Package: prepdat (via r-universe)

October 11, 2024

Title Preparing Experimental Data for Statistical Analysis

Version 1.0.8

Description Prepares data for statistical analysis (e.g., analysis of variance ;ANOVA) by enabling the user to easily and quickly merge (using the file_merge() function) raw data files into one merged table and then aggregate the merged table (using the prep() function) into a finalized table while keeping track and summarizing every step of the preparation. The finalized table contains several possibilities for dependent measures of the dependent variable. Most suitable when measuring variables in an interval or ratio scale (e.g., reaction-times) and/or discrete values such as accuracy. Main functions included are file_merge() and prep(). The file_merge() function vertically merges individual data files (in a long format) in which each line is a single observation to one single dataset. The prep() function aggregates the single dataset according to any combination of grouping variables (i.e., between-subjects and within-subjects independent variables, respectively), and returns a data frame with a number of dependent measures for further analysis for each cell according to the combination of provided grouping variables. Dependent measures for each cell include among others means before and after rejecting all values according to a flexible standard deviation criteria, number of rejected values according to the flexible standard deviation criteria, proportions of rejected values according to the flexible standard deviation criteria, number of values before rejection, means after rejecting values according to procedures described in Van Selst & Jolicoeur (1994; suitable when measuring reaction-times), standard deviations, medians, means according to any percentile (e.g., 0.05, 0.25, 0.75, 0.95) and harmonic means. The data frame prep() returns can also be exported as a txt file to be used for statistical analysis in other statistical programs.

Depends R (>= 3.0.3)

License GPL-3

LazyData true

URL http://github.com/ayalaallon/prepdat

BugReports http://github.com/ayalaallon/prepdat/issues

Imports dplyr (>= 0.4.2), reshape2 (>= 1.4.1), psych(>= 1.5.4)

Suggests knitr, testthat

RoxygenNote 5.0.1

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

RemoteUrl https://github.com/ayalaallon/prepdat

RemoteRef HEAD

RemoteSha 920bfddbcf47306017b3818270d234f8bb6739a8

Contents

`file_merge` *Vertically Merge Files in a Directory into a Single Large Dataset*

Description

Vertically concatenates files containing data tables in a long format into a single large dataset. In order for the function to work, all files you wish to merge should be in the same format (either txt or csv). This function is very useful for concatenating raw data files of individual subjects in an experiment (in which each line corresponds to a single observation in the experiment) to one raw data file that includes all subjects.

Usage

```
file_merge(
           folder_path = NULL
           , has_header = TRUE
           , new_header = c()
           , raw_file_name = NULL
           , raw_file_extension = NULL
           , file_name = "dataset.txt"
```

```
  , save_table = TRUE
  , dir_save_table = NULL
  , notification = TRUE
 )
```

Arguments

`folder_path` A string with the path of the folder in which files to be merged are searched. Search is recursive (i.e., can search also in subdirectories). `folder_path` must be provided. Default is `NULL`.

`has_header` Logical. If `TRUE`, the function takes the first line of the first file found as the header of the merged table. Default is `TRUE`.

`new_header` String vector with names for columns of the merged table. Default is `c()`. If used, `new_header` should be the same length as the number of columns in the merged table.

`raw_file_name` A string with the name of the files to be searched and then merged. File extension should NOT be included here (see `raw_file_extension`). `raw_file_name` must be provided. Default is `NULL`.

`raw_file_extension` A string with the format of the files (i.e., `csv` or `txt`) to be merged. `raw_file_extension` must be provided. Default is `NULL`.

`file_name` A string with the name of the file of the merged table the function creates in case `save_table` is `TRUE`. Extension of the the file can be txt or csv and should be included. Default is `"dataset.txt"`.

`save_table` Logical. If `TRUE`, saves the merged table. Default is `TRUE`.

`dir_save_table` A string with the path of the folder in which the merged table is saved in case `save_table` is `TRUE`. Default is the path provided in `folder_path`.

`notification` Logical. If `TRUE`, prints messages about the progress of the function. Default is `TRUE`.

Value

The merged table

`finalized_stroopdata` *Finalized Table* `prepdat::prep()` *returns for* `stroopdata` *According to the Example in* `prepdat::prep()`.

Description

A data frame containing dependent measures `prep` for each `id` calculated according to grouping variables: block and target_type. `prep()` aggregates the columns for the dependent measures by first dividing them to the levels of the first independent variable in `wthin vars`, and then within each level `prep()` divides the columns according to the next variable in `within_vars` and so forth. Thus, for each dependent measure in this example there are four columns according to the order they where entered in `within_vars` argument in `prep`. For this data frame this argument was `within_vars = c("block", "target_type")`.

