Package 'CBCgrps'

April 15, 2021

Type Package

Title Compare Baseline Characteristics Between Groups

Version 2.8.2

Date 2021-4-16

Author Zhongheng Zhang,
Sir Run Run Shaw hospital,
Zhejiang university school of medicine

Maintainer Zhongheng Zhang <zh_zhang1984@zju.edu.cn>

Depends R (>= 3.2.0), nortest (>= 1.0-4)

Description Compare baseline characteristics between two or more groups. The variables being compared can be factor and numeric variables. The function will automatically judge the type and distribution of the variables, and make statistical description and bivariate analysis.

License GPL-2

RoxygenNote 7.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2021-04-15 06:00:04 UTC

R topics documented:

CBCgrps2.8-package *Compare Baseline Characteristics Between Groups*

Description

The package aims to automate the process of comparing Baseline Characteristics between groups.

Details

The DESCRIPTION file: In clinical studies emplying electronic medical records, the variables to be investigated are usually large in number. It is sometimes cumbersome to compare these variables between two or more groups one by one. I design this package to automate the process of statistical description and bivariate statistical inference.

Author(s)

Zhongheng Zhang Department of emergency medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, China. <zh_zhang1984@zju.edu.cn> Maintainer: Zhongheng Zhang

References

Zhang Z. Univariate description and bivariate statistical inference: the first step delving into data.Ann Transl Med. 2016 Mar;4(5):91.

Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernasndez P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. Ann Transl Med. 2017 Dec;5(24):484. doi: 10.21037/atm.2017.09.39.

See Also

No ther reference.

Examples

```
data(df)
a<-twogrps(df,"mort")
```

df *simulated dataset as a working example*

Description

A data frame with 1000 observations on the following 7 variables.

Usage

```
data("df")
```

Format

A data frame with 1000 observations on the following 7 variables.

`crp` a numeric vector, C-reactive protein measured in mg/l

`hb` a numeric vector, hemoglobin

`ddimer` a numeric vector

`wbc` a numeric vector, white blood cell

`comorbid` a factor with levels `cirrhosis` `COPD` `diabetes` `heartfailure` `hypertension` `renalfailure` `stroke`

`sex` a factor with levels `female` `male`

`mort` a factor with levels `alive` `dead`

Details

The dataset is generated as a working example without clinical relevance.

Source

Simulated dataset without sources.

References

simulated dataset without reference.

Examples

```
data(df)
## maybe str(df) ; plot(df) ...
```

dt *simulated dataset as a working example*

Description

A data frame with 500 observations on the following 8 variables. Missing values are also present

Usage

```
data("dt")
```

Format

A data frame with 500 observations on the following 8 variables.

`crp` a numeric vector, C-reactive protein measured in mg/l, with missing values

`vaso` a factor with two levels `Yes` `No`, indicating the use of vasopressor or not

`wbc` a numeric vector, white blood cell count

`lac` a numeric vector, serum lactate

`age` a numeric vector, age in years

`type` a factor with levels `surgery` `medical` `emergency`

`gender` a factor with levels `female` `male`

`mort` an integer with two values `1` `0`

Details

The dataset is generated as a working example without clinical relevance.

Source

Simulated dataset without sources.

References

simulated dataset without reference.

Examples

```
data(dt)
## maybe str(df) ; plot(df) ...
```

multigrps *Compare Baseline Characteristics between three or more groups*

Description

The main function of the CBCgrps package.

