gruppe* insofern hervor, dass der Anteil der Patienten, die bereits einen Herzinfarkt erlitten haben um vier Prozent höher ist als in der aktiv *Gruppe*. Jedoch ist darauf hinzuweisen, dass sich in der placebo *Gruppe* im Vergleich zur aktiv *Gruppe* zwei Patienten mehr befinden, sodass die Beaobachtung relativiert wird. Dass sich die relativen Anteile der *Geschlechter* zwischen den *Gruppen* um 9 Prozent unterscheiden, ist kritischer zu betrachten und bietet Anlass zur Diskussion über den Einfluss des Geschlechts als mögliche Störgrößenvariable auf die Wirksamkeit des Medikaments. Die deutlichsten Unterschiede zwischen den Medikationsgruppen ließen sich in der Variable *Dauer der Herzinsuffizienz* beobachten. In der aktiv *Gruppe*, befinden sich vier Patienten die eine Dauer von über 200 Monaten aufweisen, während das Maximum in der placebo *Gruppe* 152 Monate beträgt. Diese Disbalance hätte leicht durch eine sorgfältigere Randomisierung der Patienten mit extremen Werten in dieser Variable vermieden werden können. Trotz dieser Kritik ist anzumerken, dass die weniger empfindlichen univariaten Kenngrößen für die *Dauer der Herzinsuffizenz* in beiden *Gruppen* vergleichbar sind.
Es lässt sich festhalten, dass die Randomisierung des Gesamtdatensatzes insgesamt erfolgreich verlaufen ist, wobei insbesondere das *Geschlecht* als fragliche Störgrößenvariable hinsichtlich Wirkungsunterschieden des Medikaments im Blick behalten werden sollte.

Literatur

Assenmacher, W. (2013). *Deskriptive Statistik*. Springer Berlin Heidelberg.

Becker, M. (2022). *3.6 Symmetrie- und Wölbungsmaße*. besucht am 25.10.2022. URL: https://www.lehrstab-statistik.de/online/Deskriptive-WR/Vorlesung/3-6-symmetrie-und-wlbungsmae.html.

Burkschat M., Cramer E. und Kamps U. (2012). *Beschreibende Statistik: Grundlegende Methoden der Datenanalyse*. Springer Berlin Heidelberg.

Fahrmeir L., Heumann C. Künstler R. Pigeot I. Tutz G. (2016). *Statistik - Der Weg zur Datenanalyse*. Springer Berlin Heidelberg.

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

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Tukey J.W., McGill R. und Larsen W.A. (1978). „Variations of Box Plots“. In: *The American Statistician 32* 32(1), S. 12–16.

Van Aken H., Reinhart K. Zimpfer M. Welte T. (2007). *Intensivmedizin*. Thieme.

World Health Organization (2010). *A healthy lifestyle - WHO recommendations*. besucht am 25.10.2022. URL: https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations.

Anhang

Tabelle 1: Deskriptive Kenngrößen für Variable *Geschlecht*

	Gesamt	Randomisiert	Aktiv	placebo
Anzahl männlich	131	107	51	56
Anzahl weiblich	68	57	33	24
Anzahl NA's	0	0	0	0
relativer Anteil männlich	0.66	0.65	0.61	0.70
relativer Anteil weiblich	0.34	0.35	0.39	0.30
relativer Anteil NA's	0	0	0	0
Modalwert	männlich	männlich	männlich	männlich

Tabelle 2: Deskriptive Kenngrößen für Variable *Infarkt*

	Gesamt	Randomisiert	Aktiv	placebo
Anzahl ja	72	62	30	32
Anzahl nein	127	102	54	48
Anzahl NA's	0	0	0	0
relativer Anteil ja	0.36	0.38	0.36	0.40
relativer Anteil nein	0.64	0.62	0.64	0.60
relativer Anteil NA's	0	0	0	0
Modalwert	nein	nein	nein	nein

Tabelle 3: Deskriptive Kenngrößen für Variable *Gruppe*

	Gesamt
Anzahl aktiv	84
Anzahl placebo	80
Anzahl NA's	35
relativer Anteil aktiv	0.42
relativer Anteil placebo	0.40
relativer Anteil NA's	0.18
Modalwert	aktiv

Tabelle 4: univariate Kenngrößen für metrische Variablen aus Gesamtdatensatz

	Größe(cm)	*Gewicht*(kg)	*BMI*($^{kg}/_{m^2}$)	*Alter*(Jahre)	*Dauer*(Monate)
arithm. Mittel	168.86	76.27	26.69	73.00	48.67
Median	169.00	75.00	25.95	72.86	25.23
Minimum	146.00	46.00	17.79	56.58	0.57
Maximum	191.00	132.00	41.20	89.60	315.03
Spannweite	45.00	86.00	23.41	33.02	314.47
1.Quartil	164.00	66.50	23.88	68.21	9.03
3.Quartil	175.00	84.50	28.95	77.46	69.60
IQR	11.00	18.00	5.07	9.26	60.57
Standardabw.	8.56	13.75	4.05	6.20	57.45
MAD	8.90	13.34	3.63	6.92	31.21
Schiefe	-0.24	0.74	0.71	0.05	2.30
Wölbung	3.01	4.28	3.46	2.82	9.61

Tabelle 5: univariate Kenngrößen für metrische Variablen aus randomisierten Datensatz getrennt nach *Gruppe* (a.≙ aktiv, p. ≙ placebo

	Größe (a.)	*Größe* (p.)	*Gewicht* (a.)	*Gewicht* (p.)	*BMI* (a.)	*BMI* (p.)	*Alter* (a.)	*Alter* (p.)	*Dauer* (a.)	*Dauer* (p.)
arithm. Mittel	167.31	170.32	75.40	76.89	26.79	26.49	72.85	73.55	52.29	43.33
Median	168.00	170.00	73.00	75.50	25.79	25.73	72.64	73.31	21.03	26.83
Minimum	146.00	150.00	46.00	52.00	18.36	17.79	56.58	63.67	0.90	0.57
Maximum	191.00	189.00	132.00	121.00	41.20	36.93	89.60	86.71	311.20	152.10
Spannweite	45.00	39.00	86.00	69.00	22.84	19.14	33.02	23.04	310.30	151.53
1.Quartil	161.00	166.00	64.00	68.75	23.67	24.21	67.85	68.22	8.61	10.87
3.Quartil	173.25	175.00	85.00	82.00	29.14	28.67	76.92	77.94	77.22	63.23
IQR	12.25	9.00	21.00	13.25	5.47	4.47	9.07	9.72	68.61	52.37
Standardabw.	9.77	7.75	16.05	11.59	4.37	3.47	6.33	6.04	65.76	39.71
MAD	10.38	7.41	14.83	9.64	3.94	2.75	7.02	7.22	26.32	30.59
Schiefe	-0.14	-0.26	0.82	1.01	0.75	0.67	0.04	0.27	2.09	1.06
Wölbung	2.57	3.11	3.98	5.12	3.44	3.82	2.98	2.29	7.67	3.24