Problems with Synthpop

Jonathan P. Latner

2024-07-18

```
 # Set the global seed here
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy = TRUE, tidy.opts = list(width.cutoff = 80))
options(max.print = 80) # Adjust the limit as needed
set.seed(123)

# load libraries
library(tidyverse)
library(synthpop)
```

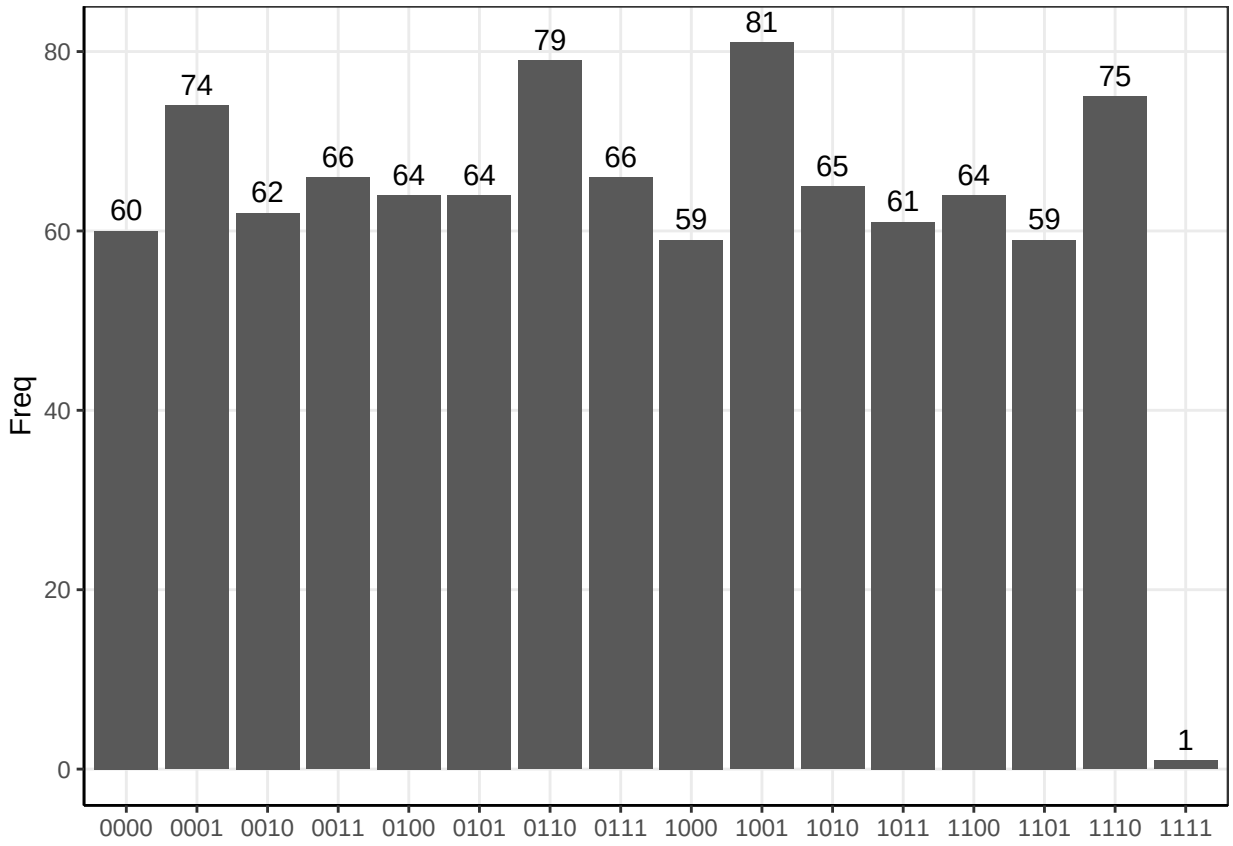
Generate simulated data

Following Reiter et al., 2014 (pg. 20):

"We use a simple simulation scenario that illustrates many of the main issues: pro- tecting a 24 binary table with fully synthetic data. For $i = 1, \ldots, 1000 = n$, let $y_i = (y_{1i}, y_{2i}, y_{3i}, y_{4i})$ comprise four binary variables. Let each of the K = 16 possible combinations be denoted ck, where $k = 1, \ldots, 16$. Let $c_{16} = (0, 0, 0, 0)$, and let $C_{-16} = (c_1, \ldots, c_{15})$. We generate an observed dataset D as follows. For $i = 1, \ldots, n_1 = 999$, sample yi from a multinomial distribution such that $p(y_i = c_k) = 1/15$ for all $c_k \in C_{-16}$. Set $y_{1000} = c_{16}$. Since we do full synthesis, $X = \emptyset$.' '

