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

2 Problemstellung (noch nicht vollständig)

Im Folgenden werden die größtenteils demografischen Daten einer multinationalen, multizentrischen, doppelblinden, placebo-kontollierten Phase III Studie zur Prüfung der Wirksamkeit eines Medikamentes untersucht.
Dazu ist Datenmatrial in Form eines Datensatzes, 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*, *Zentrum*, *Screeningnummer*, *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 oder ein Placebo erhalten. Die Variablen *Safety-Analysis Population*, *Intention-To-Treat* und *Per-Protocol-Analysis Population* geben dabei weitere klinisch relavante Informationen.
Im Fokus stehen in diesem Projekt jedoch die demografischen Variablen. Dazu gehören das *Geschlecht*, *Größe*, *Gewicht*, *Alter*, *Body-Mass-Index*, *Dauer der bestehenden Herzinsuffizienz* und *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. Dabei bezeichnen „ja“ und „nein“ ob ein Herzinfarkt vorliegt.
Die restlichen demografischen Variablen liegen auf einer Kardinalskala vor. Dabei werden die stetigen Variablen *Größe* in cm und das *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* handelt es sich um das Verhältniss aus Körpergröße und Körpergewicht. Dabei werden Werte des Body-Mass-Index (BMI) 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.

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 Varible mit $x_1, \ldots, x_n$ bezeichnet. Hierbei berzeichnet 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 $\bar{x} := \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$. Neben diesem klassischem Lagemaß werden Quantile verwendet, die in Abhängigkeit des Parameters $p \in (0, 1)$, folgendermaßen definiert sind:

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

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) 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:

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

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

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:

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

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 der beobachteten Werte. Hingegen gibt der Interquartilsabstand $IQA := Q_{0.75} - Q_{0.25}$ einen Bereich an, in dem 50% der Beobachtungen liegen. Auch bei den Streuungsmaßen wird ein gegen Ausreißer robustes Maß eingestzt, welches auf den robsuten Eigenschaften des Medians beruht. Diese Maß wird als Mittlere absolute Abweichung vom Median (MAD) bezeichnet. Bei der Definition wird auf die in R implememtierte Version mit dem Vorfaktor 1.4826 zurückgegriffen:

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

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 $g_1 := \frac{m_3}{s^3}$ 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 $g_2 := \frac{m_4}{s^4}$ unter Verwendung des vierten Momentes definiert. Verglichen wird die Wölbung mit 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

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. 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.

Die empirische Verteilungsfunktion stellt die relativen kumulierten Hufigkeiten dar. Dabei bezeichnet $v(x)$ die absolute Häufigkeit der Werte x_i für die gilt, dass $x_i \leq x$ ist. Betrachtet man nun die georneten Beobachtungen $x_{(i)}$, lässt sich die empirische Verteilungsfunktion folgendermaßen definieren:

$$F(x) = \sum_{i:\ x_{(i)} \leq x} \frac{v(x)}{n}$$

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. Somit wird das Merkmal auf der x-Achse und die relative Häufigkeit $h_{i,j}$ auf der y-Achse abgetragen.

4 Statistische Auswertung

Hier muss ich mich in 4.1 noch für Grafiken oder Tabellen oder nur Grafiken, Formatierung natürlich auch noch nicht da

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 Verteilung der Geschlechter zu erkennen, da etwa 66% der gescreenten Patienten Männer und nur 34% Frauen sind. Somit ist der Modalwert in diesem Fall „männlich“.

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

Tabelle 1: Absolute und relative Häufigkeiten - Geschlecht (n = 199)

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

Tabelle 2: Absolute und relative Häufigkeiten - Herzinfarkt (n = 199)

Eine ähnliche 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 bis zum Zeitpunkt des Screenings keinen Herzinfarkt gehabt zu haben. Bei etwa 64%, was einer absoluten Häufigkeit von 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 grafsiche 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 Verteilung vor, die Anlass dazu gibt eine Normalverteilung zu vermuten. Die univariaten Kennzahlen der beiden Variablen ist den Tabellen 3 und 4 zu entnehmen. Auffällig ist dabei, dass bei bei den Variablen die klassichen und robusten Methoden kaum voneinander abweichen. Dies gilt sowohl für die Lage als auch die Streuungsmaße.

