Kategoriale Merkmale analysieren

Link zum Datensatz in CODAP: <https://tinyurl.com/you-pb50>

Das Analysieren von kategorialen Merkmalen bietet spannende Explorationsmöglichkeiten, kommt jedoch im Unterricht selten vor. Deshalb empfehlen wir, diese kurze Einführung zu lesen, um Explorationsmöglichkeiten und mögliche Lernschwierigkeiten einschätzen zu können.

Inhalt

[1. Einleitung 1](#_Toc190683403)

[2. Verteilungen kategorialer Merkmale untersuchen 2](#_Toc190683404)

[2.1 Kategoriale Verteilungen darstellen 2](#_Toc190683405)

[2.2 Kategoriale Verteilungen lesen und interpretieren 3](#_Toc190683406)

[3 Zusammenhänge von zwei kategorialen Merkmalen untersuchen 4](#_Toc190683407)

[3.1 Zusammenhänge zwischen zwei binären Merkmalen darstellen 4](#_Toc190683408)

[3.1.1 Verbalisierungen zur Vierfeldertafel 4](#_Toc190683409)

[3.1.2 Anteile in der Vierfeldertafel 5](#_Toc190683410)

[3.1.3 Unterschiedliche Aussagekraft der verschiedenen Analysen 7](#_Toc190683411)

[3.2 Zusammenhänge zwischen zwei Merkmalen mit mehreren Ausprägungen 7](#_Toc190683412)

[3.3 Anteilsdiagramme 8](#_Toc190683413)

[4 Zum Nachlesen 10](#_Toc190683414)

# Einleitung

Kategoriale Merkmale haben endlich viele Ausprägungen, die in der Regel durch Wörter oder Buchstabenfolgend angegeben werden. Ein Beispiel für ein kategoriales Merkmal im YOU-PB Datensatz ist *Ebook\_lesen*. Hinter diesem Merkmal steht die geschlossene Frage „Wie häufig liest du Ebooks?“ mit den vorgegebenen Antwortmöglichkeiten *täglich, mehrmals pro Woche, einmal pro Woche, einmal in 14 Tagen, einmal pro Monat, seltener, nie*. Diese Antwortmöglichkeiten kann man durch umcodieren (siehe Unterrichtsstunde 2+3) zusammenfassen, z. B. zu *Häufig* (beinhaltet die Ausprägungen *täglich, mehrmals pro Woche, einmal pro Woche)* und *Selten* (beinhaltet die Ausprägungen *einmal in 14 Tagen, einmal pro Monat, seltener, nie)*. Bei diesem speziellen Merkmal kann man eine Reihenfolge der Ausprägungen erkennen, deshalb spricht man von einem ordinalen kategorialen Merkmal.

Von Statistikprogrammen wie CODAP werden kategoriale Merkmale häufig lexikalisch oder numerisch aufsteigend sortiert. Aus diesem Grund wurden den Ausprägungen im Datensatz Ziffern vorangestellt, um die logische Reihenfolge, die den Ausprägungen zugrunde liegt, automatisch zu erzeugen in einer Visualisierung: *1\_nie, 2\_seltener, 3\_einmal pro Monat, etc*. Ohne die vorangestellten Ziffern wären die Ausprägungen alphabetisch aufsteigend sortiert worden, beginnend mit *einmal\_im\_Monat*…

Eine solche Reihenfolge muss aber nicht bei allen kategorialen Merkmalen vorliegen (die Ausprägungen des Merkmals *Augenfarbe* lässt sich schwierig ordnen). Die Ausprägungen von kategorialen Merkmalen unterliegen häufig einer gewissen Beliebigkeit, man hätte die ursprünglichen Ausprägungen auch anders wählen können (z. B. ergänzen: *einmal im halben Jahr*) oder auch anders zusammenfassen (z. B. *nie* von allen anderen abgrenzen).

Im Folgenden wird gezeigt, wie Verteilungen kategorialer Merkmale erkundet werden können.

# Verteilungen kategorialer Merkmale untersuchen

Merkmalsnamen sind nicht unbedingt selbsterklärend. Die Bezeichnung *Ebooks\_lesen* umfasst die Frage des Fragebogens nach der Häufigkeit des Lesens von Ebooks mit den oben vorgestellten Antwortmöglichkeiten. Um die Bedeutung einer Merkmalbezeichnung in einem Datensatz erfassen zu können, ist es oft hilfreich, mit der Variablenliste zu arbeiten, in der diese erklärt ist.