"With this design, we create a record that is guaranteed to be unique in the sample. Intuitively, we expect such records potentially to face higher risks, since they can offer information to the synthesis model that is not available from other records. Whether or not this is true depends on the nature of the synthesizer; to illustrate this, we examine results for different types of synthesizers, which we now describe ...' '

```
##    combination Freq
## 1         0000   60
## 2         0001   74
## 3         0010   62
## 4         0011   66
## 5         0100   64
## 6         0101   64
## 7         0110   79
## 8         0111   66
## 9         1000   59
## 10        1001   81
## 11        1010   65
## 12        1011   61
## 13        1100   64
## 14        1101   59
## 15        1110   75
## 16        1111    1
```

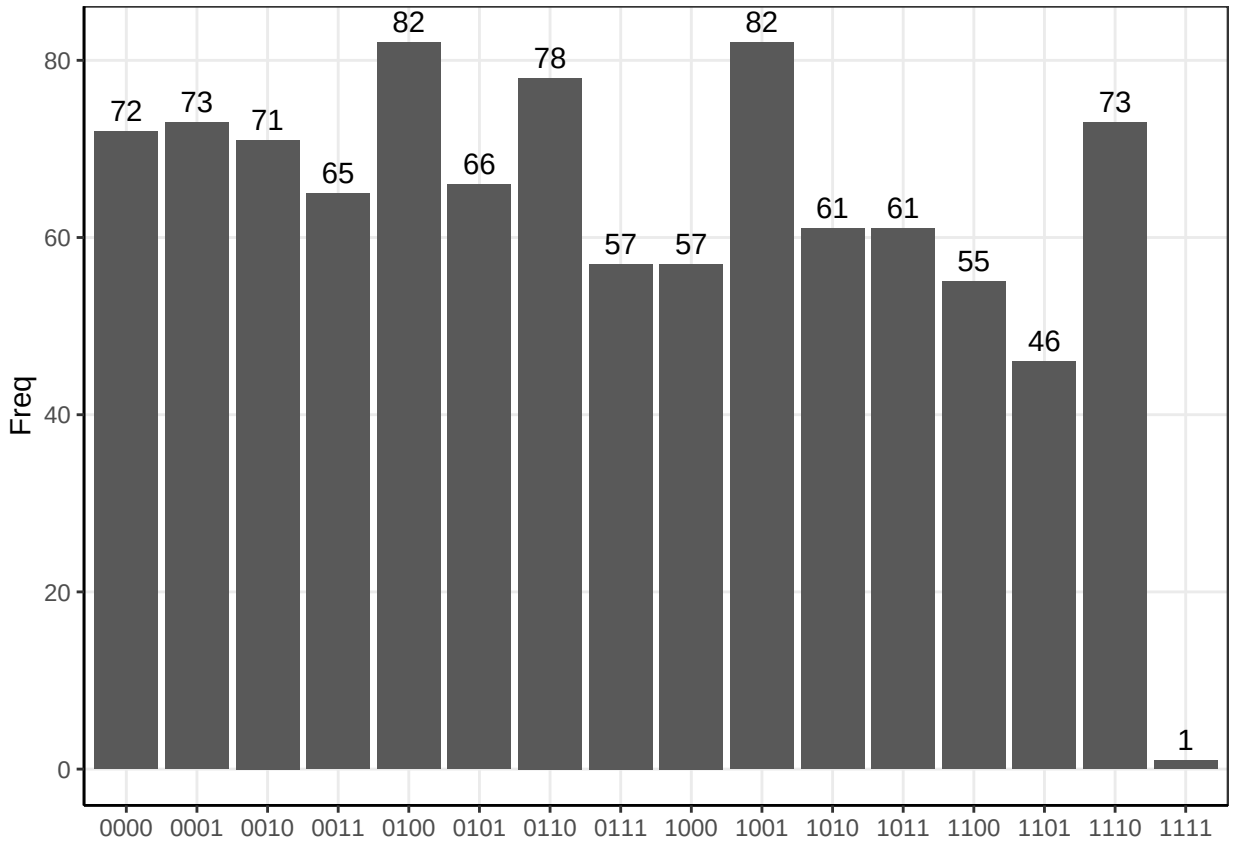


Generate synthetic data (default values)

First step is to estimate synthetic dat from Synthpop. Here, we use default values. In this case (with this seed), the synthesizer estimates values for all combinations. Notice that Synthpop creates 2 observations with the same combination of values as the unique value in the original data. This indicates that Synthpop provides low levels of protection for unique values in the data. However, also notice the warning suggesting that we change our numeric variables into factors because they have 5 or fewer levels.

