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

Gruppenmitglieder:
Caroline Baer
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, Hugo et al.).
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 bzw. die Patientin 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 in diesem Bericht soll zeigen, ob die Randomisierung der Studienprobanden insofern erfolgreich war, dass sich die Verteilung der interessierenden demographischen Variablen zwischen den Medikationsgruppen nicht stark unterscheidet.
Nachdem zunächst die zugrunde liegenden Daten und das genaue Ziel des Berichts beschrieben werden, werden sowohl graphische wie auch numerische statistische Methoden vorgestellt, die für die Analyse erforderlich sind. Im Anschluss erfolgt in der statistischen Auswertung der Vergleich der Verteilungen der interessierenden Variablen, wobei besonders auffällige Erkenntnisse hervorgehoben betrachtet werden. Dies ermöglicht schließlich die Zusammenfassung 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* wurden eine Stichprobe aus 200 Patienten und Patientinnen 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 primäre Teilerhebung einer Beobachtungsstudie.
Die Variablen Land (landnr), Zentrum (zentrum), Screeningnummer (screennr), Patientennummer (patnr), Medikationsgruppe (gruppe), im Folgenden als Gruppe bezeichnet, Safety-Analysis Population (saf), Intention-To-Treat Population (itt) und Per-Protocol-Analysis Population (ppa) sollen lediglich erwähnt sein, werden im Folgenden aber nicht weiter betrachtet.
Für dieses Projekt sind sieben Variablen von Interesse: Die nominalen Variablen Geschlecht (sex) mit den Ausprägungen 1 für männlich und 2 für weiblich und erlittener Herzinfarkt (myo_infarct), im Folgenden als Infarkt bezeichnet, mit den Ausprägungen 1 für ja und 2 für nein wurden diskret erhoben. Die metrischen Variablen Größe (groesse) in cm und Gewicht (gewicht) in kg wurden diskret erhoben. Die metrischen Variablen Alter (alter) in Jahren, Body-Mass-Index (bmi) in kg/m^2, im Folgenden als BMI bezeichnet und Dauer der bestehenden Herzinsuffizienz (dauer_insuff) in Monaten, im Folgenden als Dauer bezeichnet, wurden stetig erhoben. Das Alter und die Dauer ergeben sich als Differenz aus dem Datum des Screenings und dem Datum der Geburt bzw. der Erstdiagnose. Der BMI wird mit der allgemeinen Definition Körpergewicht in Kilogramm geteilt durch die quadrierte Körptergröße in Metern berechnet (WHO) und gibt den Ernährungsstatus einer Person an. BMI Werte zwischen 18.5 und 24.9 zeigen ein Normalgewicht an, kleinere Werte Untergewicht, höhere Werte Übergewicht.
Um die Qualität des Datensatzes einschätzen zu können, werden die Daten auf fehlende Werte in R (R Core Team, ???) untersucht. Die mit 86 insgesamt relative hohe Zahl fehlender Werte im gesamten Datensatz (43%) lässt sich damit begründen, dass für 36 Patienten im Screeningverfahren die interessierenden Variablen Geschlecht, Größe, Gewicht, Alter, BMI, Dauer und Infarkt erhoben wurden, diese Patienten im Anschluss aber nicht randomisiert in die Gruppen 1 für Placebo und 2 für aktives Medikament eingeteilt werden, sodass bei ihnen in zwei der insgesamt 15 Variablen fehlenden Werte generiert werden.
Anders sieht es bei Beobachtung 179 aus, bei der in neun von 15 Variablen fehlende Werte generiert werden. Diese Beobachtung wird als Messfehler betrachtet und aus dem

Datensatz entfernt. Da der Fragestellung einer erfolgreichen Randomisierung mit einem Vergleich der in die beiden Gruppen eingeteilten Patienten nachgegangen wird, werden nur die 164 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. Eine Entfernung dieser Daten führt zu einer Stichprobe von N = 159 Patienten, was eine hinreichend große Stichprobengröße für den Vergleich der interessierenden Variablen darstellt.