## 2.1 Kategoriale Verteilungen darstellen

Stellt man nun die Verteilung des Merkmals *Ebook\_lesen* mit Hilfe von CODAP dar wie in Abbildung 1, so lässt sich leicht erkennen, dass CODAP eine Ordnung der Ausprägungen vornimmt.

Ein Bild, das Text, Screenshot, Rechteck, Reihe enthält.

Automatisch generierte Beschreibung

Abbildung : Punkt-Säulendiagramm zur Verteilung des Merkmals Ebook\_lesen

Jeder Fall (=jede befragte Person) ist durch einen Punkt in Abbildung 1 repräsentiert, entsprechend seiner Ausprägung in einem gestapelten Punkt-Säulendiagramm angeordnet. Ein herkömmliches Säulendiagramm kann durch den Befehl „Punkte verschmelzen“ am Graph erzeugt werden (Abbildung 2, die Einstellung „Punkte rechteckig“ wird direkt am Graph vorgenommen) und wird durch das Einblenden von Anzahl oder Prozent noch informativer (Abbildung 3).

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

Abbildung : Säulendiagramm zur Verteilung des Merkmals Ebook\_lesen

Ein Bild, das Text, Diagramm, Software, Zahl enthält.

Automatisch generierte Beschreibung

Abbildung : Säulendiagramm zur Verteilung des Merkmals Ebook\_lesen mit eingeblendeten Anzahlen und Anteilen

Der wesentliche Unterschied von Abbildung 1 zu den Abbildungen 2 und 3 ist, dass der Aufbau der Säulen aus den einzelnen Datenpunkten in Abbildung 3 nicht mehr sichtbar ist.

Bei diesem Diagrammtyp werden fehlende Werte in CODAP nicht extra gekennzeichnet und werden bei der Berechnung als nicht existent behandelt.

## 2.2 Kategoriale Verteilungen lesen und interpretieren

Ist eine Verteilung eines kategorialen Merkmals visualisiert, so muss sie auch gelesen und interpretiert werden. Zum Lesen und Interpretieren kann man sich an den Stufen „reading the data“ (Lesen der Daten), „reading between the data“ (Lesen zwischen den Daten) und „reading beyong the data“ (Lesen über die Daten hinaus) orientieren, die auf Friel, Bright and Curcio 2001[[1]](#footnote-2) zurückgehen. Aus Abbildung 3 lässt sich beispielsweise auf drei Stufen interpretieren:

* „802 SuS haben angegeben, *nie* Ebooks zu lesen“ (erste Interpretationsstufe, da gegebene Informationen direkt abgelesen werden)
* „Kombiniert man *einmal pro Woche, mehrmals pro Woche* und *täglich* (was als Antwort nicht vorgekommen ist)*,* so erhält man 228 Jugendliche, die Ebooks *häufig* lesen (zweite Interpretationsstufe, da mathematische Fertigkeiten verwenden werden, um Beziehungen in den Daten zu entdecken)
* „Die Daten sind von 2021. Würde man heute noch einmal eine Erhebung durchführen, so würden vermutlich noch mehr Personen *nie* Ebooks lesen, weil insgesamt weniger gelesen wird“ (dritte Interpretationsstufe, weil Schlussfolgerungen über die Daten hinaus gezogen werden unter Einbezug von Hintergrundwissen)

Didaktischer Kommentar

Im Unterricht sollte das Lesen und Interpretieren von Verteilungen kategorialer Merkmale geübt werden. Erprobungen haben gezeigt, dass SuS die erste Interpretationsstufe häufig leicht fällt, die anderen beiden jedoch erfordern, dass die Verteilung „als ganzes“ in den Blick genommen wird, was eine Herausforderung für SuS darstellen kann. Dieser Herausforderung kann man durch Üben entgegenwirken.

# Zusammenhänge von zwei kategorialen Merkmalen untersuchen

Interessant ist es über die Untersuchung der Verteilung eines kategorialen Merkmals hinaus den Zusammenhang zwischen zwei kategorialen Merkmalen zu untersuchen. Im einfachsten Fall untersucht man den Zusammenhand zwischen zwei binären Merkmalen.