Usage

```
data(finalized_stroopdata)
```

Format

A data frame with 15 rows and 98 columns.

Details

The complete list of names of the dependent measures is:

mdvc: mean dvc.

sdvc: SD for dvc.

meddvc: median dvc.

tdvc: mean dvc after rejecting observations above standard deviation criteria specified in sd_criterion.

ntr: number of observations rejected for each standard deviation criterion specified in sd_criterion.

ndvc: number of observations before rejection.

ptr: proportion of observations rejected for each standard deviation criterion specified in sd_criterion.

rminv: harmonic mean of dvc.

prt: dvc according to each of the percentiles specified in percentiles.

mdvd: mean dvd.

merr: mean error.

nrmc: mean dvc according to non-recursive procedure with moving criterion.

nnrmc: number of observations rejected for dvc according to non-recursive procedure with moving criterion.

pnrmc: percent of observations rejected for dvc according to non-recursive procedure with moving criterion.

tnrmc: total number of observations upon which the non-recursive procedure with moving criterion was applied.

mrmc: mean dvc according to modified-recursive procedure with moving criterion.

nmrmc: number of observations rejected for dvc according to modified-recursive procedure with moving criterion.

pmrmc: percent of observations rejected for dvc according to modified-recursive procedure with moving criterion.

tmrmc: total number of observations upon which the modified-recursive procedure with moving criterion was applied.

hrmc: mean dvc according to hybrid-recursive procedure with moving criterion.

nhrmc: number of observations rejected for dvc according to hybrid-recursive procedure with moving criterion.

thrmc: total number of observations upon which the hybrid-recursive procedure with moving criterion was applied.

Examples

```
data(finalized_stroopdata)
head(finalized_stroopdata)
```

---

hybrid_recursive_mc *Hybrid-recursive Outlier Removal Procedure with Moving Criterion*

Description

Hybrid-recursive outlier removal procedure with moving criterion according to Van Selst & Jolicoeur (1994).

Usage

```
hybrid_recursive_mc(exp_cell)
```

Arguments

`exp_cell`	Numeric vector on which the outlier removal method takes place. If experimental cell has 4 trials or less it will result in `NA`.

Value

A vector with the mean of `exp_cell` after removing outliers, percent of trials removed, and total number of trials in `exp_cell` before outlier removal.

References

Grange, J.A. (2015). trimr: An implementation of common response time trimming methods. R Package Version 1.0.0. https://cran.r-project.org/package=trimr

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.

---

modified_recursive_mc *Modified-recursive Outlier Removal Procedure with Moving Criterion*

Description

Modified-recursive outlier removal procedure with moving criterion according to Van Selst & Jolicoeur (1994).

Usage

```
modified_recursive_mc(exp_cell)
```

Arguments

`exp_cell` Numeric vector on which the outlier removal method takes place. If experimental cell has 4 trials or less it will result in `NA`.

Value

A vector with the mean of `exp_cell` after removing outliers, percent of trials removed, number of trials removed in the procedure,and total number of trials in `exp_cell` before outlier removal.

References

Grange, J.A. (2015). trimr: An implementation of common response time trimming methods. R Package Version 1.0.0. https://cran.r-project.org/package=trimr

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.

`non_recursive_mc` *Non-recursive Outlier Removal Procedure with Moving Criterion*

Description

Non-recursive outlier removal procedure with moving criterion according to Van Selst & Jolicoeur (1994).

Usage

```
non_recursive_mc(exp_cell)
```

Arguments

`exp_cell` Numeric vector on which the outlier removal method takes place. If experimental cell has 4 trials or less it will result in `NA`.

Value

A vector with the mean of `exp_cell` after removing outliers, percent of trials removed, number of trials removed in the procedure,and total number of trials in `exp_cell` before outlier removal.

References

Grange, J.A. (2015). trimr: An implementation of common response time trimming methods. R Package Version 1.0.0. https://cran.r-project.org/package=trimr

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.

prep *Creates One Finalized Table Ready for Statistical Analysis*

Description

prep() aggregates a single dataset in a long format according to any number of grouping variables. This makes prep() suitable for aggregating data from various types of experimental designs such as between-subjects, within-subjects (i.e., repeated measures), and mixed designs (i.e., experimental designs that include both between- and within- subjects independent variables). prep() returns a data frame with a number of dependent measures for further analysis for each aggregated cell (i.e., experimental cell) according to the provided grouping variables (i.e., independent variables). Dependent measures for each experimental cell include among others means before and after rejecting observations according to a flexible standard deviation criteria, number of rejected observations according to the flexible standard deviation criteria, proportions of rejected observations according to the flexible standard deviation criteria, number of observations before rejection, means after rejecting observations according to procedures described in Van Selst & Jolicoeur (1994; suitable when measuring reaction-times), standard deviations, medians, means according to any percentile (e.g., 0.05, 0.25, 0.75, 0.95) and harmonic means. The data frame prep() returns can also be exported as a txt or csv file to be used for statistical analysis in other statistical programs.