2.2 Ziele des Projekts

In diesem Bericht soll deskriptiv untersucht werden, ob die Randomisierung der Studienpatienten in die beiden Gruppen erfolgreich verlaufen ist. Dieser Fragestellung wird nachgegangen indem die Verteilungen der interessierenden Variablen zwischen den Gruppen verglichen werden. 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

Betrachtet man die geordneten Werte $x_{1j}, ..., x_{nj}$ metrisch skalierter Variablen $x_1, ..., x_n$ mit Stichprobenumfang N, dann ist das **p - Quantil** $\tilde{x}_p$ für $p \in (0, 1)$ gegeben durch

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

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. Diese kennt 9 verschiedene Definitionen von Quantilen. Die Definition für type=2 entspricht der aus (1) (Quelle: R oder Ligges?) Das 0.5-Quantil $\tilde{x}_{0.5}$ entspricht dem Wert in der Mitte aller geordneten Beobachtungswerte und wird **Median** genannt (Burkschat).
Der **Quartilsabstand** 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.
Ein Lagemaß, das in seiner Ausprägung empfindlich auf auf **Ausreißer**, das heißt auf (für die Verteilung) ungewöhnlich große oder kleine Werte, und auf Schiefe der Verteilung

reagiert ist das **arithmetischen Mittel** gegeben durch

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

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

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

berechnen (Becker). Eine Verteilung heißt rechtsschief, wenn $skewness > 0$, linksschief wenn $skewness < 0$ und symmetrisch falls $skewness = 0$.
s steht in Formel 3 für die Standardabweichung. Die Standardabweichung ergibt sich als Wurzel der **Varianz**, die gegeben ist durch (Ligges)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4}$$

Die Varianz ist die durchschnittliche quadratische Abweichung der Beobachtungen von $\bar{x}$ (Assemacher). Wegen des Quadrierens besitzt sie jedoch eine andere Dimension des betrachteten Merkmals. 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 (Assemacher).
Eine weitere Verteilungscharakterisierung ist die **Wölbung** der Verteilung. Die Wölbung gibt an, wie gleichmäßig die Beobachtungen um den Median streuen. Der empirische Wölbungkoeffizient gegeben durch

$$kurtosis = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \tag{5}$$

(Becker) ist $= 3$ falls die Beobachtungen im Sinne der **Normalverteilung** (????) um den Median gelegen sind, $\in [1, 3)$ falls die Beobachtungen eher gleichmäßig (flach) zwischen dem kleinsten Wert, dem **Minimum**, und dem größten Wert, dem **Maximum**, der Stichprobe verteilt sind und > 3 falls die Beobachtungen dicht (spitz) um den Median verteilt sind.
Schiefe und Wölbe von Verteilungen lassen sich in R mit dem Paket „moment“ nach den in Formel 4 und 5 gegebenen Definitionen berechnen.
Die **Spannweite** ist die Differenz aus Maximum und Minimum.

3.2 deskriptive Graphiken

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 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. In der R Funktion boxplot() wird mit der von Tukey et. al. vorgestellten Definition von Quantilen gearbeitet. (Wie viel weiter erläutern?)

Ein **Säulendiagramm** (english Barplot) ist eine einfache grafische Methode, um die Häufigkeiten der 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).

Metrische Daten mit sehr vielen Beobachtungswerten lassen sich übersichtlich darstellen, indem sie (unter Inkaufnahme eines gewissen Informationsverlusts) in Klassen zusammengefasst werden. Für die Klassifikation der Merkmalausprägungen mindestens ordinaler skalierter Daten in Intervalle, wobei das erste und das letzte Intervall keine offene Klasse sein dürfen, kann ein **Histogramm** als grafisches Hilfsmittel hinzugezogen werden. Die Klassenintervalle für m=1,...,M Klassen werden wie folgt definiert:

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

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. Auf der x-Achse eines Histogramms befinden sich die Klassengrenzen $\nu_0, ..., \nu_M$ der Intervalle, die y-Achse zeigt die absolute Häufigkeit der Beobachtungen in der jeweiligen Klasse an. Zwischen den Klassengrenzen, also in jedem Intervall K_j, wird ein Kasten gezeichnet. Die Breite des Kastens entspricht der

Länge des Intervalls, also der Klassenbreite b_j, die Höhe h_j wird als Quotient $\frac{f(K_j)}{b_j}$ der relativen Klassenhäufigkeit und der Klassenbreite berechnet (Burkschat). Daraus ergibt sich, dass die Fläche des Kastens die relative Häufigkeit der Klasse ist (Ligges).
Die **empirische Verteilungsfunktion** gibt für mindestens ordinale Daten die Folge der Summenhäufigkeiten S_m mit $M = 1, ..., m$ Merkmalsklassen an. Die Summenhäufigkeiten ist definiert als

$$S_k = \begin{cases} 0 & \forall x < \nu_1 \\ \sum_{j=1}^{m} h_j & x \in \text{[Ende von Klasse m, Ende von Klasse m+1]} \\ 1 & \forall x \geq \nu_M \end{cases} \tag{6}$$