## 3.1 Zusammenhänge zwischen zwei binären Merkmalen darstellen

Wir wollen den Zusammenhang zwischen *Ebook\_lesen* und *Geschlecht* in den YOU-PB Daten untersuchen. Man könnte die allgemeine Hypothese aufstellen, dass Mädchen häufiger als Jungen Bücher lesen und deshalb auch häufiger Ebooks lesen. Wenn dies stimmen würde, würde man in unserer Population in den Daten auch einen Anteil unter den Mädchen, die Ebook lesen, erwarten, der höher ist als bei den Jungen.

Betrachten wir zunächst einmal die vier Teilgruppen, die entstehen und die absoluten Anzahlen der Personen in diesen Teilgruppen, um diese näher zu verstehen. Mit dem binär codierten Merkmal *Ebook\_lesen* und dem Merkmal *Geschlecht* können wir den Datensatz in zwei Gruppen (männlich/weiblich) und zwei Untergruppen (Ebook lesen häufig/selten) einteilen. Es ergeben sich vier Teilgruppen wie in der Vierfeldertafel in Abbildung 4. Schaut man zunächst einmal nur auf die absoluten Anzahlen in den vier Teilgruppen, so kann entnommen werden, dass 143 der befragten Personen weiblich sind und häufig Ebooks lesen und 56 sind männlich und lesen häufig Ebooks.

Ein Bild, das Text, Screenshot, Reihe, Rechteck enthält.

Automatisch generierte Beschreibung

Abbildung : Vierfeldertafel zur Untersuchung des Zusammenhangs zwischen Geschlecht und Ebook\_lesen

Somit gibt es absolut mehr Mädchen, die Ebooks lesen als Jungen, die Ebooks lesen in dieser Stichprobe.

### 3.1.1 Verbalisierungen zur Vierfeldertafel

Betrachtet man die Vierfeldertafel zeilenweise, so lässt sich verbalisieren: Von den weiblichen Befragten lesen 143 Ebooks häufig und 569 lesen selten Ebooks (obere Zeile). Von den männlichen Befragten (untere Zeile) lesen 56 häufig Ebooks und 505 selten. Es gibt also in beiden Gruppen mehr selten Lesende als häufig Lesende.

Betrachtet man die Vierfelder spaltenweise, so lässt sich verbalisieren: In der Teilgruppe derjenigen, die häufig Ebooks lesen (linke Spalte), sind 143 weibliche und 56 männliche Befragte.

Mit Verteilungen lässt sich das so verbalisieren:

* Zeilenweise: In der ersten Zeile findet man die Verteilung des Merkmals *Ebook\_lesen* unter den weiblichen Befragten, in der zweiten Zeile ist die Verteilung des Merkmals *Ebook\_lesen* unter den männlichen Befragten zu sehen.
* Spaltenweise: In der ersten Spalte steht die Verteilung des Merkmals *Geschlecht* in der Teilgruppe der häufig Ebook Lesenden, in der zweiten Spalte steht die Verteilung des Merkmals *Geschlecht* in der Teilgruppe der selten Ebook Lesenden.

### 3.1.2 Anteile in der Vierfeldertafel

Wir haben oben gesehen, dass es mehr weibliche Befragte als männliche innerhalb der Gruppe der häufig Ebook Lesenden gibt. Man hat allerdings auch gesehen, dass die Gesamtgruppe der weiblichen Befragten 143+569=712 Personen umfasst und die der männlichen nur 56+505=561. Wenn die Gruppen nicht gleich groß sind, müssen Anteile verglichen werden.

#### 3.1.2.1 Spaltenprozente

Betrachten wir den Anteil der weiblichen unter den häufig Ebook Lesenden. Dieser beträgt 143/(143+56)72%. Der Anteil der männlichen unter den häufig Ebook Lesenden beträgt 56/(143+56)28%, ist also deutlich geringer. Da diese Prozente spaltenbezogen ermittelt wurden, spricht man auch von **Spaltenprozenten**. Die oben geäußerte Vermutung ist also für die Befragten dieser Stichprobe bestätigt.

In CODAP lassen sich Spaltenprozente über das Menü einblenden wie in Abbildung 5.