Usage

```
prep(
   dataset = NULL
   , file_name = NULL
   , file_path = NULL
   , id = NULL
   , within_vars = c()
   , between_vars = c()
   , dvc = NULL
   , dvd = NULL
   , keep_trials = NULL
   , drop_vars = c()
   , keep_trials_dvc = NULL
   , keep_trials_dvd = NULL
   , id_properties = c()
   , sd_criterion = c(1, 1.5, 2)
   , percentiles = c(0.05, 0.25, 0.75, 0.95)
   , outlier_removal = NULL
   , keep_trials_outlier = NULL
   , decimal_places = 4
   , notification = TRUE
   , dm = c()
   , save_results = TRUE
   , results_name = "results.txt"
   , results_path = NULL
   , save_summary = TRUE
)
```

Arguments

`dataset` Name of the data frame in R that contains the long format table after merging the individual data files using `file_merge()`. Either `dataset` or `file_name` must be provided. Default is `NULL`.

`file_name` A string with the name of a txt or csv file (including the file extension, e.g. `"my_data.txt"`) with the merged table in case the user already merged the individual data files. Either `dataset` or `file_name` must be provided. Default is `NULL`.

`file_path` A string with the path of the folder in which `file_name` is located. If `file_name` was used, then `file_path` must be provided. Default is `NULL`.

`id` A string with the name of the column in `file_name` or in `dataset` that contains the variable specifying the case identifier (i.e., the variable upon which the measurement took place; e.g., `"subject_number"`). This should be a unique value per case. Values in this column must be numeric. Argument must be provided. Default is `NULL`.

`within_vars` String vector with names of grouping variables in `file_name` or in `dataset` that contain independent variables manipulated (or observed) within-ids (i.e., within-subjects, repeated measures). Single or multiple values must be specified as a string (e.g., `c("SOA", "condition")`) according to the hierarchical order you wish. Note that the order of the names in `within_vars()` is important because `prep()` aggregates the data for the dependent measures by first dividing them to the levels of the first grouping variable in `witin_vars()`, and then within each of those levels `prep()` divides the data according to the next variable in `within_vars()` and so forth. Values in these columns must be numeric. Either `within_vars` or `between_vars` (or both) arguments must be provided. Default is `c()`.

`between_vars` String vector with names of grouping variables in `file_name` or in `dataset` that contain independent variables manipulated (or observed) between-ids (i.e., between-subjects). Single or multiple values must be specified as a string (e.g., `c("order")`). Order of the names in `between_vars()` does not matter. Values in this column must be numeric. Either `between_vars` or `within_vars` (or both) arguments must be provided. Default is `c()`.

`dvc` A string with the name of the column in `file_name` or in `dataset` that contains the dependent variable (e.g., "rt" for reaction-time as a dependent variable). Values in this column must be in an interval or ratio scale. Either `dvc` or `dvd` (or both) arguments must be provided. Default is `NULL`.

`dvd` A string with the name of the column in `file_name` or in `dataset` that contains the dependent variable (e.g., `"ac"` for accuracy as a dependent variable). Values in this column must be numeric and discrete (e.g., 0 and 1). Either `dvc` or `dvd` (or both) arguments must be provided. Default is `NULL`.

`keep_trials` A string. Allows deleting unnecessary observations and keeping necessary observations in `file_name` or in `dataset` according to logical conditions specified as a string. For example, if the dataset contains practice trials for each subject, these trials should not be included in the aggregation. The user should remove these trials by specifying how they were coded in the raw data (i.e., data before aggregation). For example, if practice trials are the ones for which the "block" column in the raw data tables equals to zero, the `keep_trials` argument should be `"raw_data$block !== 0"`. `raw_data` is the internal object in `prep()` representing the merged table. All logical conditions in `keep_trials` should be put in the same string and be concatenated by `&` or `|`. Logical conditions for this argument can relate to different columns in the merged table. Note that all further arguments of `prep()` will relate to the remaining observations in the merged table. Default is `NULL`.

`drop_vars` String vector with names of columns to delete in `file_name` or in `dataset`. Single or multiple values must be specified as a string (e.g., `c("font_size")`). Order of the names in `drop_vars` does not matter. Note that all further arguments of `prep()` will relate to the remaining variables in the merged table. Default is `c()`.

