Package: syntrial (via r-universe)

September 1, 2024

Title synthesize data from real clinical trial

Version 0.4.2.9001

Description Syntrial takes as input the data of a clinical trial in SDTM format (https://www.cdisc.org/standards/foundational/sdtm). For each synthesized patient a number of source patients is selected. From the source patients' data randomly weighted numeric values or, for categorical variables, one value is selected in a modification of classical bootstrap.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.5)

Imports dplyr (>= 1.0), lubridate, magrittr, philentropy, tibble, tidyr

Suggests testthat (>= 2.1.0)

RoxygenNote 7.2.3

Repository https://pharmaverse.r-universe.dev

RemoteUrl https://github.com/openpharma/syntrial

RemoteRef HEAD

RemoteSha 52bd03b2f66190260e877ecde3073a2aa25cfd87

Contents

`categorical_vars` *categorical variables*

Description

Determine all categorical variables of a dataframe and return a vector of their names.

Usage

```
categorical_vars(df)
```

Arguments

`df`	dataframe to be evaluated

Value

logical

Examples

```
categorical_vars(CRC305ABC_DM)
```

`CRC305ABC_AE` *adverse event data from clinical trials CRC305ABC*

Description

An anonymized dataset from 123 patients in SDTM format (https://en.wikipedia.org/wiki/SDTM).

Usage

```
CRC305ABC_AE
```

Format

A dataframe with 552 rows and 32 variables:

STUDYID Study Identifier

DOMAIN Domain Abbreviation

USUBJID Unique Subject Identifier

AESEQ Sequence Number

AESPID Sponsor-Defined Identifier

AETERM Reported Term for the Adverse Event

AEMODIFY Modified Reported Term

AELLT Lowest Level Term

AELLTCD Lowest Level Term Code

AEDECOD Dictionary-Derived Term

AEPTCD Preferred Term Code

AEHLT High Level Term

AEHLTCD High Level Term Code

AEHLGT High Level Group Term

AEHLGTCD High Level Group Term Code

AECAT Category for Adverse Event

AEPRESP Pre-Specified Adverse Event

AESOC Primary System Organ Class

AESOCCD Primary System Organ Class Code

AELOC Location of Event

AESEV Severity/Intensity

AESER Serious Event

AEACN Action Taken with Study Treatment

AEREL Causality

AEPATT Pattern of Adverse Event

AEOUT Outcome of Adverse Event

AESTDTC Start Date/Time of Adverse Event

AEENDTC End Date/Time of Adverse Event

AESTDY Study Day of Start of Adverse Event

AEENDY Study Day of End of Adverse Event

AEDUR Duration of Adverse Event

AESIZE Measure of Adverse Event

Source

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QPHMKX

```
CRC305ABC_AE <- readr::read_tsv('https://dataverse.harvard.edu/api/access/datafile/3462713?gbrecs=fa
                                 col_types = 'ccciccccicicicicccccccccccciici',
                                 locale = readr::locale(encoding = 'latin1')
                                )
```

CRC305ABC_DM *demographic data from clinical trials CRC305ABC*

Description

An anonymized dataset of 123 patients in SDTM format (https://en.wikipedia.org/wiki/SDTM).

Usage

```
CRC305ABC_DM
```

Format

A dataframe with 123 rows and 19 variables:

STUDYID Study Identifier

DOMAIN Domain Abbreviation

USUBJID Unique Subject Identifier

SUBJID Subject Identifier for the Study

RFSTDTC Subject Reference Start Date/Time

RFENDTC Subject Reference End Date/Time

RFXSTDTC Date/Time of First Study Treatment

RFXENDTC Date/Time of Last Study Treatment

RFICTC Date/Time of Informed Consent

RFPENDTC Date/Time of End of Participation

SITEID Study Site Identifier

AGE Age

AGEU Age Units

SEX Sex

RACE Race

ETHNIC Ethnicity

ARMCD Planned Arm Code

ARM Description of Planned Arm

COUNTRY Country

Source

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QPHMKX

```
CRC305ABC_DM <- readr::read_tsv('https://dataverse.harvard.edu/api/access/datafile/3462712?gbrecs=fa
                                  col_types = 'ccccccccccicccccc',
                                  locale = readr::locale(encoding = 'latin1')
                                )
```

CRC305ABC_LB *lab data from clinical trials CRC305ABC*

Description

An anonymized dataset from 123 patients in SDTM format (https://en.wikipedia.org/wiki/SDTM).

Usage

```
CRC305ABC_LB
```

Format

A dataframe with 55660 rows and 31 variables:

STUDYID Study Identifier

DOMAIN Domain Abbreviation

USUBJID Unique Subject Identifier

LBSEQ Sequence Number

LBREFID Specimen ID

LBTESTCD Lab Test or Examination Short Name

LBTEST Lab Test or Examination Name

LBCAT Category for Lab Test

LBORRES Result or Finding in Original Units

LBORRESU Original Units

LBORNRLO Reference Range Lower Limit in Orig Unit

LBORNRHI Reference Range Upper Limit in Orig Unit

LBSTRESC Character Result/Finding in Std Format

LBSTRESN Numeric Result/Finding in Standard Units

LBSTRESU Standard Units

LBSTNRLO Reference Range Lower Limit-Std Units

LBSTNRHI Reference Range Upper Limit-Std Units

LBNRIND Reference Range Indicator

LBSTAT Completion Status

LBREASND Reason Test Not Done

LBNAM Vendor Name

LBSPEC Speciment Type

LBSPCCND Specimen Condition

LBBLFL Baseline Flag

LBFAST Fasting Status

VISITNUM Visit Number

VISIT Visit Name

LBDTC Date/Time of Specimen Collection

LBDY Study Day of Specimen Collection

LBTPT Planned Time Point Name

LBTPTNUM Planned Time Point Number

Source

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QPHMKX

```
CRC305ABC_LB <- readr::read_tsv('https://dataverse.harvard.edu/api/access/datafile/3462715?gbrecs=fa
                                col_types = 'ccciccccccccdcddccccccciccicci',
                                locale = readr::locale(encoding = 'latin1')
                                )
```

`CRC305ABC_VS` *vital sign data from clinical trials CRC305ABC*

Description

An anonymized dataset from 123 patients in SDTM format (https://en.wikipedia.org/wiki/SDTM).