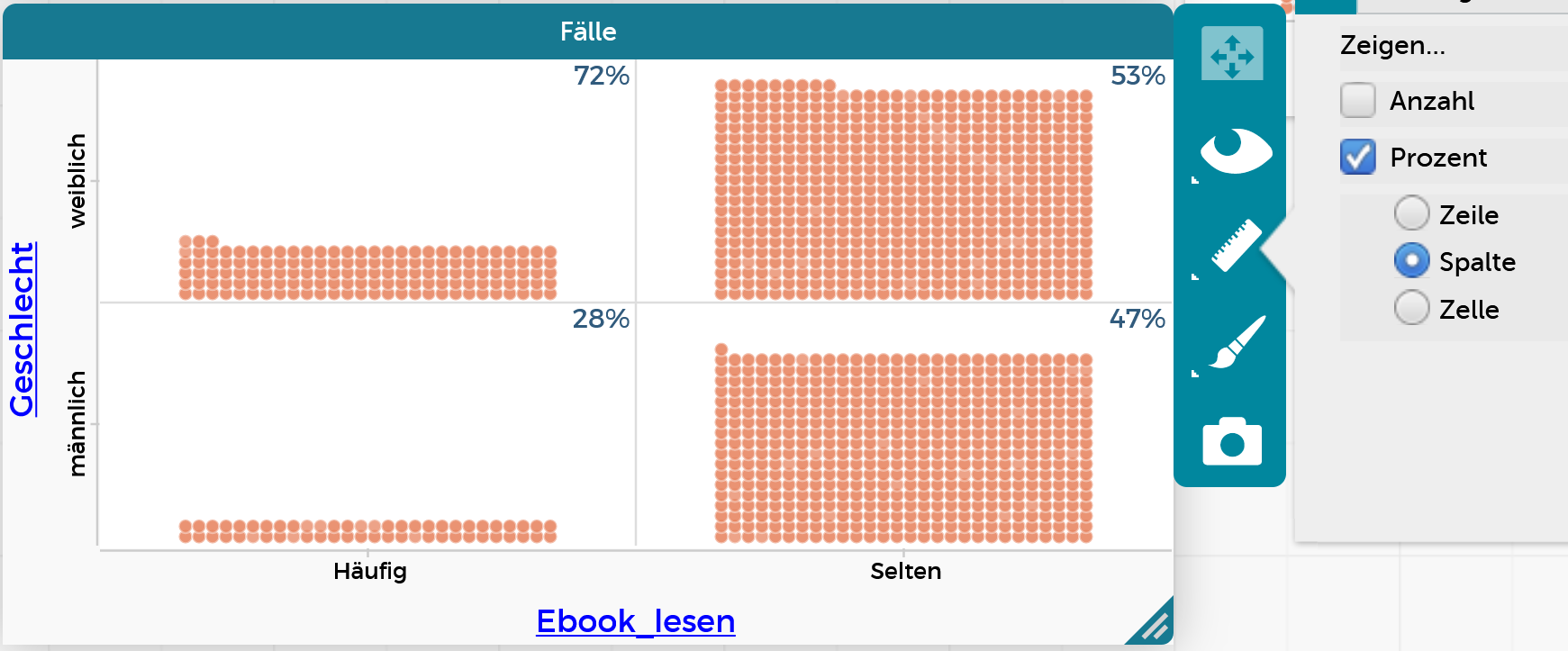


Abbildung : Vierfeldertafel mit eingeblendeten Spaltenprozenten

#### 3.1.2.2 Zeilenprozente

Kommen wir zurück auf die Ausgangshypothese und fragen danach, ob unter den weiblichen Befragten eher häufig Ebook Lesende zu finden sind als unter den männlichen Befragten, so muss die Analyse geändert werden. Bezogen auf die Darstellung in Abbildung 4 entspricht nun eine Zeile der Bezugsgruppe. Dementsprechend werden für die Auswertung Zeilenprozente benötigt wie in Abbildung 6.

Ein Bild, das Text, Screenshot, Zahl, Software enthält.

Automatisch generierte Beschreibung

Abbildung : Vierfeldertfel mit eingeblendeten Zeilenprozenten

In Abbildung 6 lässt sich nun verbalisieren: von den weiblichen Befragten (obere Zeile) lesen 20% häufig Ebooks und 80% tun dies selten. Bei den männlichen Befragten (untere Zeile) ist dies noch weniger: nur 10% lesen häufig Ebooks, und 90% tun dies selten. Also kann man als Ergebnis formulieren, dass die Ausgangshypothese für die Befragten in dieser Stichprobe zutrifft, da ein größerer Anteil (nämlich 20%) der weiblichen Befragten häufig Ebooks liest, denn bei den männlichen Befragten beträgt dieser Anteil nur 10%.

