Package 'DasGuptR'

August 4, 2026

Type Package

Title Das Gupta Standardisation and Decomposition

Version 2.2.0

Date 2026-08-04

Maintainer Josiah King <josiah.king@ed.ac.uk>

Description Implementation of Das Gupta's standardisation and decomposition of population rates, as set out ``Standardization and decomposition of rates: A user's manual", Das Gupta (1993) <https://www2.census.gov/library/publications/1993/demographics/p23-186.pdf>. The goal of these methods is to calculate adjusted rates based on compositional 'factors' and quantify the contribution of each factor to the difference in crude rates between populations. The package offers functionality to handle various scenarios for any number of factors and populations, where said factors can be comprised of vectors across sub-populations (including cross-classified population breakdowns), and with the option to specify user-defined rate functions.

License GPL (>= 3)

URL https://github.com/josiahpjking/DasGuptR

BugReports https://github.com/josiahpjking/DasGuptR/issues

Depends R (>= 4.1.0)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.3

Suggests R.rsp, testthat (>= 3.0.0)

VignetteBuilder R.rsp

Config/testthat/edition 3

NeedsCompilation no

Author Josiah King [aut, cre],
Ben Matthews [aut],
Prithwis Das Gupta [aut]

Repository CRAN

Date/Publication 2026-08-04 17:10:15 UTC

Contents

ccwrap . . . 2
dg2pop . . . 3
dg354 . . . 4
dg611 . . . 5
dg612 . . . 6
dgcc . . . 6
dgeg2_1 . . . 7
dgeg2_2 . . . 8
dgeg2_3 . . . 9
dgeg2_4 . . . 10
dgeg3_1 . . . 11
dgeg4_1 . . . 12
dgeg4_3 . . . 13
dgeg4_4 . . . 14
dgeg5_1 . . . 15
dgeg5_3 . . . 16
dgeg6_5 . . . 17
dgeg6_6 . . . 18
dgnpop . . . 20
dg_check . . . 24
dg_plot . . . 25
dg_table . . . 25
reconv . . . 26
split_popstr . . . 27

Index **28**

`ccwrap` *Wrapper for cross-classified data that standardises rates across a pair of populations. Because these are (r+r')/2 * Q(a_i), this requires 1) doing the rate standardisation on each sub-population, 2) performing the standardisation on the cross classified structure variables, 3) multiplying and (optionally) aggregating up*

Description

Wrapper for cross-classified data that standardises rates across a pair of populations. Because these are (r+r')/2 * Q(a_i), this requires 1) doing the rate standardisation on each sub-population, 2) performing the standardisation on the cross classified structure variables, 3) multiplying and (optionally) aggregating up

Usage

```
ccwrap(
  pw,
  pop,
  factors,
  id_vars,
  crossclassified,
  agg,
  ratefunction = NULL,
  quietly = TRUE
)
```

Arguments

pw	dataframe containing two populations worth of factor data, with columns specifying 1) population and 2) each rate-factor to be considered. must have column named "pop" indicating the population ID.
pop	name (character string) of variable indicating population
factors	names (character vector) of variables indicating compositional factors
id_vars	character vector of variables indicating sub-populations
crossclassified	character string of variable indicating size of sub-population. If specified, the proportion of each population in a given sub-population (e.g. each age-sex combination) is re-expressed as a product of symmetrical expressions representing the different variables (age, sex) constituting the sub-populations.
agg	logical indicating whether, when cross-classified data is used, to output should be aggregated up to the population level
ratefunction	user defined character string in R syntax that when evaluated specifies the function defining the rate as a function of factors. if NULL then will assume rate is the product of all factors.
quietly	logical indicating whether interim messages should be outputted indicating progress through the P factors

Value

data.frame that includes K-a standardised rates for each population and each factor a, along with differences between standardised rates

---

dg2pop *Standardisation and decomposition of rates over K rate-factors and 2 populations. We suggest using dgnpop, which will internally call this function.*

Description

Standardisation and decomposition of rates over K rate-factors and 2 populations. We suggest using dgnpop, which will internally call this function.

