ETH zürich

Schweizer Umweltpanel
Panel suisse de l'environnement
Panel svizzero dell'ambiente

# Codebook functions (cb__) of sep R package

**Written by Janek Bruker**

*Last modified 25/02/2022*

# Contents

```r
# Packages needed to load data and run functions
library(readxl)
library(sep)
library(haven)
library(knitr)
library(tidyverse)

# Load metadata
upanelmeta <- read_excel("demo-data/upanel_w7_metadata.xlsx")

# Load survey responses
upanel = read_dta("demo-data/upanel_w7.dta") %>% zap_labels()
```

```r
knitr::opts_chunk$set(options(
  # Leave empty space in table if NA
  knitr.kable.NA = " "),
  # Not displaying anything except output
  echo=TRUE,
  warning=FALSE,
  message=FALSE,
  # Display tables (and not latex code)
  results = 'asis',
  # Setting for summary stats
  fig.width = 10,
  fig.height = 4)
```

It is very important to adapt to the structure of your metadata. The vectors numerically indicate the position of the column groups. The codebook functions rely on the correct indication of these columns.

```r
# General variable information (e.g., question type, instruction text)
meta = 1:13

# German value labels
codes_de = 14:37

# English value labels
codes_en = 38:61

# Missing labels
miscodes = 62:70
```

# Examples of introduction text

Writing in Rmarkdown, outside of the chunks, you can add text sections. For instance, add a text on the missing code scheme…

## Missing Code Scheme

The number of hashes in front of the headers determines the level of the header. The first header on this page uses one hash, the second uses two hashes. This difference becomes visible in size and in the table of contents.

```
cb_mistable(metadata = upanelmeta)
```

| | |
|---|---|
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |
| -99 | item nonresponse |

# Examples of codebook table functions

## Question 1: Year of birth

```r
# This function prints one correctly formatted page of the codebook.
# It includes a page break at the end of the page.
# Use a numeric indicator to point at the variable.
cb_page(metadata = upanelmeta, num.var = 1, comment = "Add a comment to point at
        particularities of this variable that do not become clear from the
        table.")
```

### w7_q1

| | |
|---|---|
| Variable name | w7_q1 |
| Variable label | birthyear |
| Dataset | upanel_w7 |
| Item source | Umweltpanel Welle 7 |
| Question type | Open Text Box (numeric) |
| Intro text | |
| Intro text (EN) | |
| Instruction | Bitte Geburtsjahr eintragen |
| Instruction (EN) | Please enter year of birth |
| Question Text | Wann wurden Sie geboren? |
| Question Text (EN) | When were you born? |
| Item Text | |
| Item Text (EN) | |
| *Missing Labels* | |
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |
| -99 | item nonresponse |

Add a comment to point at particularities of this variable that do not become clear from the table.

## Question 7: Children in household

```
# This function prints multiple correctly formatted page of the codebook.
# It includes a page break at the end of the page.
# Use a vector of numeric indicators to point at the variables.
# Define comment in character vector that you add to argument.
comment_q7 = c("When refering to another
        variable in this comment, it is best practice to link the variable.
        For instance, I can refer to the [Year of birth](#year).")
cb_pages(metadata = upanelmeta, multi.vars = 9:12, comment = comment_q7)
```

### w7_q7x1

| | |
|---|---|
| Variable name | w7_q7x1 |
| Variable label | child1 |
| Dataset | upanel_w7 |
| Item source | Umweltpanel Welle 7 |
| Question type | Single Choice |
| Intro text | |
| Intro text (EN) | |
| Instruction | |
| Instruction (EN) | |
| Question Text | Haben Sie eigene oder adoptierte Kinder unter 18 Jahren? |
| Question Text (EN) | Do you have children of your own or adopted children under 18? |
| Item Text | |
| Item Text (EN) | |
| **Value Labels** | |
| 1 | Nein |
| 2 | Ja |
| **Value Labels (EN)** | |
| 1 (EN) | No |
| 2 (EN) | Yes |
| *Missing Labels* | |
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |
| -99 | item nonresponse |

When refering to another variable in this comment, it is best practice to link the variable. For instance, I can refer to the Year of birth.