Usage

```
multigrps(df, gvar, p.rd = 3, varlist = NULL,
    skewvar=NULL, norm.rd = 2,
    sk.rd = 2, tabNA = "no", cat.rd = 0, pnormtest = 0.05,
    maxfactorlevels = 30,
    minfactorlevels = 10, sim = FALSE, workspace = 2e+05,
    ShowStatistic = F)
```

Arguments

df	The data frame on which statistical descroption and inference are performed.
gvar	The group variable.
p.rd	Decimal space of p value to be displayed. If the p value is less than the minimum value of that decimal space, it will print less than that value. For instance, if p.rd = 3 and p = 0.00045, then it will print "<0.001" in the p column of the table.
varlist	Specify a vector of variable names to be compared between groups (i.e. not all variables in the *df* will be compared and users can choose which variables to be compared). This argument also allows to specify the order in which the variables will appear in the table. It will issue an error message if varlist contains variable names not found in the df.
skewvar	Specify a vector of variable names which are considered to be not normally distributed. This function is useful for some large datasets where the normality test is too sensitive and users may want to specify skew variables by their own judgement. skewvar is NULL by default. If it is not NULL and skew data were specified by the users, the statistical test for normality is switched off.
norm.rd	Decimal space of normally distributed numeric variables to be displayed.
sk.rd	Decimal space of skewed numeric variables to be displayed.
tabNA	Weather categorical variables with NA be displayed or not. "no" to be omitted, "ifany" to be displayed. The default value is "no".
cat.rd	Decimal space of categorical variables (the proportion) to be displayed.
pnormtest	Significance level for the normal test. It is 0.05 by convention (default). However, for some large datasets the test will be too sensitive that only a small deviation in magnitude from the normal distribution will give a p value less than 0.05. In this situation, users can specify smaller significance level. Note that the normality test will no longer be used to judge the normality if the skewvar argument is not NULL.

`maxfactorlevels`
: The maximum levels for factor variables, the default is 30. The argument is used to avoid treating date or time variables as factor variables.

`minfactorlevels`
: If a numeric variable has only several values, it is treated as categorical variable. The default value is 10.

`sim`
: a logical indicating whether to compute p-values by Monte Carlo simulation, in larger than 2 by 2 tables. The default is FALSE.

`workspace`
: If the fisher.test() fucntion requires more workspace, it can be defined here. The default is workspace euqals to 2e+05.

`ShowStatistic`
: logic value for whether showing statistics or not. The defaule is FALSE for not showing statistics. Statistics is used for statistical inference such as F value for Chi-square test and T value for student t test. No statistic will be shown because Fisher's exact test jumps past a test statistic and goes straight to a p-value.

Details

The function compares differences in categorical and continuous variables between three or more groups. The function automatically judges the distribution of the continuous variable and use appropriate description for them. Chi-square test is used for categorical data. Analysis of variance is used for normally distributed numeric data. Kruskal-Wallis rank sum test is used for non-normally distributed data. It is common that some categorical variables contain numeric or integer values. For example, the gender variable may contain values 1 and 2, representing male and female respectively. Such a variable can be identified by counting the number of integer values. Thus, the minfactorlevels argument is used to define the minimum value for a variable to be considered as numeric variable. For some large dataset, the normality test is extremely sensitive that a small deviation from normal distribution will lead to the rejection of the null hypothesis of normality. In such a circumstance, users may opt to switch off the normality test by explicitly specify the skewed data (i.e. skewvar=some variables names) or set a smaller p value for normal test (i.e. pnormtest=0.0001).

Value

`Table`
: The table with string values. The mean and standard error are put in a single cell, and connected by plus and minus symbol.

Note

No further notes

Author(s)

Zhongheng Zhang Department of emergency medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, China. <zh_zhang1984@zju.edu.cn>

References

Myles Hollander and Douglas A. Wolfe (1973), Nonparametric Statistical Methods. New York: John Wiley&Sons. Pages 115-120.

Zhang Z. Univariate description and bivariate statistical inference: the first step delving into data.Ann Transl Med. 2016 Mar;4(5):91.

Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernasndez P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. Ann Transl Med. 2017 Dec;5(24):484. doi: 10.21037/atm.2017.09.39.

See Also

No other reference

Examples

```
data(df)
b<-multigrps(df,"comorbid")
print(b,quote=TRUE)
```

`twogrps` *Compare Baseline Characteristics between two groups*

Description

The main function of the CBCgrps package. The function compares differences in categorical and continuous variables between two groups. The function automatically judges the distribution of the continuous variable and use appropriate description for them. Wilcoxon rank sum test is employed for non-normal data.

