KAPITEL 10

Visualisierung von Proportionen

Wir möchten oft zeigen, wie eine Gruppe, eine Menge, ein Wert oder ein Betrag in einzelne Teile zerfällt, die jeweils einen Teil des Ganzen darstellen. Häufige Beispiele sind das Verhältnis von Männern und Frauen in einer Personengruppe, der Prozentsatz der Personen, die bei einer Wahl für verschiedene politische Parteien stimmten, oder die Marktanteile von Unternehmen. Die prototypische Visualisierung hierfür ist das Kreisdiagramm (engl. *pie chart*), das in jeder geschäftlichen Präsentation allgegenwärtig ist und unter Data Scientists häufig schlecht beleumdet ist. Wie wir sehen werden, kann es eine Herausforderung sein, Proportionen zu visualisieren, insbesondere wenn das Ganze in viele unterschiedliche Anteile aufgespalten ist oder wenn wir Änderungen der jeweiligen Anteile im zeitlichen Verlauf oder über bestimmte Kategorien hinweg sehen wollen. Es gibt nicht die einzige ideale Darstellungsform, die immer funktioniert. Um dieses Problem zu veranschaulichen, werden im Folgenden einige unterschiedliche Szenarien besprochen, für die jeweils ein anderer Typ von Visualisierung erforderlich ist.

Denken Sie daran, dass Sie immer die Visualisierung auswählen müssen, die am besten zu Ihrem spezifischen Datensatz passt und die Hauptmerkmale hervorhebt, die Sie zeigen möchten.

Ein Fall für Kreisdiagramme

Von 1961 bis 1983 setzte sich der Deutsche Bundestag aus Mitgliedern der drei Parteien CDU/CSU, SPD und FDP zusammen. Während des größten Teils dieser Zeit hatten CDU/CSU und SPD ungefähr eine vergleichbare Anzahl von Sitzen, während die FDP typischerweise nur einen kleinen Bruchteil der Sitze innehatte. Zum Beispiel hatte die CDU/CSU im achten Bundestag von 1976 bis 1980 243 Sitze, die SPD 214 und die FDP 39 von insgesamt 496 Sitzen. Solche parlamentarischen Daten werden am häufigsten als Kreisdiagramm dargestellt (Abbildung 10-1).

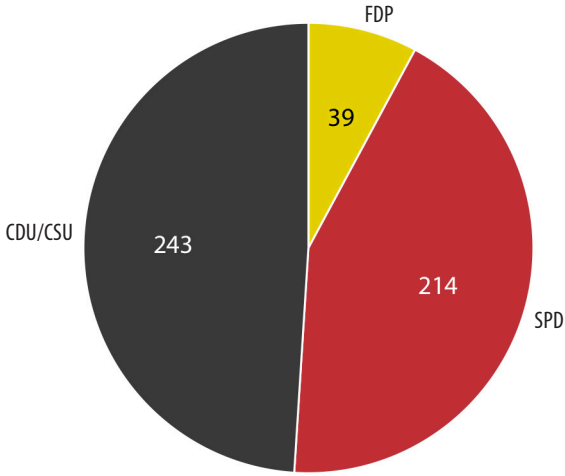

Abbildung 10-1: Parteizusammensetzung des 8. Deutschen Bundestages 1976–1980, dargestellt als Kreisdiagramm. Diese Visualisierung verdeutlicht, dass die Regierungskoalition von SPD und FDP eine geringe Mehrheit gegenüber der Opposition von CDU/CSU hatte. (Datenquelle: Wikipedia)

Ein Kreisdiagramm zerlegt einen Kreis in Segmente, sodass die Fläche jedes Segments proportional zum Bruchteil der Summe ist, die es darstellt. Das gleiche Verfahren kann für ein Rechteck durchgeführt werden. Das Ergebnis ist ein gestapeltes Balkendiagramm (Abbildung 10-2). Je nachdem, ob wir den Balken vertikal oder horizontal differenzieren, erhalten Sie vertikal (Abbildung 10-2a) oder horizontal gestapelte Balken (Abbildung 10-2b).

