Basic descriptive statistics useful for psychometrics

Description

There are many summary statistics available in R; this function provides the ones most useful for scale construction and item analysis in classic psychometrics. Range is most useful for the first pass in a data set, to check for coding errors.

Usage

```
describe(x, na.rm = TRUE, interp=FALSE,skew = TRUE, ranges = TRUE,trim=.1,
type=3,check=TRUE,fast=NULL)
describeData(x,head=4,tail=4)
```

Arguments

`x`
: A data frame or matrix

`na.rm`
: The default is to delete missing data. na.rm=FALSE will delete the case.

`interp`
: Should the median be standard or interpolated

`skew`
: Should the skew and kurtosis be calculated?

`ranges`
: Should the range be calculated?

`trim`
: trim=.1 – trim means by dropping the top and bottom trim fraction

`type`
: Which estimate of skew and kurtosis should be used? (See details.)

`check`
: Should we check for non-numeric variables? Slower but helpful.

`fast`
: if TRUE, will do n, means, sds, ranges for an improvement in speed. If NULL, will switch to fast mode for large (ncol * nrow > 10^7) problems, otherwise defaults to fast = FALSE

`head`
: show the first 1:head cases for each variable in describeData

`tail`
: Show the last nobs-tail cases for each variable in describeData

Details

In basic data analysis it is vital to get basic descriptive statistics. Procedures such as `summary` and hmisc::describe do so. The describe function in the `psych` package is meant to produce the most frequently requested stats in psychometric and

psychology studies, and to produce them in an easy to read data.frame. The results from describe can be used in graphics functions (e.g., `error.crosses`).

The range statistics (min, max, range) are most useful for data checking to detect coding errors, and should be found in early analyses of the data.

Although describe will work on data frames as well as matrices, it is important to realize that for data frames, descriptive statistics will be reported only for those variables where this makes sense (i.e., not for alphanumeric data).

If the check option is TRUE, variables that are categorical or logical are converted to numeric and then described. These variables are marked with an * in the row name. This is somewhat slower. Note that in the case of categories or factors, the numerical ordering is not necessarily the one expected. For instance, if education is coded "high school", "some college" , "finished college", then the default coding will lead to these as values of 2, 3, 1. Thus, statistics for those variables marked with * should be interpreted cautiously (if at all).

In a typical study, one might read the data in from the clipboard (`read.clipboard`), show the splom plot of the correlations (`pairs.panels`), and then describe the data.

na.rm=FALSE is equivalent to describe(na.omit(x))

When finding the skew and the kurtosis, there are three different options available. These match the choices available in skewness and kurtosis found in the e1071 package (see Joanes and Gill (1998) for the advantages of each one).

If we define *m_r = [sum(X- mx)^r]/n* then

Type 1 finds skewness and kurtosis by *g_1 = m_3/(m_2)^{3/2}* and *g_2 = m_4/(m_2)^2 -3*.

Type 2 is *G1 = g1 * √{n *(n-1)}/(n-2)* and *G2 = (n-1)*[(n+1)g2 +6]/((n-2)(n-3))*.

Type 3 is *b1 = [(n-1)/n]^{3/2} m_3/m_2^{3/2}* and *b2 = [(n-1)/n]^{3/2} m_4/m_2^2)*.

The additional helper function `describeData` just scans the data array and reports on whether the data are all numerical, logical/factorial, or categorical. This is a useful check to run if trying to get descriptive statistics on very large data sets where to improve the speed, the check option is FALSE.

The fast=TRUE option will lead to a speed up of about 50% for larger problems by not finding all of the statistics (see NOTE)

Value

A data.frame of the relevant statistics:
item name
item number
number of valid cases
mean
standard deviation
trimmed mean (with trim defaulting to .1)
median (standard or interpolated
mad: median absolute deviation (from the median)
minimum
maximum
skew
kurtosis
standard error

Note

For very large data sets that are data.frames, describe can be rather slow. Converting the data to a matrix first is recommended. However, if the data are of different types, (factors or logical), this is not possible. If the data set includes columns of character data, it is also not possible. Thus, a quick pass with `describeData` is recommended.

For the greatest speed, at the cost of losing information, do not ask for ranges or for skew and turn off check. This is done automatically if the fast option is TRUE or for large data sets.

Note that by default, fast=NULL. But if the number of cases x number of variables exceeds (ncol * nrow > 10^7), fast will be set to TRUE. This will provide just n, mean, sd, min, max, range, and standard errors. To get all of the statistics (but at a cost of greater time) set fast=FALSE.

The problem seems to be a memory limitation in that the time taken is an accelerating function of nvars * nobs. Thus, for a largish problem (72,000 cases with 1680 variables) which might take 330 seconds, doing it as two sets of 840 variable cuts the time down to 80 seconds.

Author(s)

http://personality-project.org/revelle.html

Maintainer: William Revelle revelle@northwestern.edu

References

Joanes, D.N. and Gill, C.A (1998). Comparing measures of sample skewness and kurtosis. The Statistician, 47, 183-189.

See Also

`describe.by`, `skew`, `kurtosi` `interp.median`, `pairs.panels`, `read.clipboard`, `error.crosses`

Examples

```
data(sat.act)
describe(sat.act)

describe(sat.act,skew=FALSE)
describeData(sat.act)
```
