Analyze categorical variables

Link to the data record in CODAP: <https://tinyurl.com/you-pb50>

Note: This document has been automatically translated using DeepL.com. As a result, some content may not be fully accurate or correctly displayed.

Analyzing categorical variables offers exciting opportunities for exploration, but rarely occurs in the classroom. We therefore recommend reading this short introduction in order to be able to assess exploration possibilities and possible learning difficulties.

Contents

[1. Introduction 1](#_Toc190691454)

[2. Examine distributions of categorical variables 2](#_Toc190691455)

[2.1 Representing categorical distributions 2](#_Toc190691456)

[2.2 Reading and interpreting categorical distributions 3](#_Toc190691457)

[3 Examine the relationships between two categorical variables 4](#_Toc190691458)

[3.1 Representing relationships between two binary variables 4](#_Toc190691459)

[3.1.1 Verbalizations for the four-field table 4](#_Toc190691460)

[3.1.2 Shares in the four-field table 5](#_Toc190691461)

[3.1.3 Differences in the significance of the various analyses 7](#_Toc190691462)

[3.2 Relationships between two variables with multiple expressions 7](#_Toc190691463)

[3.3 Share diagrams 8](#_Toc190691464)

[4 For further reading 10](#_Toc190691465)

# Introduction

Categorical variables have a finite number of values, which are usually indicated by words or letter sequences. An example of a categorical variable in the YOU-PB data set is *Ebook\_reading*. Behind this variable is the closed question "How often do you read ebooks?" with the given answer options *daily, several times a week, once a week, once a fortnight, once a month, less often, never*. These answer options can be summarized by recoding (see lesson 2+3), e.g. to *Frequently* (includes the options *daily, several times a week, once a week)* and *Rarely* (includes the options *once a fortnight, once a month, less often, never)*. With this particular variable, it is possible to recognize an order of the values, which is why it is referred to as an ordinal categorical variable.

Statistical programs such as CODAP often sort categorical variables lexically or numerically in ascending order. For this reason, the characteristics in the data set were preceded by numbers in order to automatically generate the logical order on which the characteristics are based in a visualization: *1\_never, 2\_once a month, 3\_once a month, etc*. Without the preceding digits, the characteristics would have been sorted alphabetically in ascending order, starting with *once\_per\_month*...

However, not all categorical variables have to be ordered in this way (it is difficult to order the values of the variable *eye color*). The expressions of categorical variables are often subject to a certain arbitrariness; the original expressions could have been chosen differently (e.g. add: *once every six months*) or summarized differently (e.g. *never* separate from all others).

The following shows how distributions of categorical variables can be explored.

# Examine distributions of categorical variables

Feature names are not necessarily self-explanatory. The variable *Ebooks\_reading* covers the questionnaire's question about the frequency of reading ebooks with the answer options presented above. In order to understand the meaning of a characteristic name in a data set, it is often helpful to work with the variable list in which it is explained.

## 2.1 Representing categorical distributions

If we now display the distribution of the variable *Ebook\_reading* with the help of CODAP as in Figure 1, it is easy to see that CODAP orders the values.

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

Automatisch generierte Beschreibung

Figure 1: Point-column diagram for the distribution of the variable Ebook\_reading

Each case (=each person interviewed) is represented by a point in Figure 1, arranged according to its characteristics in a stacked point-column chart. A conventional bar chart can be created using the "Merge points" command on the graph (Figure 2, the "Rectangular points" setting is made directly on the graph) and becomes even more informative by displaying the number or percentage (Figure 3).

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

Automatisch generierte Beschreibung

Figure 2: Bar chart showing the distribution of the variable Ebook\_reading

Ein Bild, das Text, Software, Diagramm, Multimedia-Software enthält.

Automatisch generierte Beschreibung

Figure 3: Bar chart showing the distribution of the variable Ebook\_reading with numbers and proportions shown

The main difference between Figure 1 and Figures 2 and 3 is that the structure of the columns from the individual data points is no longer visible in Figure 3.

With this diagram type, missing values are not specially marked in CODAP and are treated as non-existent in the calculation.

## 2.2 Reading and interpreting categorical distributions

Once a distribution of a categorical variable has been visualized, it must also be read and interpreted. For reading and interpreting, one can be guided by the stages "reading the data", "reading between the data" and "reading beyond the data", which go back to Friel, Bright and Curcio 2001[[1]](#footnote-2) . Figure 3, for example, can be interpreted in three stages:

* "802 students stated that they *never* read e-books" (first level of interpretation, as given information is read directly)
* "Combining *once a week, several times a week* and *daily* (which did not occur as an answer), we get 228 young people who read ebooks *frequently* (second level of interpretation, as math skills are used to discover relationships in the data)
* "The data is from 2021. If a survey were to be conducted again today, even more people would probably *never* read ebooks because less is being read overall" (third level of interpretation, because conclusions are drawn beyond the data, taking background knowledge into account)

Didactic commentary

Reading and interpreting distributions of categorical variables should be practiced in class. Tests have shown that pupils often find the first level of interpretation easy, but the other two require them to look at the distribution "as a whole", which can be a challenge for pupils. This challenge can be counteracted by practicing.