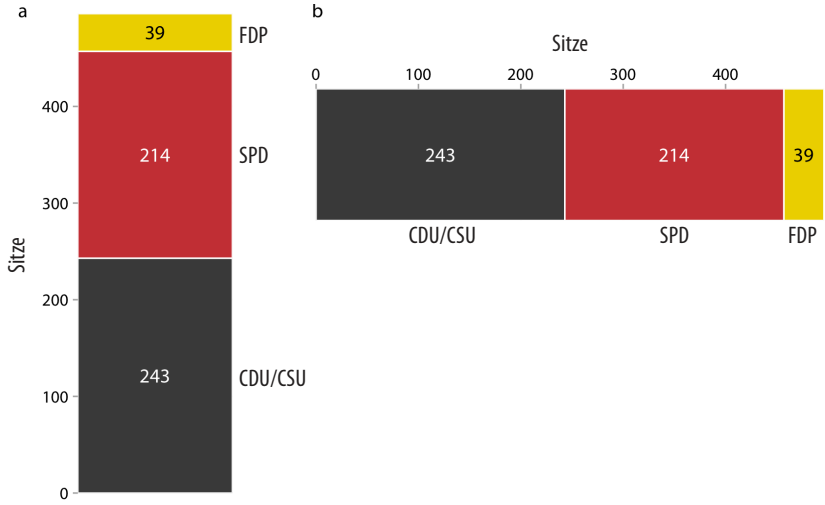

Abbildung 10-2: Parteienzusammensetzung des achten Deutschen Bundestages 1976–1980, dargestellt als gestapelte Balken: (a) vertikale Ausrichtung (b); horizontale Ausrichtung. In dieser Darstellung ist nicht sofort ersichtlich, dass SPD und FDP gemeinsam mehr Sitze hatten als CDU/CSU. (Datenquelle: Wikipedia)

Wir können die Balken auch aus Abbildung 10-2a entnehmen und nebeneinander platzieren, anstatt sie übereinander zu stapeln. Diese Visualisierung erleichtert den direkten Vergleich der drei Gruppen, verdeckt jedoch andere Aspekte der Daten (Abbildung 10-3). Insbesondere ist in einem solchen Balkendiagramm (mit nebeneinander gezeichneten Balken) die Beziehung der einzelnen Balken zum Ganzen optisch nicht ersichtlich.

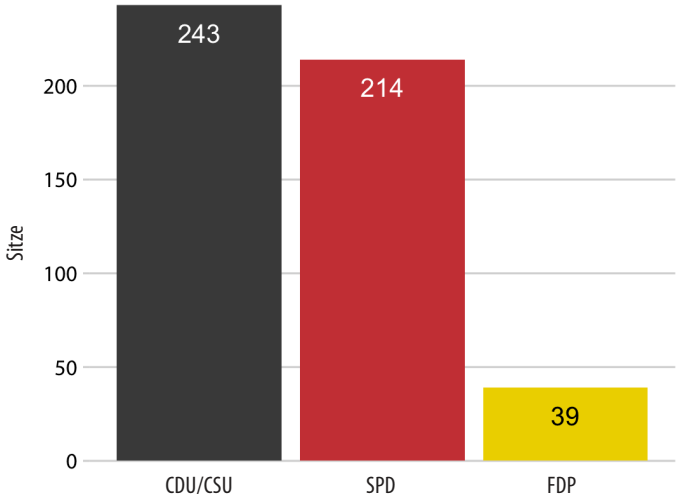

Abbildung 10-3: Parteienzusammensetzung des 8. Deutschen Bundestages 1976–1980, dargestellt als seitlich nebeneinander angeordnete Balken. Wie in Abbildung 10.2 ist nicht sofort ersichtlich, dass SPD und FDP gemeinsam mehr Sitze hatten als CDU/CSU. (Datenquelle: Wikipedia)

Viele Autoren lehnen Kreisdiagramme kategorisch ab und plädieren für nebeneinander angeordnete oder gestapelte Balken. Andere verteidigen die Verwendung von Kreisdiagrammen in einigen Anwendungsfällen. Meine Meinung ist, dass keine dieser Darstellungen anderen durchgehend überlegen ist. Abhängig von den Merkmalen des Datensatzes und der spezifischen Botschaft, die Sie übermitteln möchten, können Sie möglicherweise den einen oder anderen Ansatz bevorzugen. Im Falle des achten Deutschen Bundestages denke ich, dass ein Kreisdiagramm die beste Option ist: Es wird deutlich, dass die Regierungskoalition aus SPD und FDP gemeinsam eine geringe Mehrheit gegenüber der CDU/CSU hatte (Abbildung 10-1). Diese Tatsache ist in keinem der anderen Diagramme sichtbar (Abbildungen 10-2 und 10-3).