Usage

```
dg2pop(pw, pop, factors, id_vars, ratefunction = NULL, quietly = TRUE)
```

Arguments

pw	dataframe containing two populations worth of factor data, with columns specifying 1) population and 2) each rate-factor to be considered. must have column named "pop" indicating the population ID.
pop	name (character string) of variable indicating population
factors	names (character vector) of variables indicating compositional factors
id_vars	character vector of variables indicating sub-populations
ratefunction	user defined character string in R syntax that when evaluated specifies the function defining the rate as a function of factors. if NULL then will assume rate is the product of all factors.
quietly	logical indicating whether interim messages should be outputted indicating progress through the P factors

Value

named list along set of K factors included in the standardisation. Each list element contains a data.frame that includes K-a standardised rates for each population, along with differences between standardised rates

---

dg354 *Das Gupta equation 3.54. internal function called by dg2pop.*

Description

Das Gupta equation 3.54. internal function called by dg2pop.

Usage

```
dg354(df2, i, pop, factors, id_vars, ratefunction, quietly = TRUE)
```

Arguments

df2	list of 2 population dataframes, in which each one contains data for all factors for the relevant population, along with variables indicating population and sub-populations
i	the index of the factors vector which is not being adjusted for (the alpha in "P-alpha standardised rates")
pop	name (character string) of variable indicating population
factors	names (character vector) of variables indicating compositional factors
id_vars	character vector of variables indicating sub-populations
ratefunction	user defined character string in R syntax that when evaluated specifies the function defining the rate as a function of factors. if NULL then will assume rate is the product of all factors.
quietly	logical indicating whether interim messages should be outputted indicating progress

Value

data.frame object including K-a standardised rates for each population for given factor a, along with differences between standardised rates

---

dg611 *Das Gupta equation 6.11: Standardises rates across populations*

Description

Das Gupta equation 6.11: Standardises rates across populations

Usage

```
dg611(srates, all_p, y, factor)
```

Arguments

srates	a dataframe/tibble object of standardised rates from dg2pop()
all_p	character or numeric vector of all N populations
y	character/numeric indicating a single population
factor	string indicating rate-factor being standardised.

Value

data.frame object including K-a standardised rates for each population for given factor a, across N populations

dg612 *Das Gupta equation 6.12 for a differences between 2 populations when standardised across N populations.*

Description

Das Gupta equation 6.12 for a differences between 2 populations when standardised across N populations.

Usage

```
dg612(srates, all_p, ps, factor, std = TRUE)
```

Arguments

srates	a dataframe output from dg2pop()
all_p	character or numeric vector of all N populations
ps	vector of length 2 specifying a possible pairwise comparison of populations
factor	character string indicating name of factor
std	logical indicating whether standardisation across multiple populations has occurred

Value

data.frame object including K-a standardised-rate-differences for each population for given factor a, across N populations

dgcc *Das Gupta equation 5.36 across N populations: Decomposes cross-classified population structures into a set of symmetric proportions indicating contribution of individual structural variables.*

Description

Das Gupta equation 5.36 across N populations: Decomposes cross-classified population structures into a set of symmetric proportions indicating contribution of individual structural variables.

Usage

```
dgcc(x, pop, id_vars, crossclassified)
```

Arguments

x	dataframe consisting of one population, including variables indicating cross-classified structure, and a variable indicating size of each cell
pop	variable name (character string) containing population identifier
id_vars	character vector of variables indicating cross-classified structure.
crossclassified	variable name (character string) containing cell sizes or proportions

Value

inputted data.frame is returned with the addition of variables for each of the the cross-classified variables representing the contribution to the population size.

