Package 'sampleVADIR'

October 14, 2022

Title Draw Stratified Samples from the VADIR Database

Version 1.0.0

Maintainer Trevor Swanson <trevorswanson222@gmail.com>

Description Affords researchers the ability to draw stratified samples from the U.S. Department of Veteran's Affairs/Department of Defense Identity Repository (VADIR) database according to a variety of population characteristics. The VADIR database contains information for all veterans who were separated from the military after 1980. The central utility of the present package is to integrate data cleaning and formatting for the VADIR database with the stratification methods described by Mahto (2019) <https://CRAN.R-project.org/package=splitstackshape>. Data from VADIR are not provided as part of this package.

License GPL (>= 3)

URL https://github.com/tswanson222/sampleVADIR

BugReports https://github.com/tswanson222/sampleVADIR/issues

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Imports lubridate, methods, splitstackshape

Suggests haven, rio

Depends R (>= 2.10)

NeedsCompilation no

Author Trevor Swanson [aut, cre],
Kelsie Forbush [aut],
Joanna Wiese [ctb],
Melinda Gaddy [ctb],
Mary Oehlert [ctb]

Repository CRAN

Date/Publication 2021-10-27 14:40:04 UTC

R topics documented:

`checkData` *Compare old and new versions of VADIR to find repeats*

Description

Can be used to identify whether a new version of VADIR contains any old responses. Can also automatically remove repeated responses.

Usage

```
checkData(old, new, fix = FALSE, dates = FALSE)
```

Arguments

`old`	Past version of VADIR
`new`	New version of VADIR
`fix`	Logical. Determines whether to automatically remove repeated responses.
`dates`	Logical. Determines whether to include date variables when comparing datasets. Recommended to keep `FALSE`.

Value

Returns a message that no repeated responses exist if there are none. Otherwise, returns either a warning that repeated responses exist, or returns the new VADIR dataset without repeated responses if `fix = TRUE`.

`fixTypos` *Fix typos in VADIR dataset*

Description

If there are known typos, the correct values of those incorrect responses can be provided and fixed across the dataset.

Usage

```
fixTypos(data, old, new = NULL, var = "RANK_CD")
```

Arguments

`data`	VADIR dataset
`old`	Character vector containing typos
`new`	Character vector in the same order as `old`, containing corresponding values to fix typos to.
`var`	Variable name for which typos should be corrected

Value

VADIR dataset with typos corrected

Examples

```
data <- fixTypos(data = VADIR_fake, old = c('CW02', 'CW0-2', 'PV1'),
                 new = c('CWO2', 'CWO2', 'PVT'), var = 'RANK_CD')
```

`getData` *Data import function to accommodate multiple filetypes*

Description

Allows for easy data importation. Automatically detects filetype and applies appropriate function for importing.

Usage

```
getData(filename, filetype = "csv", fixDates = FALSE, ...)
```

Arguments

filename	Character string specifying the path to desired datafile
filetype	Character string indicating filetype. Useful if no file extension is provided in filename. If file extension is provided in filename (recommended), then this argument is ignored. Accommodates "csv", "rds", "xlsx", "rdata", "sav", "txt"
fixDates	Logical. Determines whether to adjust date format.
...	Additional arguments

Value

Imported datafile

---

rankDat *Rank to pay grade data*

Description

Serves as a key for relating certain military rank designations with pay grades. Used in the sampleVADIR function for stratifying based on pay grade rather than rank.

Usage

```
rankDat
```

Format

A data frame with six variables that links pay grades to military ranks within each military branch. PayGrade indicates the pay grade associated with a specific job title (Title) within a given Branch of the military. Title designates the job title, where Initials is the shorthand for each title (this is how the RANK_CD variable is coded in the VADIR dataset). Branch designates the military branch, where "N" stands for Navy, "A" stands for Army, "M" stands for Marines, and "F" stands for Air Force. PayCat4 represents one coding scheme that categorizes different pay grades into four categories, where "E" stands for enlisted, "NCO" stands for non-commissioned officer, "W" stands for warrant officer, and "O" stands for commissioned officer. PayCat7 represents an alternative categorization that breaks pay grades into seven categories, wherein "SNCO" stands for senior non-commissioned officer, "FGO" stands for field grade officer, "CGO" stands for company grade officer, and "GO" stands for general officer.

Details

The way these data are used in the sampleVADIR function is by indexing the values of the RANK_CD variable of the VADIR dataset against the Initials variable in the present dataset, and then the RANK_CD value is replaced with the associated value in either the PayCat4 or PayCat7 variable depending on what is specified in the sampleVADIR function. The purpose of this is to make the RANK_CD variable more amenable to stratification, given the difficultly of stratifying across values of a categorical variable with so many unique values.

`sampleVADIR` *Draw stratified samples from VADIR database*

Description

Core function used to pull a stratified sample from VADIR based on a variety of parameters.