Im Allgemeinen funktionieren Kreisdiagramme gut, wenn einfache Anteile wie die Hälfte, ein Drittel oder ein Viertel hervorgehoben werden sollen. Sie funktionieren auch gut, wenn wir sehr kleine Datensätze haben. Ein einzelnes Kreisdiagramm, wie in Abbildung 10.1, sieht ansprechend aus, während eine einzelne Spalte mit gestapelten Balken, wie in Abbildung 10.2a, eher unbeholfen aussieht. Nebeneinander gestapelte Balken hingegen können mehrere Kategorien oder eine Zeitreihe miteinander vergleichen; sie werden bevorzugt, wenn die einzelnen Anteile direkt

miteinander verglichen werden sollen. Eine Zusammenfassung der verschiedenen Vor- und Nachteile von Kreisdiagrammen, gestapelten Balken und nebeneinander angeordneten Balken finden Sie in Tabelle 10-1.

Tabelle 10-1: Vor- und Nachteile der gängigen Ansätze zur Darstellung von Proportionen: Kreisdiagramme, gestapelte Balken und nebeneinander angeordnete Balken.

	Kreisdiagramme	Gestapelte Balken	Nebeneinander angeordnete Balken
Visualisiert die Daten klar als Anteile eines Ganzen.	✓	✓	×
Ermöglicht einen einfachen optischen Vergleich der relativen Anteile.	×	×	✓
Hebt einfache Anteile wie 1/2, 1/3, 1/4 optisch hervor.	✓	×	×
Sieht auch für sehr kleine Datensätze optisch ansprechend aus.	✓	×	✓
Gut geeignet, wenn ein Ganzes in viele Anteile gegliedert ist.	×	×	✓
Gut geeignet für die Darstellung mehrerer Proportionen oder für Zeitreihen von Proportionen.	×	✓	×

Ein Fall für nebeneinander angeordnete Balken (engl. Side-by-side bars)

Ich werde nun einen Fall demonstrieren, für den Kreisdiagramme nicht geeignet sind. Dieses Beispiel ist einer Kritik an Kreisdiagrammen nachempfunden, die ursprünglich auf Wikipedia [Wikipedia 2007] veröffentlicht wurden. Betrachten Sie das hypothetische Szenario von fünf Unternehmen, A, B, C, D und E, die alle einen vergleichbaren Marktanteil von ca. 20% haben. Unser hypothetischer Datensatz listet den Marktanteil jedes Unternehmens für drei aufeinanderfolgende Jahre auf. Wenn wir diesen Datensatz mit Kreisdiagrammen visualisieren, ist es schwierig, bestimmte Trends zu erkennen (Abbildung 10.4). Es scheint, dass der Marktanteil von Unternehmen A wächst und der von Unternehmen E schrumpft, aber über diese Beobachtung hinaus können wir keine weiteren Aussagen treffen. Insbesondere ist unklar, wie sich die Marktanteile der verschiedenen Unternehmen innerhalb eines Jahres genau unterscheiden.

Das Bild wird etwas klarer, wenn wir zu gestapelten Balken wechseln (Abbildung 10-5). Nun sind die Trends eines wachsenden Marktanteils für Unternehmen A und eines schrumpfenden Marktanteils für Unternehmen E deutlich zu erkennen. Die relativen Marktanteile der fünf Unternehmen innerhalb eines Jahres sind jedoch noch immer schwer zu vergleichen. Und es ist schwierig, die Marktanteile der Unternehmen B, C und D über die Jahre hinweg zu vergleichen, da die Balkensegmente

dieser Unternehmen in den Jahren gegeneinander verschoben sind. Dies ist ein allgemeines Problem von Abbildungen mit gestapelten Balken und der Hauptgrund, warum ich diese Art der Visualisierung normalerweise nicht empfehle.