`keep_trials_dvc`

A string. Allows deleting unnecessary observations and keeping necessary observations in `file_name` or in `dataset` for calculations and aggregation of the dependent variable in `dvc` according to logical conditions specified as a string. Logical conditions should be specified as a string as in the `keep_trials` argument (e.g., `"raw_data$rt > 100 & raw_data$rt < 3000 & raw_dada$ac == 1"`). All dependent measures for `dvc` except for those specified in `outlier_removal` will be calculated on the remaining observations. Defalut is `NULL`.

`keep_trials_dvd`

A string. Allows deleting unnecessary observations and keeping necessary observations in `file_name` or in `dataset` for calculations and aggregation of the dependent variable in `dvd` according to logical conditions specified as a string. Logical conditions should be specified as a string as in the `keep_trials` argument (e.g., `raw_data$rt > 100 & raw_data$rt < 3000`). All dependent measures for `dvd` (i.e., `"mdvd"` and `"merr"`) will be calculated on the remaining observations. Default is `NULL`.

`id_properties` String vector with names of columns in `dataset` or in `file_name` that describe the ids (e.g., subjects) in the data and were not manipulated within-or between-ids. For example, in case the user logged for each observation and for each id in an experiment also the age and the gender of the subject, this argument will be `c("age", "gender")`. Order of the names in `id_properties` does not matter. Single or multiple values must be specified as a string. Values in these columns must be numeric. Default is `c()`.

`sd_criterion` Numeric vector specifying a number of standard deviation criteria for which `prep()` will calculate the mean `dvc` for each cell in the finalized table after rejecting observations that did not meet the criterion (e.g., rejecting observations that were more than 2 standard deviations above or below the mean of that cell). Values in this vector must be numeric. Default is `c(1, 1.5, 2)`.

`percentiles` Numeric vector containing wanted percentiles for `dvc`. Values in this vector must be decimal numbers between 0 to 1. Percentiles are calculated according to `type = 7` (see quantile for more information). Default is `c(0.05, 0.25, 0.75, 0.95)`.

outlier_removal
: Numeric. Specifies which outlier removal procedure with moving criterion to calculate for dvc according to procedures described by Van Selst & Jolicoeur (1994). If 1 then non-recursive procedure is calculated, if 2 then modified recursive procedure is calculated, if 3 then hybrid recursive procedure is calculated. Moving criterion is according to Table 4 in Van Selst & Jolicoeur (1994). If experimental cell has 4 trials or less it will result in NA. Default is NULL.

keep_trials_outlier
: A string. Allows deleting unnecessary observations and keeping necessary observations in file_name or in dataset for calculations and aggregation of the outlier removal procedures by Van Selst & Jolicoeur (1994). Logical conditions should be specified as a string as in the keep_trials argument (e.g., "raw_data$ac == 1"). outlier_removal procedure will be calculated on the remaining observations. Defalut is NULL.

decimal_places
: Numeric. Specifies number of decimals to be written in results_name for each value of the dependent measures for dvc. Value must be numeric. Default is 4.

notification
: Logical. If TRUE, prints messages about the progress of the function. Default is TRUE.

dm
: String vector with names of dependent measures the function returns. If empty (i.e., c()) the function returns a data frame with all possible dependent measures in prep(). Values in this vector must be strings from the following list: "mdvc", "sdvc", "meddvc", "tdvc", "ntr", "ndvc", "ptr", "prt", "rminv", "mdvd", "merr". Default is c(). See Value section below for more details.

save_results
: Logical. If TRUE, the function creates a txt file containing the returned data frame. Default is TRUE.

results_name
: A string with the name of the file prep returns in case save_results is TRUE. Extension of the file can be txt or csv and should be included. Default is "results.txt".

results_path
: A string with the path of the folder in which results_name will be saved. Default is the path provided in file_path. In case no path was provided in file_path, results_path must be provided.

save_summary
: Logical. if TRUE, creates a summary file in the same format as results_name. Default is TRUE.

Value

A data frame with dependent measures for the dependent variables in dvc and dvd by id and grouping variables.

The first column in the finalized table is the id column. In case id_properties was used, the next columns will be the value of each id_properties for each id.

If between_vars was used then the next column{}s will be the value of each beween_vars for each id.

The next columns of the finalized table contain the dependent measures according to the design specified. If within_vars was used, then the data for each dependent measure was first divided according to the levels of the first grouping variable in witin_vars(), and then within each of those levels `prep()` divided the data according to the next variable in `within_vars()` and so forth. The dependent measures in the finalized table are:

`mdvc`: mean `dvc`.