Usage

```
twogrps(df, gvar, varlist = NULL, p.rd = 3,
    skewvar=NULL, norm.rd = 2,
    sk.rd = 2, tabNA = "no", cat.rd = 0, pnormtest = 0.05,
    maxfactorlevels = 30, minfactorlevels = 10, sim = FALSE,
    workspace = 2e+05,ShowStatistic = F,ExtractP = 0.05)
```

Arguments

`df`	The data frame on which statistical descroption and inference are performed.
`gvar`	The group variable.
`p.rd`	Decimal space of p value to be displayed. If the p value is less than the minimum value of that decimal space, it will print less than that value. For instance, if p.rd = 3 and p = 0.00045, then it will print "<0.001" in the p column of the table.
`varlist`	Specify a vector of variable names to be compared between groups (i.e. not all variables in the df will be compared and users can choose which variables to be compared). This argument also allows to specify the order in which the variables will appear in the table.

skewvar	Specify a vector of variable names which are considered to be not normally distributed. This function is useful for some large datasets where the normality test is too sensitive and users may want to specify skew variables by their own judgement. skewvar is NULL by default. If it is not NULL and skew data were specified by the users, the statistical test for normality is switched off.
norm.rd	Decimal space of normally distributed numeric variables to be displayed.
sk.rd	Decimal space of skewed numeric variables to be displayed.
tabNA	Weather categorical variables with NA be displayed or not. "no" to be omitted, "ifany" to be displayed. The default value is "no".
cat.rd	Decimal space of categorical variables (the proportion) to be displayed.
pnormtest	Significance level for the normal test. It is 0.05 by convention (default). However, for some large datasets the test will be too sensitive that only a small deviation in magnitude from the normal distribution will give a p value less than 0.05. In this situation, users can specify smaller significance level. Note that the normality test will be switched off by specifying skewvar argument.
maxfactorlevels	The maximum levels for factor variables, the default is 30. The argument is used to avoid treating date or time variables as factor variables.
minfactorlevels	If a numeric variable has only several values, it is treated as categorical variable. The default value is 10.
sim	A logical indicating whether to compute p-values by Monte Carlo simulation, in larger than 2 by 2 tables. The default is FALSE.
workspace	If the fisher.test() fucntion requires more workspace, it can be defined here. The default is workspace=2e+05.
ShowStatistic	logic value for whether showing statistics or not. The defaule is FALSE for not showing statistics. Statistics is used for statistical inference such as F value for Chi-square test and T value for student t test. No statistic will be shown because Fisher's exact test jumps past a test statistic and goes straight to a p-value.
ExtractP	Some variables with p value less than a certain threshoud can be extracted for subsequent multivariate regression modeling. The parameter specifies the threshoud of p value for a variable to be extrtacted.

Details

It is common that some categorical variables contain numeric or integer values. For example, the gender variable may contain values 1 and 2, representing male and female respectively. Such a variable can be identified by counting the number of integer values. Thus, the minfactorlevels argument is used to define the minimum value for a variable to be considered as numeric variable.

Value

Table	A table with string values. The mean and standard error are put in a single cell, and connected by plus and minus symbol.
VarExtract	A character vector containing variable names. These extracted variables have p value less than ExtractP in univariate analysis.

Note

No further notes

Author(s)

Zhongheng Zhang Department of emergency medicine, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, China. <zh_zhang1984@zju.edu.cn>

References

Zhang Z. Univariate description and bivariate statistical inference: the first step delving into data.Ann Transl Med. 2016 Mar;4(5):91.

Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernandez P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. Ann Transl Med. 2017 Dec;5(24):484. doi: 10.21037/atm.2017.09.39.

See Also

No other reference

Examples

```
data(df)
a<-twogrps(df,"mort")
print(a,quote = TRUE)
# define skewed variables manually
print(twogrps(df,"mort",skewvar=c("crp","wbc")),
      quote = TRUE)
```

Index