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

4 Statistische Auswertung

Die Fragestellung, inwieweit die Randomisierung der in die Studie aufgenommenen Patienten erfolgreich verlaufen ist, lässt sich mit einem Vergleich zwischen den Gruppen aktiv und placebo in den interessierenden Variablen beantworten.
In Tabelle 1 ist zu sehen, dass die Aufteilung der Probanden in die Gruppen sich nur um fünf Probanden (drei % von N) unterscheidet und auch die Geschlechterverteilung in den Gruppen vergleichbar ist. Dies sind wichtige Grundvoraussetzungen, um die anderen Variablen sinnvoll miteinander vergleichen zu können.
Die Verteilungen der laut den Risikofaktoren (NVL) für eine chronische Herzinsuffizienz bedeutenden Variablen Infarkt, Gewicht, Alter, BMI und Dauer werden im Folgenden dargestellt. In Abbildung **??** ist zu sehen, dass die Verteilung der Variable Infarkt zwischen den Gruppen nahezu ausgeglichen ist.

Tabelle 1: Vierfeldertafel für Geschlecht aus randomisierten Datensatz

	aktiv	placebo
männlich	51	53
weiblich	31	24
Summe	82	77

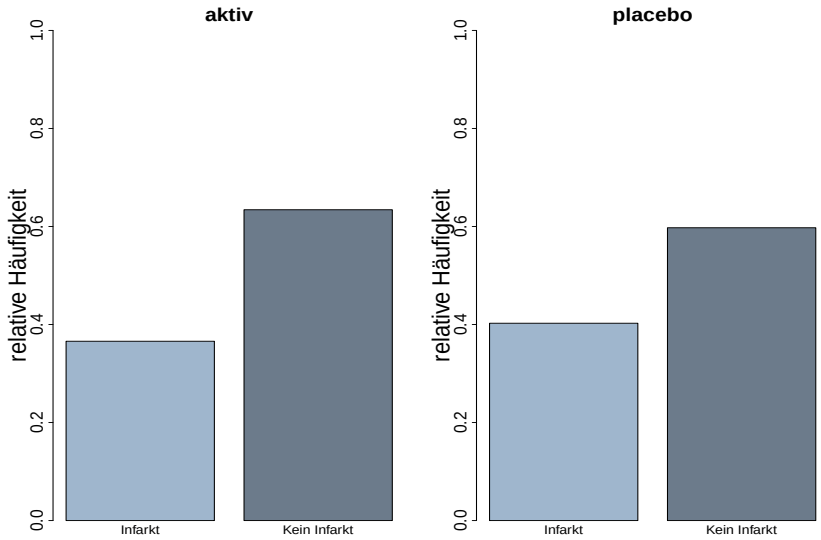

Abbildung 1: Säulendiagramm für Infarkt, aufgeteilt in Gruppen

Tabelle **??** zeigt univariate Kenngrößen für die metrischen Variablen. Die Größe scheint in beiden Gruppen gleichmäßig ausgeglichen zu sein. Die Verteilungscharakterisierungen Schiefe und Wölbung zeigen, dass die Verteilungen beider Gruppen der einer Normalverteilung nahe kommen. Dies wird durch Abbildung **??** unterstützt, in der die Häufigkeitsausprägungen der Variable Größe in einem Histogramm dargestellt wird und über die Dichtefunktion einer an die Variable angepasste Normalverteilungsfunktion gelegt wird. Beim Gewicht sieht es anders aus: Die Wölbungskoeffizienten, die in der aktiven Gruppe bei 4.04 und bei der placebo Gruppe bei 6.04 liegen lassen auf eine spitze Verteilung schließen, das heißt die Körpergewichte der Probanden liegen dicht 0um den Median 73 bzw. 75 Kilogramm. Das dies bei beiden Gruppen der Fall ist und nur in bedingtem Maß für die Auswertung der Daten von Bedeutung ist, zeigt ein Blick auf Abbildung **??**. Die Boxplots zeigen eine ähnliche Verteilung der BMIs in beiden Gruppen.