# Examine the relationships between two categorical variables

In addition to examining the distribution of a categorical variable, it is interesting to examine the relationship between two categorical variables. In the simplest case, one examines the relationship between two binary variables.

## 3.1 Representing relationships between two binary variables

We want to investigate the relationship between *ebook reading* and *gender* in the YOU-PB data. One could make the general hypothesis that girls read books more often than boys and therefore read ebooks more often. If this were true, one would also expect a higher proportion of girls in our population to read ebooks than boys.

Let's first look at the four subgroups that emerge and the absolute numbers of people in these subgroups to understand them in more detail. Using the binary coded variable *Ebook\_reading* and the variable *gender*, we can divide the data set into two groups (male/female) and two subgroups (ebook reading often/rarely). This results in four subgroups as shown in the four-field table in Figure 4. If we first look only at the absolute numbers in the four subgroups, we can see that 143 of the respondents are female and read ebooks frequently and 56 are male and read ebooks frequently.

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

Automatisch generierte Beschreibung

Figure 4: Four-field table to investigate the relationship between gender and ebook\_reading

This means that there are absolutely more girls who read ebooks than boys who read ebooks in this sample.

### 3.1.1 Verbalizations for the four-field table

Looking at the four-field table row by row, it can be verbalized: Of the female respondents, 143 read ebooks frequently and 569 rarely read ebooks (top row). Of the male respondents (bottom row), 56 read ebooks frequently and 505 rarely. There are therefore more infrequent readers than frequent readers in both groups.

Looking at the four fields column by column, it can be verbalized: In the subgroup of those who frequently read ebooks (left column), there are 143 female and 56 male respondents.

This can be verbalized with distributions:

* Row by row: The first row shows the distribution of the variable *Ebook\_reading* among female respondents, the second row shows the distribution of the variable *Ebook\_reading* among male respondents.
* Column by column: The first column shows the distribution of the variable *gender* in the subgroup of frequent ebook readers, the second column shows the distribution of the variable *gender* in the subgroup of infrequent ebook readers.

### 3.1.2 Shares in the four-field table

We have seen above that there are more female respondents than male respondents within the group of frequent ebook readers. However, we have also seen that the total group of female respondents comprises 143+569=712 people and that of males only 56+505=561. If the groups are not the same size, proportions must be compared.

#### 3.1.2.1 Column percentages

Let's look at the proportion of females among frequent ebook readers. This is 143/(143+56) 72%. The proportion of males among frequent ebook readers is 56/(143+56) 28%, which is significantly lower. As these percentages were determined in relation to columns, they are also referred to as **column percentages**. The assumption made above is therefore confirmed for the respondents in this sample.

In CODAP, column percentages can be displayed via the menu as shown in Figure 5.

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

Automatisch generierte Beschreibung

Figure 5: Four-field table with displayed column percentages

#### 3.1.2.2 Row percentages

If we return to the initial hypothesis and ask whether female respondents are more likely to be ebook readers than male respondents, the analysis must be changed. In relation to the representation in Figure 4, one row now corresponds to the reference group. Accordingly, row percentages are required for the analysis as in Figure 6.

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

Automatisch generierte Beschreibung

Figure 6: Four-field panel with displayed row percentages

Figure 6 now verbalizes: of the female respondents (top row), 20% read ebooks frequently and 80% do so rarely. This figure is even lower for the male respondents (bottom row): only 10% read ebooks frequently and 90% do so rarely. It can therefore be concluded that the initial hypothesis is true for the respondents in this sample, as a larger proportion (namely 20%) of female respondents read ebooks frequently, as this proportion is only 10% for male respondents.

#### 3.1.2.3 Cell percentages

Finally, there is another perspective that can be taken. If you ask about the proportion of those who frequently read ebooks AND are female among all respondents, then all respondents are the reference group. Corresponding proportions can be displayed in CODAP with cell percentages as in Figure 7.

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

Automatisch generierte Beschreibung

Figure 7: Four-field table with cell percentages displayed

The following can be verbalized for Figure 7: 11% of all respondents are female and frequently read ebooks, 4.4% of all respondents are male and frequently read ebooks.

### 3.1.3 Differences in the significance of the various analyses

Which of the three findings do you think best describes the relationship between gender and ebook reading? Here are the three findings in relation to female and male persons and frequent ebook readers.

column percentages: The proportion of females among frequent ebook readers **is 72%.** In contrast, the proportion of males among frequent ebook readers is **28%.**

Provide row percentages: **20%** of female respondents read ebooks frequently. Among male respondents, **10%** read ebooks frequently.

Cell percentages provide: **11%** of all respondents are female and frequently read ebooks, **4.4%** of all respondents are male and frequently read ebooks.