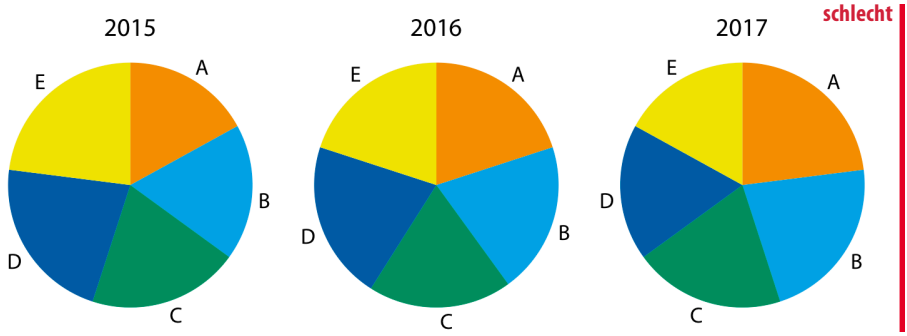

Abbildung 10-4: Marktanteil von fünf hypothetischen Unternehmen, A bis E, für die Jahre 2015–2017, dargestellt als Kreisdiagramme. Diese Darstellung weist zwei Hauptprobleme auf: (i) Ein Vergleich des relativen Marktanteils in den verschiedenen Jahren ist nahezu unmöglich und (ii) die Veränderungen des Marktanteils im zeitlichen Verlauf über mehrere Jahre sind schwer erkennbar.

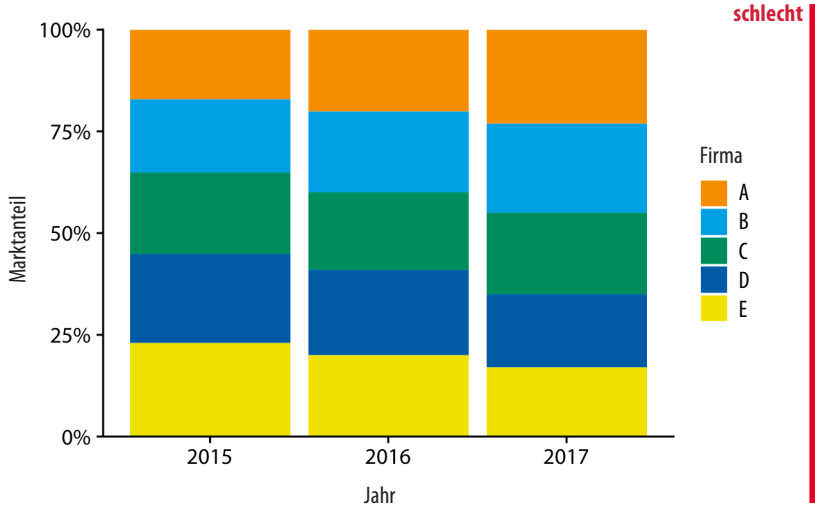

Abbildung 10-5: Marktanteil von fünf hypothetischen Unternehmen für die Jahre 2015–2017, dargestellt als gestapelte Balken. Diese Darstellung weist zwei Hauptprobleme auf: (i) Ein Vergleich der relativen Marktanteile in den verschiedenen Jahren ist schwierig, und (ii) Veränderungen der Marktanteile im zeitlichen Verlauf über mehrere Jahre für die mittleren Unternehmen (B, C und D) sind schwierig zu erkennen, da die Balkensegmente in den Jahren gegeneinander verschoben sind.

Für diesen hypothetischen Datensatz sind nebeneinander angeordnete Balken die beste Wahl (Abbildung 10.6). Diese Visualisierung zeigt, dass die Unternehmen A

und B ihren Marktanteil von 2015 bis 2017 gesteigert haben, während die Unternehmen D und E Anteile verloren haben. Es zeigt sich auch, dass die Marktanteile der Unternehmen von A nach E im Jahr 2015 sequenziell zunehmen und in ähnlicher Form in 2017 abnehmen.