#### 3.1.2.3 Zellenprozente

Schließlich gibt es noch eine weitere Perspektive, die eingenommen werden kann. Fragt man nach dem Anteil derjenigen, die häufig Ebooks lesen UND weiblich sind unter allen Befragten, so sind alle Befragte die Bezugsgruppe. Entsprechende Anteile kann man sich in CODAP mit Zellenprozenten anzeigen lassen wie in Abbildung 7.

Ein Bild, das Text, Screenshot, Zahl, Rechteck enthält.

Automatisch generierte Beschreibung

Abbildung : Vierfeldertafel mit eingeblendeten Zellenprozenten

Für Abbildung 7 kann verbalisisert werden: 11% aller Befragten sind weiblich und lesen häufig Ebooks, 4,4% aller Befragten sind männlich und lesen häufig Ebooks.

### 3.1.3 Unterschiedliche Aussagekraft der verschiedenen Analysen

Welche der drei gefundenen Ergebnisse passt aus Ihrer Sicht am besten, um den Zusammenhang zwischen Geschlecht und Ebook lesen zu beschreiben? Hier sind noch einmal die drei Befunde bezogen auf die weiblichen und männlichen Personen und die häufig Ebook Lesenden.

Spaltenprozente liefern: Der Anteil der weiblichen Personen unter den häufig Ebook Lesenden beträgt **72%.** Demgegenüber steht der Anteil der männlichen Personen unter den häufig Ebook Lesenden mit **28%.**

Zeilenprozente liefern: Von den weiblichen Befragten lesen **20%** häufig Ebooks. Bei den männlichen Befragten lesen **10%** häufig Ebooks.

Zellenprozente liefern: **11%** aller Befragten sind weiblich und lesen häufig Ebooks, **4,4%** aller Befragten sind männlich und lesen häufig Ebooks.

Didaktischer Kommentar

Man beachte, dass in den Abbildungen 5 bis 7 immer die gleichen Daten dargestellt sind. Die absoluten Anzahlen der Teilgruppen ändern sich nicht. Allerdings ist der Unterschied der Anteile deutlich, je nachdem welcher Bezug (Geschlecht, Häufigkeit des Ebook lesens oder alle Befragte) gewählt wird. Deshalb ist es wichtig, dies ausführlich mit SuS zu besprechen, für welche Frage welche Prozente die richtige Auswertung darstellen. Viele Erprobungen und Forschung in der Statistikdidaktik haben gezeigt, dass SuS hier besondere Schwierigkeiten haben.

## 3.2 Zusammenhänge zwischen zwei Merkmalen mit mehreren Ausprägungen

Hätte man nun das Merkmal *Ebook\_lesen* nicht binär zusammengefasst, sondern mit allen sieben Ausprägungen betrachtet, so ergibt sich eine Mehrfeldertafel wie in Abbildung 8. Diese Abbildung ist deutlich schwieriger zu interpretieren und wird sogar noch schwieriger, wenn beide Merkmale mehrere Ausprägungen haben.

Ein Bild, das Text, Screenshot, Zahl, Reihe enthält.

Automatisch generierte Beschreibung

Abbildung : Mehrfeldertafel zu Geschlecht und Ebook\_lesen mit eingeblendeten Zeilenprozenten

Nun muss sorgfältig jeder einzelne Prozentwert interpretiert werden. Zum Beispiel links oben: 55% der weiblichen Befragten in der Stichprobe lesen *nie* Ebooks. Unter den männlichen Befragten beträgt dieser Anteil sogar 73%.

Fragt man stattdessen nach dem Zusammenhang zwischen den Merkmalen *Ebook\_lesen* und *Podcast\_hören*, so ergibt sich eine Mehrfeldertafel mit 49 Feldern wie in Abbildung 9.

Ein Bild, das Text, Screenshot, Zahl, Schrift enthält.

Automatisch generierte Beschreibung

Abbildung : Mehrfeldertafel zu Ebook\_lesen und Podcast\_hören mit eingeblendeten Zeilenprozenten

In Abbildung 9 lassen sich zwar noch alle Anteile sinnvoll verbalisieren, allerdings bringt das wenig. Hier sucht man am besten nach auffälligen Gruppen. Dies könnte zum Beispiel die Zelle ganz links unten sein: Unter denjenigen, die nie Podcasts hören (Zeilenprozente!), lesen 72% nie Ebooks.