Usage

```
CRC305ABC_VS
```

Format

A dataframe with 20581 rows and 25 variables:

STUDYID Study Identifier

DOMAIN Domain Abbreviation

USUBJID Unique Subject Identifier

VSSEQ Sequence Number

VSTESTCD Vital Signs Test Short Name

VSTEST Vital Signs Test Name

VSPOS Vital Signs Position of Subject

VSORRES Result or Finding in Original Units

VSORRESU Original Units

VSSTRESC Character Result/Finding in Std Format

VSSTRESN Numeric Result/Finding in Standard Units

VSSTRESU Standard Units

VSSTAT Completion Status

VSREASND Reason Not Performed

VSLOC Location of Vital Signs Measurement

VSBLFL Baseline Flag

VISITNUM Visit Number

VISIT Visit Name

VSDTC Date/Time of Measurements

VSDY Study Day of Vital Signs

VSTPT Planned Time Point Name

VSTPTNUM Planned Time Point Number

VSORNRLO Reference Range Lower Limit

VSORNRHI Reference Range Upper Limit

VSNRIND Reference Range Indicator

Source

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QPHMKX

```
CRC305ABC_VS <- readr::read_tsv('https://dataverse.harvard.edu/api/access/datafile/3462714?gbrecs=fa
                                col_types = 'ccciccccccdccccciccicicccc',
                                locale = readr::locale(encoding = 'latin1')
                                )
```

domain_ids *domain identifier variables*

Description

Return identifier variables of a domain.

Usage

```
domain_ids(df)
```

Arguments

df	dataframe of one SDTM domain

Value

character vector of identifier variable names

Examples

```
domain_ids(CRC305ABC_DM)
```

Hmax *maximum entropy*

Description

Compute maximum shannon entropy `log(N, base=2)` of a number or of a dataframe or matrix. In the latter case the number of rows is used.

Usage

```
Hmax(N)
```

Arguments

N	a number, a dataframe, or a matrix

Value

log(N, base=2)

Examples

```
Hmax(CRC305ABC_DM)
```

Hstruc *Shannon entropy structure details of a dataframe*

Description

Hstruc computes the Shannon entropy of each dataframe's variable and checks for

- constant,
- record identifying, or
- 1:1 equivalent

variables.

Usage

```
Hstruc(.data)
```

Arguments

.data	dataframe

Value

A list

Examples

```
Hstruc(CRC305ABC_DM)
```

synthesize *synthesize a SDTM dataframe*

Description

Create a synthetic dataframe from syndf and a source SDTM dataframe.

Usage

```
synthesize(syndf, df, cat_fuzz = 1)
```

Arguments

syndf	synthesis dataframe
df	original SDTM dataframe
cat_fuzz	fuzz factor for noise on categorical variables, defaults to 1

Value

synthesized SDTM dataframe

Examples

```
synthesize(synth_df(CRC305ABC_DM$USUBJID), CRC305ABC_DM)
```

---

synthesize_E *synthesize events dataframe*

Description

synthesize events dataframe

Usage

```
synthesize_E(syndf, df, cat_fuzz = 1)
```

Arguments

syndf	synthesis dataframe
df	original SDTM dataframe
cat_fuzz	fuzz factor for noise on categorical variables, defaults to 1

Value

synthesized SDTM events dataframe

Examples

```
## Not run: synthesize_E(synth_df(CRC305ABC_DM$USUBJID), CRC305ABC_AE)
```

---

synth_df *generate synthesis dataframe*

Description

`synth_df()` generates the tibble that relates original and synthetic new persons. This synthesis dataframe should be kept secret to protect privacy of the original persons; it is the base for all synthetic dataframes/tibbles that are generated.

Usage

```
synth_df(USUBJID, n_new = 10, width = 3, maxweight = 2/3)
```

Arguments

USUBJID	character vector of person identifiers
n_new	number of new persons to generate, defaults to 10
width	number of persons from original trial to use for new person synthesis, defaults to 3
maxweight	maximum allowed weight for one person for synthesis of new person, defaults to 2/3

Value

The synthesis tibble relates new synthetic persons to source persons and specifies the weight of each source person's contribution.

Examples

```
synth_df(USUBJID = letters)
```

---

syntrial *Synthetic "twin trial" from real clinical trial data in SDTM format.*

Description

Stricter privacy regulations affect not only sharing of clinical trial data but also development of analysis scripts and reports. Synthetic data with same statistical properties as the original data can be used instead, possibly even for data exploration.

Details

syntrial synthesizes data in SDTM format and protects privacy via an intermediate mechanism between frequentist and Bayesian bootstrap. For each synthesized patient a number of source patients is selected. From the source patients' data randomly weighted numeric values or, for categorical variables, one value is selected.

synthdf

create the table linking real and synthetic persons

synthesize

synthesize demographic and findings data

synthesize_E

synthesize events data

Index