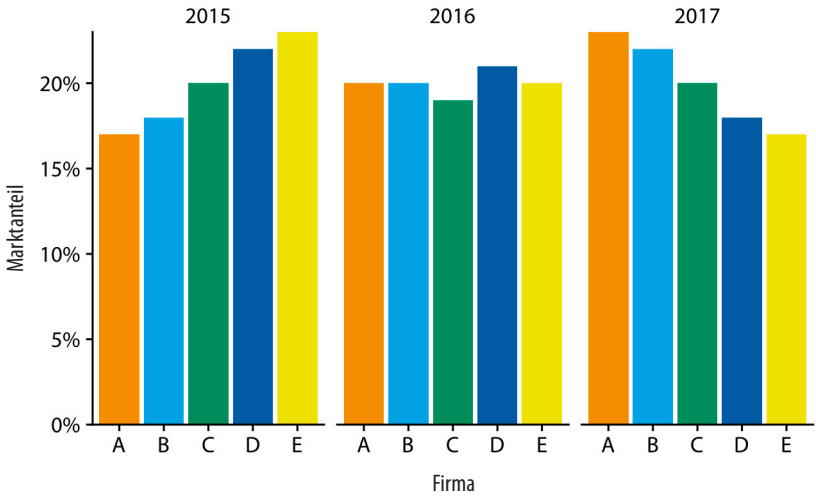

Abbildung 10-6: Marktanteil von fünf hypothetischen Unternehmen für die Jahre 2015–2017, dargestellt als nebeneinander angeordnete Balken.

Ein Fall für gestapelte Balken und gestapelte Dichten

Im vorigen Abschnitt habe ich geschrieben, dass ich normalerweise keine Sequenzen gestapelter Balken empfehle, da sich die Positionen der internen Balken entlang der Sequenz verschieben. Das Problem des Verschiebens mittlerer Balken verschwindet jedoch, wenn nur zwei Balken in jedem Stapel vorhanden sind, und in diesen Fällen kann die resultierende Visualisierung ziemlich klar sein. Betrachten Sie als Beispiel den Frauenanteil im nationalen Parlament eines Landes. Wir werden uns speziell mit dem afrikanischen Land Ruanda befassen, das ab 2016 die Liste der Länder mit dem höchsten Anteil weiblicher Abgeordneter anführt. Ruanda hat seit 2008 ein mehrheitlich weibliches Parlament und seit 2013 waren fast zwei Drittel seiner Abgeordneten weiblich. Um zu veranschaulichen, wie sich der Frauenanteil im ruandischen Parlament im Laufe der Zeit verändert hat, können wir eine Reihe gestapelter Balkendiagramme zeichnen (Abbildung 10-7). Diese Abbildung bietet eine klare visuelle Darstellung der Proportionen im zeitlichen Verlauf. Damit der Leser genau sieht, wann der Frauenanteil in der Mehrheit war, habe ich eine gestrichelte, horizontale Linie bei 50 % hinzugefügt. Ohne diese Linie wäre es nahezu unmöglich festzustellen, ob von 2003 bis 2007 die Mehrheit der Abgeordneten männlich oder weiblich war. Ich habe keine ähnlichen Linien bei 25 % und 75 % hinzugefügt, um zu vermeiden, dass die Grafik unübersichtlich wird.

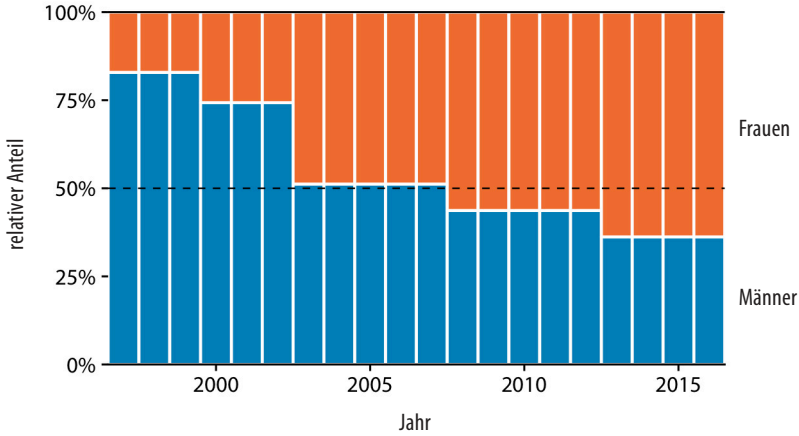

Abbildung 10-7: Veränderung der Zusammensetzung des ruandischen Parlaments nach Geschlecht im Zeitverlauf von 1997 bis 2016. Datenquelle: Interparlamentarische Union (IPU) (https://ipu.org).