`sdvc`: SD for `dvc`.

`meddvc`: median `dvc`.

`tdvc`: mean `dvc` after rejecting observations above standard deviation criteria specified in `sd_criterion`.

`ntr`: number of observations rejected for each standard deviation criterion specified in `sd_criterion`.

`ndvc`: number of observations before rejection.

`ptr`: proportion of observations rejected for each standard deviation criterion specified in `sd_criterion`.

`rminv`: harmonic mean of `dvc`.

`prt`: `dvc` according to each of the percentiles specified in `percentiles`.

`mdvd`: mean `dvd`.

`merr`: mean error.

`nrmc`: mean `dvc` according to non-recursive procedure with moving criterion.

`nnrmc`: number of observations rejected for `dvc` according to non-recursive procedure with moving criterion.

`pnrmc`: percent of observations rejected for `dvc` according to non-recursive procedure with moving criterion.

`tnrmc`: total number of observations upon which the non-recursive procedure with moving criterion was applied.

`mrmc`: mean `dvc` according to modified-recursive procedure with moving criterion.

`nmrmc`: number of observations rejected for `dvc` according to modified-recursive procedure with moving criterion.

`pmrmc`: percent of observations rejected for `dvc` according to modified-recursive procedure with moving criterion.

`tmrmc`: total number of observations upon which the modified-recursive procedure with moving criterion was applied.

`hrmc`: mean `dvc` according to hybrid-recursive procedure with moving criterion.

`nhrmc`: number of observations rejected for `dvc` according to hybrid-recursive procedure with moving criterion.

`thrmc`: total number of observations upon which the hybrid-recursive procedure with moving criterion was applied.

References

Grange, J.A. (2015). trimr: An implementation of common response time trimming methods. R Package Version 1.0.1. https://CRAN.R-project.org/package=trimr

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.

Examples

```
data(stroopdata)
finalized_stroopdata <- prep(
           dataset = stroopdata
           , file_name = NULL
           , file_path = NULL
           , id = "subject"
           , within_vars = c("block", "target_type")
           , between_vars = c("order")
           , dvc = "rt"
           , dvd = "ac"
           , keep_trials = NULL
           , drop_vars = c()
          , keep_trials_dvc = "raw_data$rt > 100 & raw_data$rt < 3000 & raw_data$ac == 1"
           , keep_trials_dvd = "raw_data$rt > 100 & raw_data$rt < 3000"
           , id_properties = c()
           , sd_criterion = c(1, 1.5, 2)
           , percentiles = c(0.05, 0.25, 0.75, 0.95)
           , outlier_removal = 2
           , keep_trials_outlier = "raw_data$ac == 1"
           , decimal_places = 0
           , notification = TRUE
           , dm = c()
           , save_results = FALSE
           , results_name = "results.txt"
           , results_path = NULL
           , save_summary = FALSE
         )
```

read_data *Reads a File in a txt or csv Format that Contains a Table and Creates a Data Frame from it*

Description

Reads a File in a txt or csv Format that Contains a Table and Creates a Data Frame from it

Usage

```
read_data(file_name, file_path = NULL, notification = TRUE)
```

Arguments

file_name A string with the name of the file to be read into R. The string should include the file extension.

file_path A string with the path to the folder in which the file to read is located. Default is NULL.

notification Logical. If TRUE, prints messages about the progress of the function. Default is TRUE.

Value

A data frame of the table specified in `file_name`.

`stroopdata` *Reaction-times and accuracy for color naming in a Stroop task (e.g., Stroop, 1935).*

Description

A dataset containing reaction-times, accuracy, and other attributes of 5400 experimental trials.

Usage

```
data(stroopdata)
```

Format

A data frame with 5401 rows and 10 columns:

subject Case identifier, in numerals

block Percent of congruent target_type trials in a block. 1 means 80 percent congruent, 2 means 20 percent congruent

age Age of subject, in integers

gender Gender of subject, in integers. 1 means male, 2 means female

order Order of blocks, in integers. 1 means subject did 80 percent congruent block first and 20 percent congruent block second. 2 means subject did 20 percent congruent block first and 80 percent congruent block second.

font_size Font size of the stimulus, in integers

trial_num Trial number, in integers

target_type Type of stimulus for a given trial. 1 means congruent stimulus, 2 means incongruent stimulus

rt Reaction time, in milliseconds

ac Accuracy, 1 means correct, 0 means incorrect

References

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643.

Examples

```
data(stroopdata)
head(stroopdata)
```

Index