Note: we refer to graphs with similar frequency as 'graph a: high risk'

```
## Warning: In your synthesis there are numeric variables with 5 or fewer levels: var1, var2, var3, var
## Consider changing them to factors. You can do it using parameter 'minnumlevels'.
##
## Synthesis
## -----------
##  var1 var2 var3 var4
```



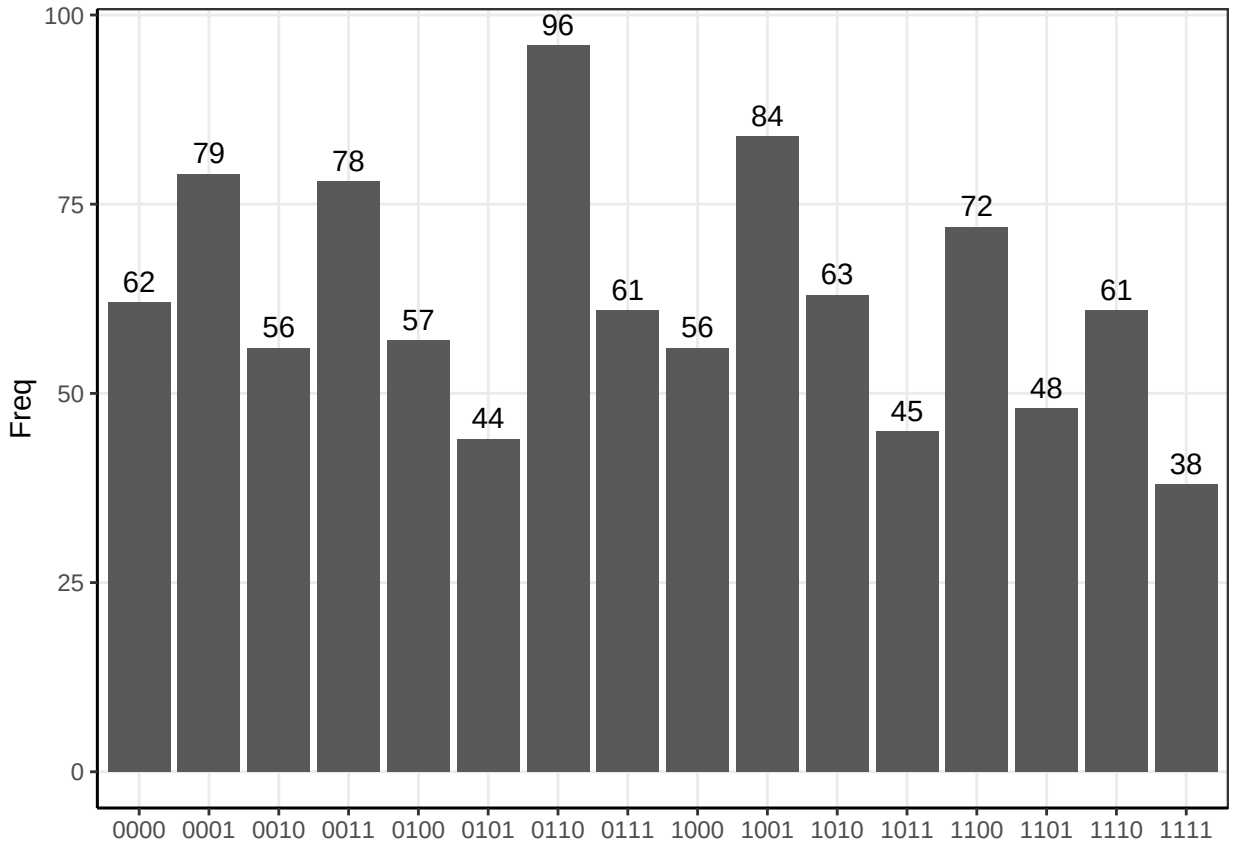
Generate synthetic data with minimum values

We follow the suggestion of Synthpop and use the paramter minnumlevels to indicate that numeric variables with less than 5 values should be treated as factor variables (default is 1). In this case, Synthpop creates 29 observations with the same combination as the unique value in the original data.

On the one hand, this would seem to be a good thing because it indicates a higher level of protection. On the other hand, this also indicates that levels of protection depends on whether original data are treated as numeric or character.