Wenn wir visualisieren möchten, wie sich die Anteile als Reaktion auf eine kontinuierliche Variable ändern, können wir von gestapelten Balken zu gestapelten Dichten wechseln. Gestapelte Dichten können als der Grenzfall von unendlich vielen, unendlich kleinen nebeneinander angeordneten gestapelten Balken angesehen werden. Die Dichten in gestapelten Dichtediagrammen werden wir typischerweise, wie in Kapitel 7 beschrieben, durch Kerndichteschätzungen erhalten. In diesem Kapitel werden die Stärken und Schwächen dieser Methode grundsätzlich beleuchtet.

Als ein gutes Beispiel dafür, wann gestapelte Dichtediagramme angebracht sein können, betrachten wir hier den Gesundheitszustand von Menschen in Abhängigkeit ihres Alters. Das Alter kann als kontinuierliche Variable betrachtet werden, und die Darstellung der Daten funktioniert auf diese Weise recht gut (Abbildung 10.8). Obwohl wir an dieser Stelle vier Gesundheitskategorien haben und ich, wie schon erwähnt, kein Fan vom Stapeln mehrerer Bedingungen bin, ist diese Darstellung hier jedoch akzeptabel. Wir können sehen, dass der allgemeine Gesundheitszustand mit zunehmendem Alter abnimmt, und wir erkennen auch, dass trotz dieses Trends die Bevölkerung mehrheitlich bis zum hohen Alter bei guter oder ausgezeichneter Gesundheit bleibt.

Diese Abbildung hat jedoch ein großes Manko: Dadurch, dass die Anteile der vier Gesundheitszustände als Prozentsätze der Gesamtzahl dargestellt werden, wird verdeckt, dass der Datensatz viel mehr jüngere Menschen als ältere Menschen enthält. Somit bleibt der *Prozentsatz* der Personen, die angeben, bei guter Gesundheit zu sein, über eine angezeigte Spanne von sieben Jahrzehnten hinweg in etwa unverändert – allerdings sinkt de facto die Gesamtzahl der Menschen, deren Gesundheitszustand gut ist, da die *tatsächliche Anzahl* der Menschen in den höheren Altersgruppen abnimmt. Ich werde im nächsten Abschnitt eine mögliche Lösung für dieses Problem vorstellen.

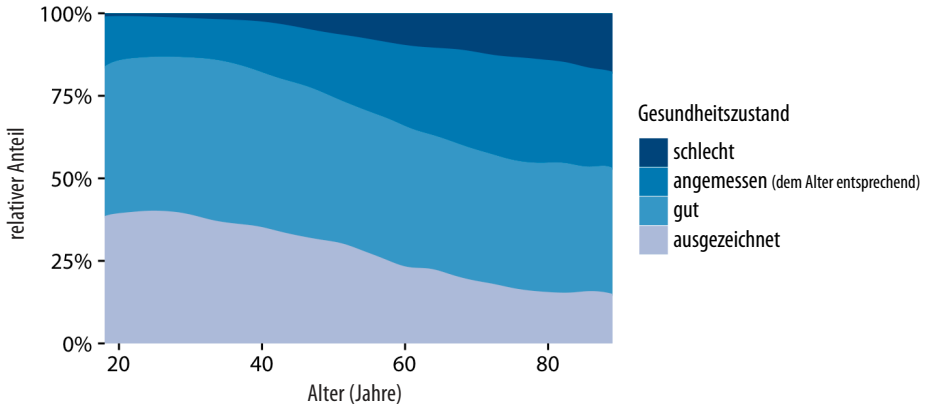

Abbildung 10-8: Gesundheitszustand nach Alter. (Datenquelle: General Social Survey [GSS])

Proportionen separat als Teile der Summe visualisieren

Nebeneinander angeordnete Balken haben das Problem, dass sie die Größe der einzelnen Teile im Verhältnis zum Ganzen nicht darstellen, und gestapelte Balken haben das Problem, dass die verschiedenen Balken nicht einfach verglichen werden können, weil sie unterschiedliche Basislinien aufweisen. Wir können diese beiden Probleme lösen, indem wir für jeden Anteil ein separates Diagramm anfertigen und in diesem den jeweiligen Anteil am Ganzen darstellen.