### w7_q7x2

| | |
|---|---|
| Variable name | w7_q7x2 |
| Variable label | child2 |
| Dataset | upanel_w7 |
| Item source | Umweltpanel Welle 7 |
| Question type | Online: Single Choice; Paper: Open Text Box (numeric) |
| Intro text | |
| Intro text (EN) | |
| Instruction | |
| Instruction (EN) | |
| Question Text | Wie viele eigene oder adoptierte Kinder unter 18 Jahren haben Sie? |
| Question Text (EN) | How many children of your own or adopted children under 18 do you have? |
| Item Text | |
| Item Text (EN) | |
| **Value Labels** | |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| 8 | 8 |
| 9 | 9 |
| 10 | 10 |
| 11 | Mehr als 10 Kinder unter 18 |
| **Value Labels (EN)** | |
| 1 (EN) | 1 |
| 2 (EN) | 2 |
| 3 (EN) | 3 |
| 4 (EN) | 4 |
| 5 (EN) | 5 |
| 6 (EN) | 6 |
| 7 (EN) | 7 |
| 8 (EN) | 8 |
| 9 (EN) | 9 |
| 10 (EN) | 10 |
| 11 (EN) | More than 10 children under 18 |
| *Missing Labels* | |
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |

| | |
|---|---|
| -99 | item nonresponse |

When refering to another variable in this comment, it is best practice to link the variable. For instance, I can refer to the Year of birth.

### w7_q7x3

| | |
|---|---|
| Variable name | w7_q7x3 |
| Variable label | child3 |
| Dataset | upanel_w7 |
| Item source | Umweltpanel Welle 7 |
| Question type | Single Choice |
| Intro text | |
| Intro text (EN) | |
| Instruction | |
| Instruction (EN) | |
| Question Text | Haben Sie eigene oder adoptierte Kinder über 18 Jahren? |
| Question Text (EN) | Do you have children of your own or adopted children over 18? |
| Item Text | |
| Item Text (EN) | |
| **Value Labels** | |
| 1 | Nein |
| 2 | Ja |
| **Value Labels (EN)** | |
| 1 (EN) | No |
| 2 (EN) | Yes |
| *Missing Labels* | |
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |
| -99 | item nonresponse |

When refering to another variable in this comment, it is best practice to link the variable. For instance, I can refer to the Year of birth.

### w7_q7x4

| | |
|---|---|
| Variable name | w7_q7x4 |
| Variable label | child4 |
| Dataset | upanel_w7 |
| Item source | Umweltpanel Welle 7 |
| Question type | Online: Single Choice; Paper: Open Text Box (numeric) |
| Intro text | |
| Intro text (EN) | |
| Instruction | |
| Instruction (EN) | |
| Question Text | Wie viele eigene oder adoptierte Kinder über 18 Jahren haben Sie? |
| Question Text (EN) | How many children of your own or adopted children over 18 do you have? |
| Item Text | |
| Item Text (EN) | |
| **Value Labels** | |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| 8 | 8 |
| 9 | 9 |
| 10 | 10 |
| 11 | Mehr als 10 Kinder über 18 |
| **Value Labels (EN)** | |
| 1 (EN) | 1 |
| 2 (EN) | 2 |
| 3 (EN) | 3 |
| 4 (EN) | 4 |
| 5 (EN) | 5 |
| 6 (EN) | 6 |
| 7 (EN) | 7 |
| 8 (EN) | 8 |
| 9 (EN) | 9 |
| 10 (EN) | 10 |
| 11 (EN) | More than 10 children over 18 |
| *Missing Labels* | |
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |

| | |
|---|---|
| -99 | item nonresponse |

When refering to another variable in this comment, it is best practice to link the variable. For instance, I can refer to the Year of birth.

```r
# This function prints the metadata table for one variable.
# It does not include a page break at the end.
# You can add a page break manually.
# It does not include a comment option.
# You can print a comment below the chunk.
cb_table(metadata = upanelmeta, num.var = 73)
```