Univaraite Kennzahlen	
Minimum	146.00
0.25-Quartil	164.00
Arithmetische Mittel	168.86
Median	169.00
0.75-Quartil	175.00
Maximum	191.00
Varianz	73.24
Standardabweichung	8.56
Spannweite	45.00
Interquartilsabstand	11.00
MAD	8.90
Schiefe	-0.24
Kurtosis	3.01

Tabelle 3: Größe

Univaraite Kennzahlen	
Minimum	56.58
0.25-Quartil	68.18
Arithmetische Mittel	73.00
Median	72.86
0.75-Quartil	77.53
Maximum	89.60
Varianz	38.42
Standardabweichung	6.20
Spannweite	33.02
Interquartilsabstand	9.26
MAD	6.92
Schiefe	0.05
Kurtosis	2.82

Tabelle 4: Alter

Bei der Varaible *Größe* lässt sich sowhohl grafisch als auch durch Betrachtung des kleinen negativen Wertes des Momentenkoeffizienten der Schiefe eine sehr leichte Tendenz zur Linksschiefe feststellen. Hingegen liegt das Wölbungsmaß mit 3.01 sehr nah bei der Wolbung einer Normalverteilung. Bei der Varaible *Alter* ist es genau anders herrum. Der Momentenkoeffizient der Schiefe liegt mit 0.05 sehr nah an Null und spricht somit für eine nahezu symmetrische Verteilung. Hingegen ist die Wolbung mit 2.82 leicht flacher als bei der Normalverteilung.

Bei Betrachtung der Variablen *Gewicht* und *Body - Mass - Index* ist jeweils eine größere Abweichung von einer symmetrischen Verteilung als bei den vorherigen betrachteten Variablen zu erkennen. Maße zur Lage, Streuung, Schiefe und Wölbung sind Tabelle 5 und 6 zu entnehmen. Auch hier gibt es noch keinen nennenswerten Unterschiede der klassichen und robusten Lage-und Streuungsmaße, wobei die Maße im Vergleich zu *Alter* und *Gewicht* schon etwas mehr voneinander abweichen. Der Grund dafür könnten die in den Boxplots beider Variablen erkennbaren Ausreißer im oberen Bereich sein. Zudem deutet sowohl die grafische Analyse, als auch die Betrachtung des Momentenkoeffizient der Schiefe, der bei beiden Variablen etwa 0.7 beträgt, auf eine leichte Rechtsschiefe hin. In Bezug auf die Wölbung sind beide Verteilungen spitzer als die Normalverteilung. Dabei ist die Variable *Gewicht* mit einer Wölbung von 4.28 noch etwas spitzer als die Varaible *Body-Mass-Index* mit einer Wölbung von 3.46.

	x
Minimum	46.00
0.25-Quartil	66.00
Arithmetische Mittel	76.27
Median	75.00
0.75-Quartil	85.00
Maximum	132.00
Varianz	189.16
Standardabweichung	13.75
Spannweite	86.00
Interquartilsabstand	18.00
MAD	13.34
Schiefe	0.74
Kurtosis	4.28

Tabelle 5: Gewicht

	x
Minimum	17.79
0.25-Quartil	23.88
Arithmetische Mittel	26.69
Median	25.95
0.75-Quartil	29.05
Maximum	41.20
Varianz	16.44
Standardabweichung	4.05
Spannweite	23.41
Interquartilsabstand	5.07
MAD	3.63
Schiefe	0.71
Kurtosis	3.46

Tabelle 6: Body-Mass-Index

	x
Minimum	0.57
0.25-Quartil	9.00
Arithmetische Mittel	48.67
Median	25.23
0.75-Quartil	69.87
Maximum	315.03
Varianz	3300.63
Standardabweichung	57.45
Spannweite	314.47
Interquartilsabstand	60.57
MAD	31.21
Schiefe	2.30
Kurtosis	9.61