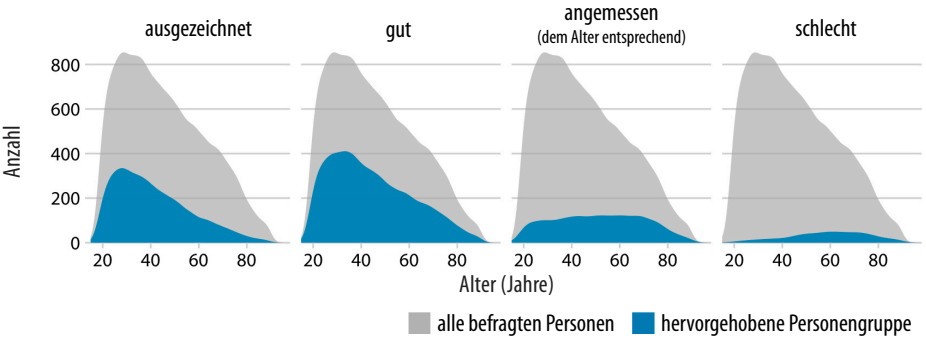

Abbildung 10-9: Gesundheitszustand nach Alter, dargestellt als Anteil an der Gesamtzahl der befragten Personen. Die farbigen Bereiche zeigen die Dichteschätzungen des Alters von Personen mit dem jeweiligen Gesundheitszustand an, der graue Bereich die Gesamtverteilung. (Datenquelle: GSS)

Für den Datensatz aus Abbildung 10-8 führt dieses Vorgehen zu Abbildung 10-9. Die gesamte Altersverteilung im Datensatz wird als grau getönter Bereich angezeigt, und die Altersverteilungen für jeden Gesundheitszustand erscheinen in Blau. Diese Abbildung zeigt, dass in absoluten Zahlen die Anzahl der Menschen mit aus-

gezeichneter oder guter Gesundheit ab einem Alter von 30 bis 40 Jahren abnimmt, während die Anzahl der Menschen, die ihre Gesundheit als gut einstuft, über alle Altersgruppen hinweg ungefähr konstant ist.

Als zweites Beispiel betrachten wir eine andere Variable aus derselben Umfrage: den Familienstand. Der Familienstand ändert sich mit dem Alter viel drastischer als der Gesundheitszustand, und eine gestapelte Dichtedarstellung des Familienstands im Vergleich zum Alter ist nicht sehr aufschlussreich (Abbildung 10-10).

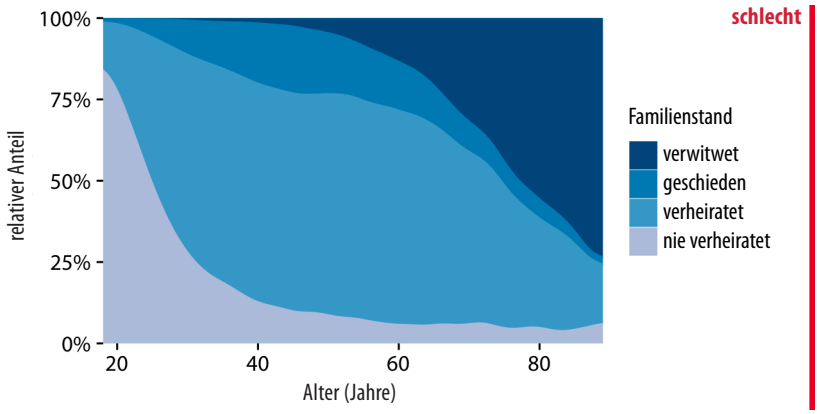

Abbildung 10-10: Familienstand nach Alter. Um die Abbildung zu vereinfachen, habe ich eine kleine Anzahl von Fällen entfernt, die sich als »getrennt lebend« definieren. Ich habe diese Abbildung als »schlecht« eingestuft, weil sich die Häufigkeit des Auftretens von Menschen, die noch nie verheiratet waren oder verwitwet sind, mit dem Alter so drastisch ändert, dass die Altersverteilung von verheirateten und geschiedenen Menschen stark verzerrt und schwer zu interpretieren ist. (Datenquelle: GSS)