Tabelle 2: univariate Kenngrößen für metrische interessierende Variablen aus randomisierten Datensatz (a. $\triangleq$ aktiv, p. $\triangleq$ placebo)

	Größe (a.)	Größe (p.)	Gewicht (a.)	Gewicht (p.)	Alter (a.)	Alter (p.)	BMI (a.)	BMI (p.)	Dauer (a.)	Dauer (p.)
1.Quartil	161.00	166.00	64.00	68.00	68.23	68.23	23.88	24.17	8.61	10.87
3.Quartil	173.75	175.00	85.00	81.00	76.69	77.93	29.32	28.37	77.22	63.23
Median	168.00	170.00	73.00	75.00	72.64	73.60	25.91	25.51	21.03	26.83
IQR	12.75	9.00	21.00	13.00	8.46	9.70	5.44	4.21	68.61	52.37
emp. Schiefekoef.	-0.19	-0.22	0.85	1.15	0.04	0.25	0.80	0.75	2.09	1.06
Standardabw.	9.82	7.72	15.83	11.16	6.33	6.04	4.31	3.43	65.76	39.71
Wölbung	2.59	3.16	4.04	6.01	3.04	2.32	3.44	4.12	7.67	3.24
Minimum	146.00	150.00	46.00	52.00	56.58	63.67	19.65	17.79	0.90	0.57
Maximum	191.00	189.00	132.00	121.00	89.60	86.71	41.20	36.93	311.20	152.10
Spannweite	45.00	39.00	86.00	69.00	33.02	23.04	21.55	19.14	310.30	151.53

Insbesondere, dass die sich Mediane, wie in Tabelle **??** ersichtlich, in den Gruppen nur marginal unterscheiden (25.91 $\frac{kg}{m^2}$ vs. 25.51 $\frac{kg}{m^2}$) ist ein bedeutender Faktor, der für eine erfolgreiche Randomisierung der Probanden spricht.

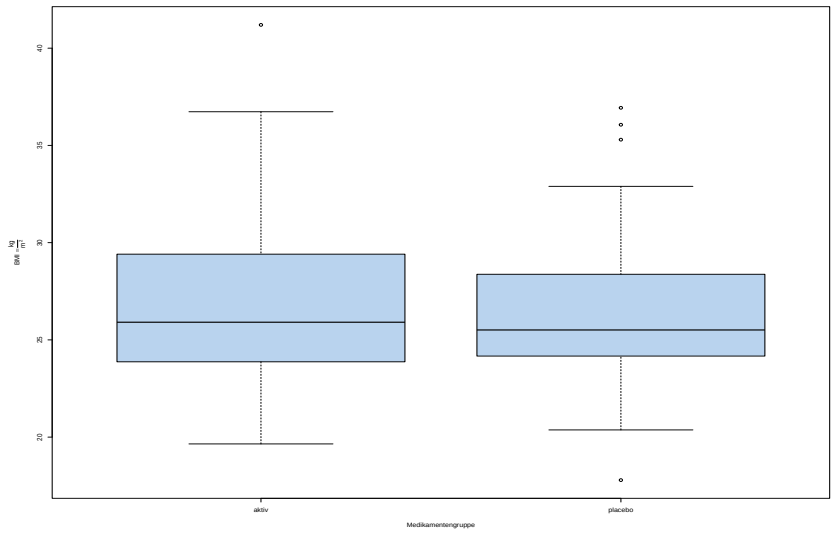

Abbildung 2: Boxplots für BMI aufgeteilt in Gruppen

5 Anhang

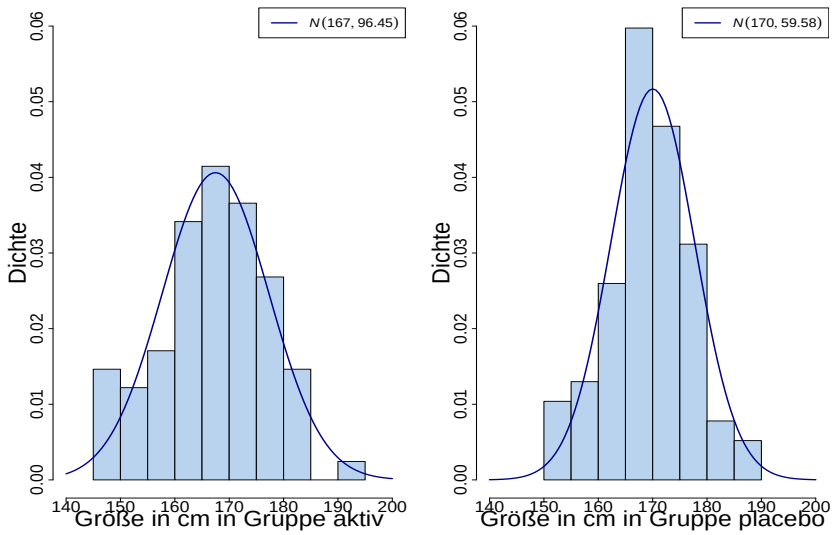

Abbildung 3: Histogramme für Größe aufgeteilt in Gruppen