Die Zeilenprozente erlauben eigentlich einen Verteilungsvergleich in den sieben Teilgruppen des Merkmals *Podcasts hören*, aber der wird visuell dadurch gestört, dass absolute Häufigkeiten dargestellt werden. Für einen besseren visuellen Verteilungsvergleich muss man Anteilsdiagramme heranziehen.

## 3.3 Anteilsdiagramme

Anteilen kann in CODAP in zwei Schritten visualisiert werden. Nehmen wir dazu noch einmal die Graphik aus Abbildung 8, in der nach dem Zusammenhang von *Geschlecht* und *Ebook\_lesen* geschaut wurde. Man setzt also ein binäres Merkmal (*Geschlecht*) mit einem Merkmal mit sieben Ausprägungen in Beziehung, wobei nach den Anteilen der Ausprägungen von *Ebook lesen* innerhalb der Gruppe der weiblichen und innerhalb der Gruppe der männlichen Befragten geschaut wurde.

Diese Anteile kann man in einem rechteckbasierten Anteilsdiagramm visualisieren und somit leichter vergleichen. Dazu benötigt man in CODAP zwei Schritte. Zuerst visualisiert man die Verteilung des gruppierenden Merkmals, hier *Geschlecht*, und verschmelzt *prozentual* (Abbildung 10).

Ein Bild, das Text, Screenshot, Software enthält.

Automatisch generierte Beschreibung

Abbildung : Verteilung des Merkmals Geschlecht als Balkendiagramm, wobei die Länge der Balken den entsprechenden Anteil der Ausprägung an der Gesamtanzahl repräsentiert

Dann zieht man das Merkmal *Ebook\_lesen* in die Mitte der Graphik in Abbildung 10, so dass die entsprechenden Anteile des Merkmals *Ebook\_lesen* innerhalb der Teilgruppe der weiblichen Befragten (oberer Balken) und der männlichen Befragten (unterer Balken) dargestellt werden (Abbildung 11).

Hinweis zu fehlenden Werten in rechteckbasierten Anteilsdiagrammen

Nur in rechteckbasierten Anteilsdiagrammen werden fehlende Werte wie folgt berücksichtigt. Wenn das gruppierende Merkmal fehlende Werte enthält, finden diese keinerlei Beachtung in der Darstellung und der Berechnung der Anteile. Wenn allerdings das Merkmal, nach welchem ausgewertet wird, fehlende Werte enthält, so werden diese als eigener Anteil grauer Farbe entsprechend dargestellt und in die Berechnung mit einbezogen.

Ein Bild, das Text, Screenshot, Software, Zahl enthält.

Automatisch generierte Beschreibung

Abbildung : Rechteckbasierte Anteilsgraphik zum Zusammenhang zwischen Geschlecht und Ebook\_lesen, eingefärbt nach Ebook\_lesen

In CODAP kann man mit der Maus über die entsprechenden Anteile fahren, um deren konkreten Wert abzulesen. In Abbildung 11 ist dies exemplarisch für den Anteil derjenigen geschehen, die nie Ebooks lesen in der Gruppe der weiblichen Befragten (lila Anteil rechts oben). Dieser Anteil liegt (wie auch schon in Abbildung 8) bei etwa 55%. Dies kann man in der Abbildung 11 einfacher als in Abbildung 8 mit dem entsprechenden Anteil in der Gruppe der männlichen Befragten vergleichen. Der lilafarbene Anteil ist in der Gruppe der männlichen Befragten deutlich größer als in der Gruppe der weiblichen Befragten und so lässt sich auch hieraus schlussfolgern, dass männliche Befragte in dieser Stichprobe eher dazu tendieren, nie Ebooks zu lesen als weibliche Befragte.

Didaktischer Hinweis

Die Darstellung bietet einen wesentlichen Fortschritt beim Verteilungsvergleich gegenüber Abbildung 8 (und auch gegenüber Abbildung 9), auch wenn die Grafik komplex ist und sorgfältig interpretiert werden muss.

## 4 Zum Nachlesen

Für eine hervorragende Vertiefung dieses Themas empfehlen wir das Buch „Daten-Spürnasen auf Spurensuche“ von Daniel Frischemeier und Rolf Biehler, erschienen bei Klett|Kallmeyer 2024: <https://www.friedrich-verlag.de/shop/daten-spuernasen-auf-spurensuche-31448>. Die obige Darstellung ist an dieses Buch angelehnt.

1. Friel, S. N., Curcio, F. R. & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158. https://doi.org/10.2307/749671 [↑](#footnote-ref-2)