Derselbe Datensatz ist, als Teildichte dargestellt, deutlich aussagekräftiger (Abbildung 10-11). Insbesondere sehen wir, dass der Anteil der Verheirateten mit Ende 30, der Anteil der Geschiedenen mit Anfang 40 und der Anteil der Verwitweten mit Mitte 70 am höchsten ist.

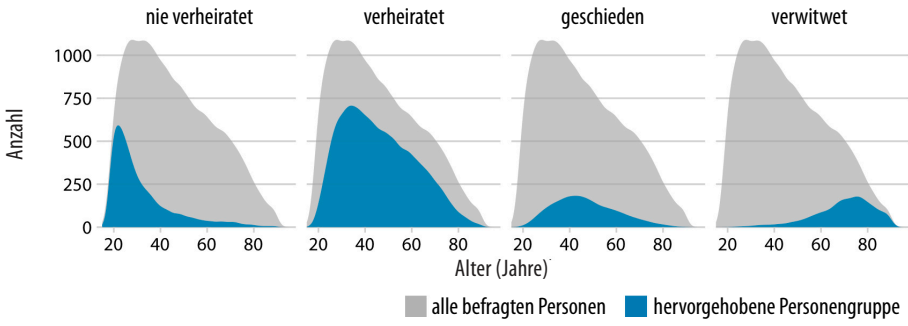

Abbildung 10-11: Familienstand nach Alter, dargestellt als Anteil an der Gesamtzahl der befragten Personen. Die farbigen Bereiche zeigen die Dichteschätzungen des Alters von Personen mit dem jeweiligen Familienstand, der graue Bereich die Gesamtverteilung. (Datenquelle: GSS)

Ein Nachteil von Abbildung 10-11 ist jedoch, dass es mit dieser Darstellung nicht einfach ist, die relativen Anteile zu einem bestimmten Zeitpunkt zu bestimmen. Wenn wir beispielsweise wissen möchten, in welchem Alter mehr als 50% aller befragten Personen verheiratet sind, können wir dies nicht einfach aus Abbildung 10-11 ableiten. Um diese Frage zu beantworten, können wir dieselbe Darstellungsform verwenden, aber entlang der *y*-Achse relative Proportionen anstelle von absoluten Zählwerten anzeigen (Abbildung 10-12). Jetzt sehen wir, dass die Mehrheit der Verheirateten ab Ende der 20er Jahre und die Mehrheit der Verwitweten ab Mitte der 70er Jahre besteht.

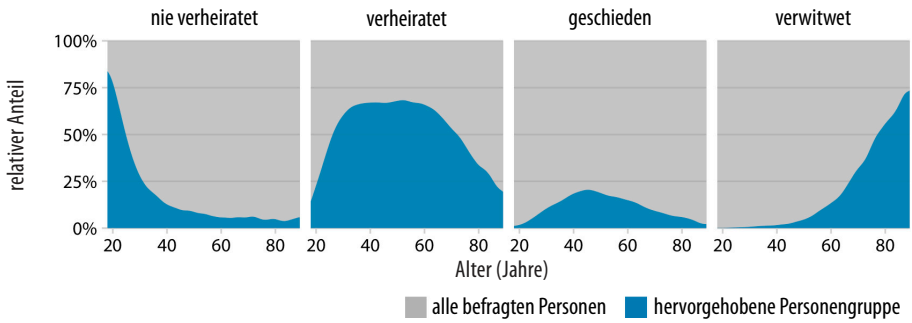

Abbildung 10-12: Familienstand nach Alter, angegeben als Anteil an der Gesamtzahl der befragten Personen. Die blau gefärbten Bereiche zeigen den Prozentsatz der Personen im gegebenen Alter mit dem jeweiligen Familienstand an, im Verhältnis zu den grau gefärbten Bereiche, welche den Prozentsatz der Personen mit allen anderen Familienständen zeigen. (Datenquelle: GSS)