### w7_q19x1

| | |
|---|---|
| Variable name | w7_q19x1 |
| Variable label | biodiv1 |
| Dataset | upanel_w7 |
| Item source | Umweltpanel Welle 7 |
| Question type | Single Choice |
| Intro text | |
| Intro text (EN) | |
| Instruction | |
| Instruction (EN) | |
| Question Text | Es gibt unzählige Tier- und Pflanzenarten auf der Erde, die in verschiedenen Ökosystemen leben. Wenn Arten aussterben, führt dies zu einem Verlust der Artenvielfalt (Biodiversität). Wenn sich neue Arten bilden oder in einem Gebiet neu ansiedeln, führt dies zu einem Anstieg der Artenvielfalt. Was denken Sie, hat die Artenvielfalt (Biodiversität) in den letzten 20 Jahren zu- oder abgenommen? Bitte antworten Sie auf einer Skala von 1 bis 7. |
| Question Text (EN) | There are countless animal and plant species on earth that live in different ecosystems. When species become extinct, this leads to a loss of biodiversity. When new species form or settle in an area, this leads to an increase in biodiversity. What do you think, has biodiversity increased or decreased over the last 20 years? Please answer on a scale from 1 to 7. |
| Item Text | In dem Kanton, in dem ich lebe |
| Item Text (EN) | In the canton where I live |

**Value Labels**

| | |
|---|---|
| 1 | Stark abgenommen |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | Stark zugenommen |

**Value Labels (EN)**

| | |
|---|---|
| 1 (EN) | Decreased greatly |
| 2 (EN) | 2 |
| 3 (EN) | 3 |
| 4 (EN) | 4 |
| 5 (EN) | 5 |
| 6 (EN) | 6 |
| 7 (EN) | Increased greatly |

| | |
|---|---|
| *Missing Labels* | |
| -22 | not in panel |
| -33 | unit nonresponse |
| -44 | missing by m.o.p. |
| -55 | missing by technical error |
| -66 | missing by design |
| -77 | not reached |
| -88 | missing by filter |
| -97 | nonvalid answer in survey (e.g. ambigious) |
| -99 | item nonresponse |

This comment is inserted below the chunk. One advantage is that you can add vertical space in comments outside of the chunk using \vspace{}. Below I insert a vertical space of 18pt.

After the space, I can write a second comment. You need to add \newpage after all comments to insert a page break.

# Examples of codebook summary functions

The coder must take the decision on which summary table or plot fits best the variable type. The examples for the summary tables and summary plots use the same variable respectively. This serves only for illustrating the design of the plot and use of the function. Particularly for the last plot (density plot), you can see that the plotting does not well fit the discrete type of the variable.

## Summary Tables

```r
# This function summarises variable information in a table.
# Besides the metadata, you have to provide the response data.
# Specifying the type argument summarises the information in different ways.
# For type "factor", the function creates a count table.
cb_sumtab(metadata = upanelmeta, response = upanel, num.var = 73, type = "factor")
```