Note: we refer to graphs with similar frequency as 'graph b: low risk'

```
##
## Variable(s): var1, var2, var3, var4 numeric but with only 5 or fewer distinct values turned into fac
##
##
## Synthesis
## -----------
##  var1 var2 var3 var4
```

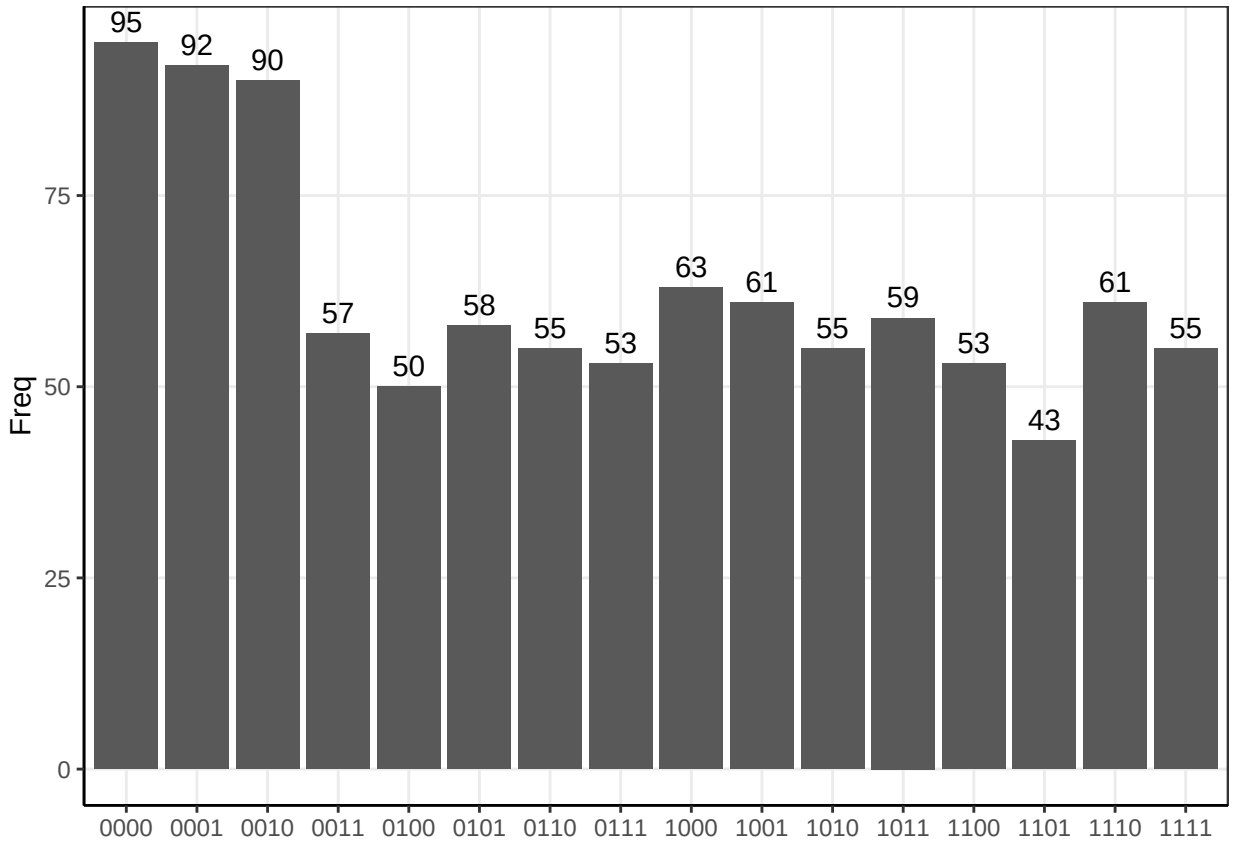


Generate synthetic data with cp = 0.1 (default is 1e-8)

To test this, we estimate new synthetic data, but without minnumlevels. Instead, we alter the complexity parameter to restrict the depth of the tree. In so doing, we can replicate graph a (high risk), but using numeric values.

The point: there is a difference between the depth of trees for numeric and factor variables created by synthpop.

```
## Warning: In your synthesis there are numeric variables with 5 or fewer levels: var1, var2, var3, var4
## Consider changing them to factors. You can do it using parameter 'minnumlevels'.
##
## Synthesis
## -----------
##  var1 var2 var3 var4
```



Initial conclusions

Synthpop appears to synthesize factor variables differently than numeric variables. It seems that for numeric variables, it does not build trees that are as deep as the trees it builds for factor variables. We do not think this is a feature of Synthpop, but rather a bug as it should build trees of similar depth regardless of the variable type.

If we assume this is a feature, then Synthpop provides lower levels of protection for unique combinations of numeric variables relative to unique combinations of factor variables. It seems unusual to think that this would be intentional.

If we assume this is a bug, then (conditional on fixing it) there is a broader problem. In our simulated data with limited number of variable combinations and 1 unique value, then Synthpop provides almost no protection for the unique value.