Tabelle 7: Dauer Herzinsuffizienz

Bei der Variable *Dauer der Herzinsuffizienz* ist erstmals eine deutlich Abweichung zwischen den robusten und klassischen Lage- und Streuungsmaßen erkennbar. Das arithmetische Mittel ist mit 48.67 knapp doppelt so groß wie der Median, der 25.57 beträgt. Ein ähnliches Bild ist bei der Standardabweichung erkennbar, die mit 57.45 ebenfalls deutlich größer als der MAD mit 31.21 ist. Dies könnte an der stark ausgeprägten Reschtsschiefe liegen, welche sowohl grafisch als auch am recht großen positiven Wert von 2.30 des Momentenkoeffizienten der Schiefe erkennbar ist. Auch die im Boxplot erkennbaren Ausreißer können ein Grund für die Diskrepanz der robusten und klassischen Methoden sein. Auch die Wölbung hat bei dieser Verteilung einen recht hohen Wert von 9.61, welcher für eine deutlich spitzere Verteilung als die Normalverteilung spricht.

4.2 Vergleich der Verteilungen der interessierenden Variablen zwischen den Medikationsgruppen

Zur Beurteilung des Erfolges der Randomisierung erfolgt pro Varaible ein Vergleich der Verteilungen zwischen den Medikationsgruppen. Wie in Abbildung 1 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. Genau anders herrum ist es bei den Frauen, welche in der aktiven MG etwa 10% mehr als in der placebo MG sind. Somit bewegen sich Abweichungen immer im Rahmen von etwa 10%. Bei der Variable *Herzinfarkt* bewegt sich die Abweichungen im Rahmen von etwa 5%. Denn in der placebo MG hatten, mit etwa 40%, mehr Personen einen Herzinfarkt als in der aktiven MG mit etwa 36%.

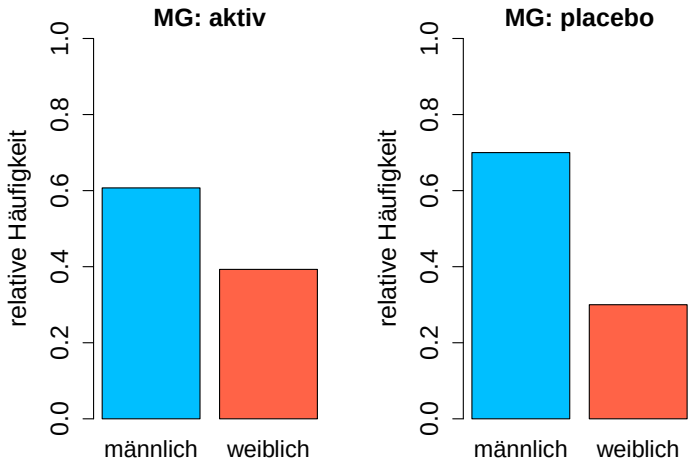

Abbildung 1: Geschlecht - Vergleich in Medikationsgruppen

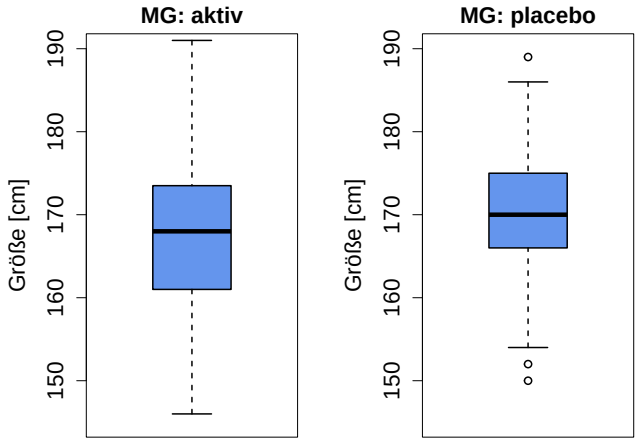

Abbildung 2: Boxplots der Variable Größe getrennt nach MG

Bei der Variable *Größe* ist in Abblidung 2 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. Im Mittel ist die Größe in der aktiven MG bei einem Median von 168 minimal niedriger als ein Median von 170 in der placebo MG.