Usage

```
sampleVADIR(
  data,
  n = 4500,
  vars = "all",
  rankDat = "rankDat",
  payRanks = 4,
  post911 = TRUE,
  dischargedAfter = FALSE,
  until = NULL,
  ageDischarge = TRUE,
  ageEnlist = FALSE,
  ageNow = FALSE,
  yearsServed = FALSE,
  dateformat = "%m/%d/%Y",
  params = NULL,
  formats = "default",
  typos = list(),
  rmDeviates = FALSE,
  timeCats = FALSE,
  saveData = TRUE,
  onlyIDs = FALSE,
  oversample = FALSE,
  exclude = FALSE,
  seed = NULL
)
```

Arguments

`data`	VADIR dataset
`n`	Total desired sample size
`vars`	Character vector indicating which variables to use in stratification
`rankDat`	Dataset linking ranks to pay grade, or character string indicating where to pull that dataset from. Recommended to leave as `"rankDat"` in order to use package-supplied dataset.
`payRanks`	Number of pay grades to use when converting rank variable. Only options are either 4 or 7.

post911	Logical. Determines whether to only consider individuals deployed after 9/11/2001
dischargedAfter	Character string indicating what date to restrict sampling to based on discharge date. Can set to FALSE if this is to be ignored. Can also set to 'past-year' in order to only sample people who were discharged within the past year (given the current date).
until	Upper limit to when service was started. NULL means there is no upper limit
ageDischarge	Logical. Determines whether to use age at discharge as a stratum.
ageEnlist	Logical. Determines whether to use age at enlist as a stratum.
ageNow	Logical. Determines whether to use current age as a stratum.
yearsServed	Logical. Determines whether to use total years served as a stratum.
dateformat	Character string indicating the expected date format. Should be automatically detected.
params	Optional list of parameters to override defaults in function. Creates an easy way to interface with the function if performing the stratification multiple times. Allows the user to avoid writing the same arguments multiple times.
formats	Should be "default"
typos	List containing typos to be fixed, as well as what they should be changed to. Leave at list() to ignore. Typos can also be fixed prior to stratification by using the fixTypos function.
rmDeviates	Logical. Determines whether rows with unexpected response values are removed. If FALSE, and deviate response values are detected, the function will stop.
timeCats	Logical or numeric. Determines whether the time-related variables should be treated as categorical variables. If TRUE, this defaults to 4.
saveData	Logical. Determines whether to save the full dataset in the output. Specifically, returns the full dataset of candidates (i.e., some people may be removed from consideration due to errors or unexpected responses).
onlyIDs	Logical. Determines whether to only return ID values for selected individuals rather than a full dataset.
oversample	Logical. Determines whether to oversample or undersample based on limitations due to available proportions of strata in subsample.
exclude	Logical. Determines whether to exclude people missing a zip code, as well as people with "NTC" as their zip code value.
seed	Numeric value indicating the seed to set for the stratification procedure. Allows for reproducible results.

Details

Performs stratification separately for males and females, where males and females are sampled at a 1:1 ratio, regardless of population ratio.

With a large dataset (which is typical for VADIR), setting any of the date-related variables to TRUE can drastically increase computation time. The relevant arguments include: ageDischarge, ageEnlist, ageNow,yearsServed.

Value

A list containing the males and females who were sampled from VADIR

Examples

```
params <- list(
  n = 7000,
  vars = c('PN_Sex_CD', 'PN_BRTH_DT', 'SVC_CD', 'PNL_CAT_CD', 'RANK_CD',
           'PNL_TERM_DT', 'PNL_BGN_DT', 'OMB_RACE_CD',
           'OMB_ETHNC_NAT_ORIG_CD', 'POST_911_DPLY_IND_CD'),
  rankDat = 'rankDat',
  payRanks = 4,
  post911 = FALSE,
  until = NULL,
  dischargedAfter = FALSE,
  ageDischarge = TRUE,
  ageEnlist = FALSE,
  ageNow = FALSE,
  yearsServed = FALSE,
  dateformat = '%m/%d/%Y',
  formats = 'default',
  rmDeviates = FALSE,
  timeCats = TRUE,
  saveData = TRUE,
  onlyIDs = FALSE,
  oversample = TRUE,
  exclude = FALSE,
  typos = list()
)

out <- sampleVADIR(VADIR_fake, params = params, seed = 19)
```

---

testStrata *Return correlations for demographics between population and sample*

Description

Used to evaluate the representativeness of the sample with regard to the population. Males and females evaluated separately.

Usage

```
testStrata(out, data = NULL, metric = cor, zeros = FALSE)
```

Arguments

out	Output of sampleVADIR
data	Original VADIR data
metric	Function for measuring similarity between population and sample
zeros	Should empty strata be included?

Value

Similarity values for males and females

VADIR_fake *Fake VADIR data*

Description

Simulated VADIR data based solely on the variable names and appropriate response options for each. Values of variables were generated based on population proportions identified in a subsample of approximately 140,000 veterans from a version of the VADIR database obtained in 2020. However, this simulated dataset does not fully represent population characteristics of VADIR, and is simply meant as a faux tool for testing functions in the sampleVADIR package.

Usage

```
VADIR_fake
```

Format

A data frame with ten variables, representing variables as they are formatted within the actual VADIR database.

Index