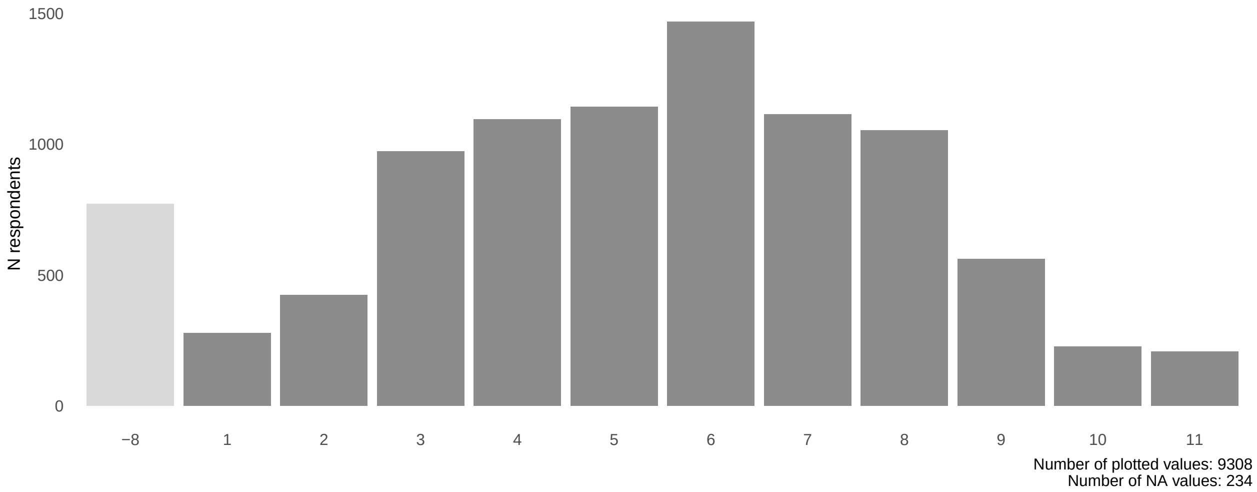
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | NA’s |
|---|---|---|---|---|---|---|---|
| 1279 | 1641 | 3148 | 2332 | 698 | 124 | 68 | 252 |

```
# This function summarises variable information in a table.
# Besides the metadata, you have to provide the response data.
# Specifying the type argument summarises the information in different ways.
# For type "numeric", the function provides distribution statistics.
cb_sumtab(metadata = upanelmeta, response = upanel, num.var = 73, type = "numeric")
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---:|---:|---:|---:|---:|---:|---:|
| 1 | 2 | 3 | 3.018622 | 4 | 7 | 252 |

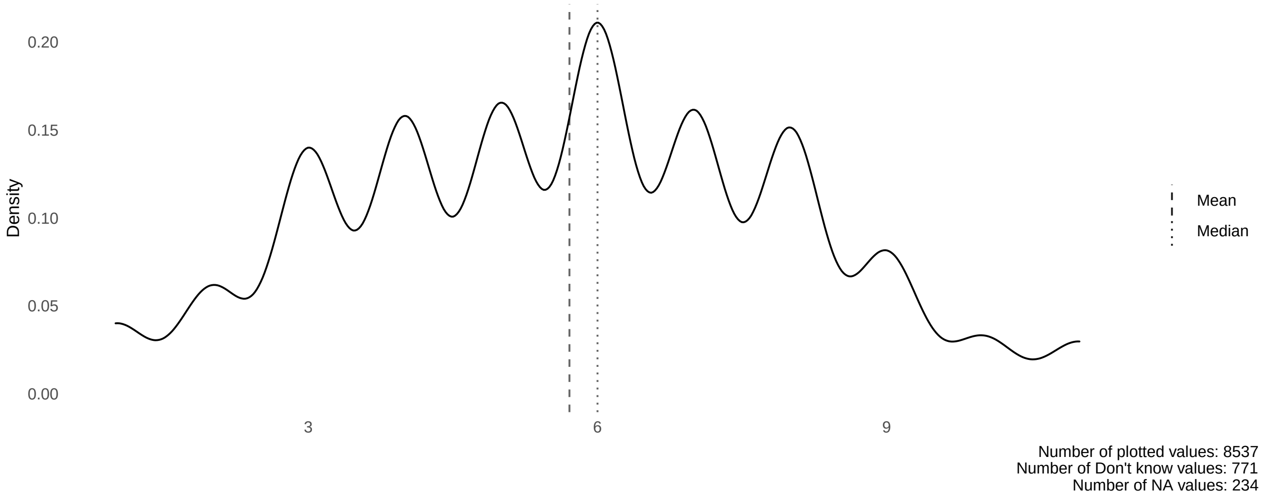
## Summary plots

```
# This function summarises variable information in a plot.
# Besides the metadata, you have to provide the response data.
# Specifying the type argument summarises the information in different ways.
# For stats "count", the function provides a vertical bar plot with value frequency.
cb_sumplot(metadata = upanelmeta, response = upanel, num.var = 97, stats = "count")
```



The example serves for illustrating the design of the plot and use of the function. The plot itself shows that density plots are not well suited for variables with discrete scales.