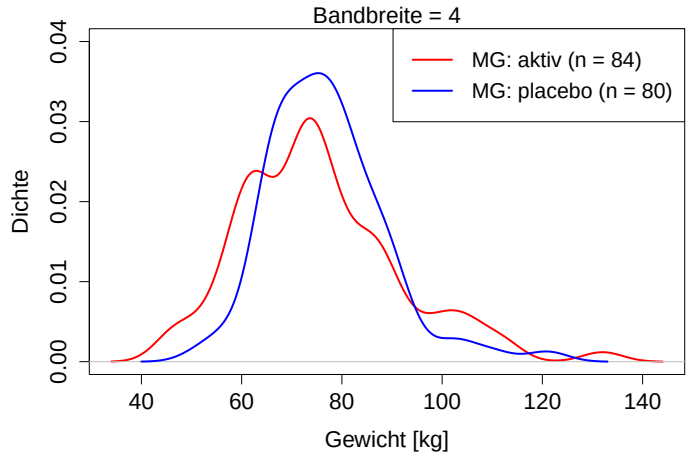

Abbildung 3: Kerndichteschätzer der Variable Gewicht getrennt nach MG

Auch beim der Variable *Gewicht* ist eine etwas größere Streuung in der aktiven MG erkennbar. Der MAD liegt in der aktiven 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 1.01 und in der placebo MG etwa 0.82. 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.

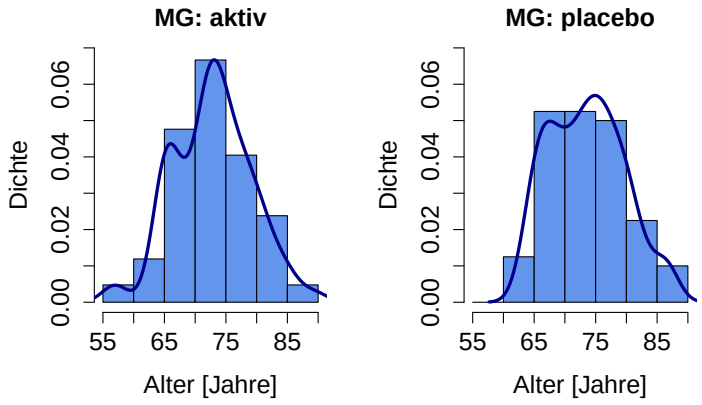

Abbildung 4: Kerndichteschätzer (Bandbreite = 2) und Histogramm der Variable Alter getrennt nach MG

In Abbildung 4 ist zu erkennen, dass bei der Variable *Alter* die Symmetrie Eigenschaften

aus der Grundgesamtheit aller gescreenten Personen nahezu beibehalten wird. Lediglich in der placebo MG ist eine leichte Tendenz zur Rechtsschiefe erkennbar (vgl. $g_1 = 0.27$). Auch im Mittel weicht das Gewicht in der aktiv MG nicht stark von dem Gewicht in der placebo MG ab. Das arithmetische Mittel beträgt 72.85 und 73.55. Ein Unterschied der Verteilungen ist jedoch in der Kurtosis zu erkennen. Die Verteilung in der placebo MG ist flacher ($g_2 = 2.2856380$) als in der aktiv MG ($g_2 = 2.98401344$). Obgleich die Streuung mit robusten oder klassischen Maßen gemessen ist, variiert sie nicht nenenswert.

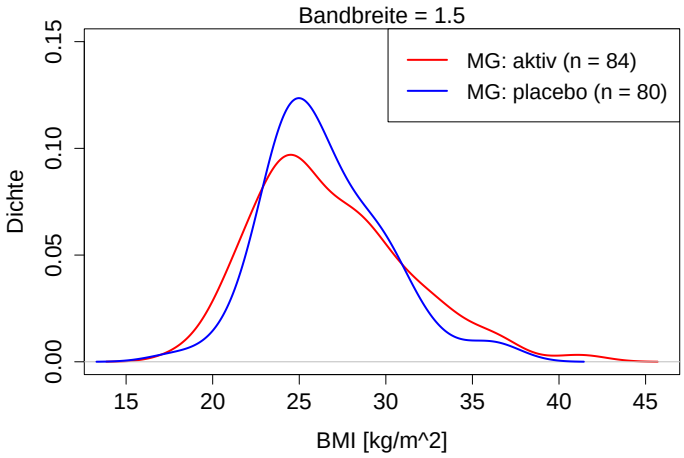

Abbildung 5: Kerndichteschätzer der Variable BMI getrennt nach MG

Bei der Variable *Body-Mass-Index* ist in Abbildung 5 zu erkennen, dass sich die Verteilung in den beiden MG nur geringfügig unterscheiden. In der aktiven MG (MAD = 3.94, $s = 4.37$) ist die Streuung minimal größer als in der placebo MG (MAD = 2.75, s = 3.47), 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 26.79287 (aktiv) und 26.49465 (placebo) nahezu gleich.