Didactic commentary

Note that Figures 5 to 7 always show the same data. The absolute numbers of the subgroups do not change. However, the difference in the percentages is clear, depending on which reference (gender, frequency of ebook reading or all respondents) is chosen. It is therefore important to discuss this in detail with pupils, for which question which percentages represent the correct evaluation. Many tests and research in statistics didactics have shown that pupils have particular difficulties here.

## 3.2 Relationships between two variables with multiple expressions

If the variable *Ebook\_reading* had not been summarized in binary form, but had been considered with all seven values, the result would be a multi-field table as shown in Figure 8. This figure is much more difficult to interpret and becomes even more difficult if both variables have several values.

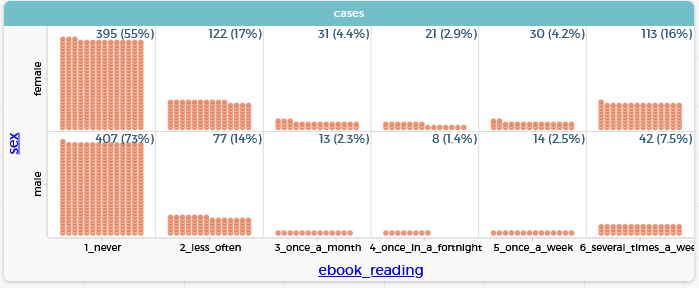


Figure 8: Multi-field table on gender and ebook\_reading with row percentages displayed

Each individual percentage value must now be interpreted carefully. For example, top left: 55% of the female respondents in the sample *never* read ebooks. Among male respondents, this proportion is as high as 73%.

If we instead ask about the relationship between the variables *Ebook\_reading* and *Podcast\_listening*, we get a multi-field table with 49 fields as shown in Figure 9.

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

Automatisch generierte Beschreibung

Figure 9: Multi-field table for Ebook\_reading and Podcast\_listen with displayed row percentages

In Figure 9, all proportions can still be verbalized in a meaningful way, but this is of little use. The best thing to do here is to look for conspicuous groups. This could be the cell at the bottom left, for example: Among those who never listen to podcasts (row percentages!), 72% never read ebooks.

The row percentages actually allow a distribution comparison in the seven subgroups of the variable *listening to podcasts*, but this is visually disturbed by the fact that absolute frequencies are shown. For a better visual distribution comparison, you have to use proportion diagrams.

## 3.3 Share diagrams

proportions can be visualized in CODAP in two steps. Let's take the graph from Figure 8 again, in which we looked at the relationship between *gender* and *Ebook\_reading*. In other words, a binary variable (*gender)* is related to a variable with seven values, whereby the proportions of the values of *ebook reading* within the group of female and within the group of male respondents were looked at.

These proportions can be visualized in a rectangle-based proportion diagram, making it easier to compare them. This requires two steps in CODAP. First, the distribution of the grouping variable, in this case *gender*, is visualized and merged *as a percentage* (Figure 10).

Ein Bild, das Text, Screenshot, Software, Multimedia-Software enthält.

Automatisch generierte Beschreibung

Figure 10: Distribution of the variable gender as a bar chart, where the length of the bars represents the corresponding share of the variable in the total number

The variable *Ebook\_reading* is then dragged to the center of the graph in Figure 10, so that the corresponding proportions of the variable *Ebook\_reading* within the subgroup of female respondents (upper bar) and male respondents (lower bar) are shown (Figure 11).

Note on missing values in rectangle-based proportion charts

Missing values are only taken into account in rectangle-based proportion charts as follows. If the grouping variable contains missing values, these are not taken into account in the display and calculation of the proportions. However, if the variable being evaluated contains missing values, these are displayed as a separate proportion in grey and included in the calculation.

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

Automatisch generierte Beschreibung

Figure 11: Rectangle-based proportion graph of the relationship between gender and Ebook\_readinging, colored by Ebook\_readinging

In CODAP, you can move the mouse over the corresponding percentages to see their specific value. In Figure 11, this is done as an example for the proportion of those who never read ebooks in the group of female respondents (purple proportion top right). This proportion is around 55% (as in Figure 8). This can be compared more easily in Figure 11 than in Figure 8 with the corresponding proportion in the group of male respondents. The purple proportion is significantly larger in the group of male respondents than in the group of female respondents, and so it can also be concluded from this that male respondents in this sample tend to never read ebooks more than female respondents.

Didactic note

The illustration offers a significant improvement in the distribution comparison compared to Figure 8 (and also compared to Figure 9), even if the graph is complex and must be interpreted carefully.

## 4 For further reading

For an excellent in-depth look at this topic, we recommend the book "Daten-Spürnasen auf Spurensuche" by Daniel Frischemeier and Rolf Biehler, published by Klett|Kallmeyer 2024: <https://www.friedrich-verlag.de/shop/daten-spuernasen-auf-spurensuche-31448>. The above presentation is based on this book.

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)