```r
# This function summarises variable information in a plot.
# Besides the metadata, you have to provide the response data.
# Specifying the type argument summarises the information in different ways.
# For type "density", the function provides a density plot.
cb_sumplot(metadata = upanelmeta, response = upanel, num.var = 97, stats = "density")
```



## The SEP missing code scheme

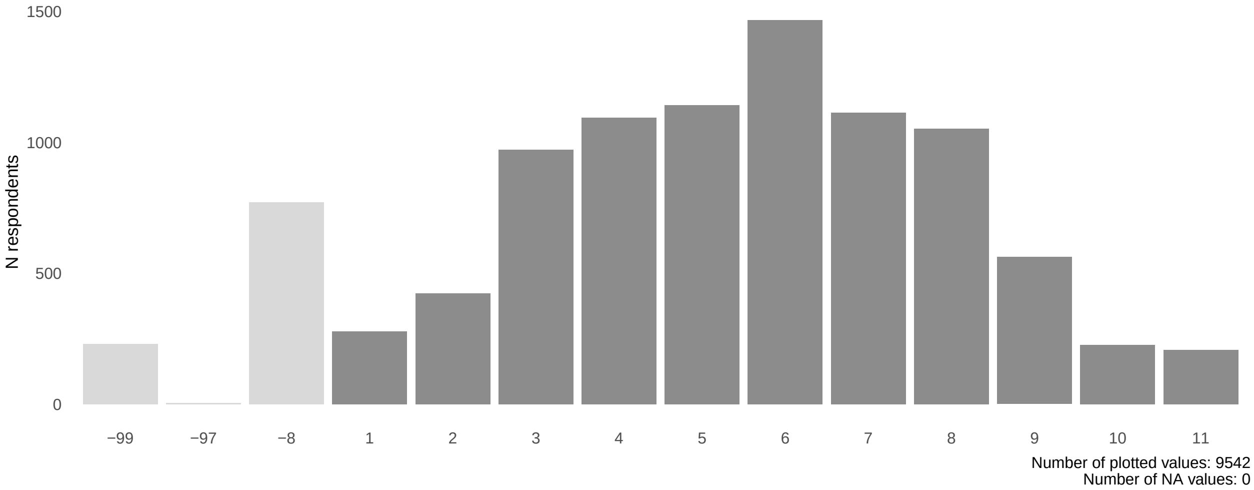
The SEP data has a specific missing code scheme. Missing values are coded as two-digit negative numbers (see the missing code table in this document) and not as NA. There are also special negative values for *Don't know* (-8) and *None* (-9) responses. The summary statistics function respect these special coding rules through the na_sep argument. It defaults to TRUE. If you do no want to use the SEP coding, you can set the argument to FALSE. The functions then consider only NA values as missing.

The example below uses SEP data, but setting na_sep to FALSE. You can see how the summary statistics get distorted.

```
cb_sumtab(metadata = upanelmeta, response = upanel, num.var = 97, type = "factor",
          na_sep = FALSE)
```

| -99 | -97 | -8 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | NA's |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 230 | 4 | 771 | 278 | 423 | 972 | 1094 | 1142 | 1467 | 1114 | 1052 | 562 | 226 | 207 | 0 |

```
cb_sumplot(metadata = upanelmeta, response = upanel, num.var = 97, stats = "count",
           na_sep = FALSE)
```


N respondents
Number of plotted values: 9542
Number of NA values: 0