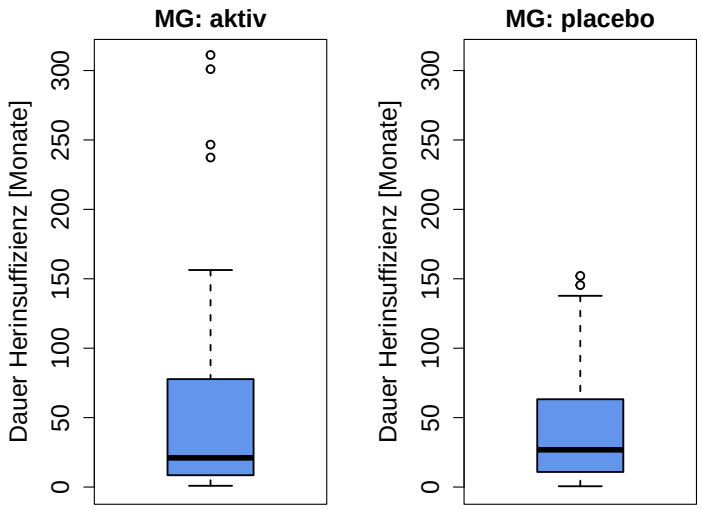

Abbildung 6: Boxplots der Variable Dauer der Herzinsuffizienz getrennt nach MG

Analog zur Verteilung der Variable *Dauer der Herzinsuffizienz* in der Gesamtheit aller gescreenten Patienten, kann man in Abbildung 6 auch in den beiden Medikationsgruppen eine starke 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 ($g_2 = 3.24$) ist. Zudem ist auffällig, dass in der aktiven MG deutlich größere Ausreißer auftreten als in der placebo MG.
In Bezug auf die Streuung liefern klassische und Robuste Methoden wiedersprüchliche Ergebnisse. So beträgt der MAD in der aktiven MG etwa 26.32 und ist somit kleiner als in der placebo MG mt einem MAD von etwa 30.59. Die Standardabweichung ist jedoch in der aktiven MG mit etwa 65.76 größer gegnüber etwa 39.71 in der placebo MG. 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, trotz der robusten Eigenschaften des Medians, etwas voneinander ab (aktiv: $Q_{0.5} = 21.03$, placebo: $Q_{0.5} = 26.83$).

5 Notizen

- Größe der Teildatensätze **Vergleich der nominalen Varaiblen**

Geschlecht: - kleiner Unterschied - Etwas mehr Männer/weniger Frauen in der placebo gruppe

Alter - sehr geringer Unterschied - etwas mehr mit Herzinfarkt in placebo Gruppe

stetige Variablen

Größe - BOXPLOT und Verteilungsfunktion - aktiv streut mehr als placebo (1) - placebo linksschiefer als aktiv - aktiv im Mittel etwas kleiner

Gewicht - KERNDICHTESCHÄTZER - aktiv streut mehr als placebo (2) - placebo rechtsschiefer und spitzer - aktiv im Mittel (ein ganz wenig) kleiner

Alter - HISTOGRAMM - aktiv streut kaum mehr als placebo - placebo flacher als aktiv - aktiv symmetrisch, placebo rechtsschief - aktiv im Mittel etwas kleiner

BMI - aktiv streut minimal mehr als placebo (4) - aktiv rechtsschiefer als placebo - placebo spitzer als aktiv - im Mittel aktiv und placebo nahezu gleich

Herzinsuffizienz - KERNDICHTESCHÄTZER - aktiv streut mehr (mehr Ausreißer) als placebo (sd und mad unterschied!) (teilweise (5)) - aktiv rechtsschiefer als placebo - aktiv spitzer als placebo - unterschied median, a.m.

R Core Team 2021 Fahrmeir et al. 2011

6 Zusammenfassung

7 Literaturverzeichnis

Literatur

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

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/`.

8 Anhang