---

dgeg2_1 *Das Gupta Example 2.1*

Description

Mean Earnings as the Product of Two Factors for Black Males and White Males, 18 Years and Over: United States, 1980

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg2_1
```

Format

A data frame with 2 rows and 3 variables:

pop Population indicator

avg_earnings Mean earnings based on those who earned

earner_prop Proportion of persons who earned

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg2_1)
dgnpop(dgeg2_1,
       pop = "pop",
       factors = c("avg_earnings", "earner_prop")) |>
 dg_table()
```

---

dgeg2_2 *Das Gupta Example 2.2*

Description

Crude Birth Rates as the Product of Three Factors: Austria and Chile, 1981

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg2_2
```

Format

A data frame with 2 rows and 4 variables:

pop Population indicator

birthsw1549 Births per 1000 women aged 15-49

propw1549 Proportion of women aged 15-49

propw Proportion of women out of total population

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg2_2)
dgnpop(dgeg2_2,
       pop = "pop",
       factors = c("birthsw1549", "propw1549", "propw")) |>
  dg_table()
```

dgeg2_3 *Das Gupta Example 2.3*

Description

Percentage Having Nonmarital Live Births as the Product of Four Factors for White Women Aged 15 to 19: United States, 1971 and 1979

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg2_3
```

Format

A data frame with 2 rows and 5 variables:

pop Population indicator

birth_preg Number of nonmarital live births per 100 nonmarital pregnancies

preg_actw Nonmarital pregnancy rate among sexually active single women

actw_prop Proportion of single women who are sexually active

w_prop Proportion of total women who are single

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg2_3)
dgnpop(dgeg2_3,
       pop = "pop",
       factors = c("birth_preg", "preg_actw", "actw_prop", "w_prop")) |>
  dg_table()
```

dgeg2_4 *Das Gupta Example 2.4*

Description

Total Fertility Rate as the Product of Five Factors: South Korea, 1960 and 1970

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg2_4
```

Format

A data frame with 2 rows and 6 variables:

pop Population indicator

prop_m Index of proportion married

noncontr Index of noncontraception

abort Index of induced abortion

lact Index of lactational infecundability

fecund Total fecundity rate

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg2_4)
dgnpop(dgeg2_4,
       pop = "pop",
       factors = c("prop_m", "noncontr", "abort", "lact", "fecund")) |>
  dg_table()
```

dgeg3_1 *Das Gupta Example 3.1*

Description

Crude Rate of Natural Increase as a Function of Crude Birth Rate and Crude Death Rate: United States, 1940 and 1960

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg3_1
```

Format

A data frame with 2 rows and 3 variables:

pop Population indicator

crude_birth Crude birth rate (number of births per 1000 people)

crude_death Crude death rate (number of deaths per 1000 people)

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg3_1)
dgnpop(dgeg3_1,
       pop = "pop",
       factors = c("crude_birth", "crude_death"),
       ratefunction = "crude_birth-crude_death") |>
  dg_table()
```

dgeg4_1 *Das Gupta Example 4.1*

Description

Female Intrinsic Growth Rate per Person as a Function of Two Vector-Factors: United States, 1960 and 1965

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg4_1
```

Format

A data frame with 18 rows and 4 variables:

age_group Age group: 1 = `[10,15)`, 2 = `[15,20)`, ..., 8 = `[45,50)`, 9 = `[50,55)`

pop Population indicator

Lx Stationary population values (5Lx)

mx Age-specific fertility rates for females

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg4_1)
## rate function:
RF4.1 <- function(A, B) {
  idx <- seq_len(length(A))
  mu0 <- sum(A * B / 100000)
  mu1 <- sum((5 * idx + 7.5) * A * B / 100000)
  r1 <- log(mu0) * (mu0 / mu1)
  while (TRUE) {
    Nr1 <- 0
    Dr1 <- 0
    Nr1 <- Nr1 + sum(exp(-r1 * (5 * idx + 7.5)) * A * (B / 100000))
    Dr1 <- Dr1 - sum((5 * idx + 7.5) * exp(-r1 * (5 * idx + 7.5)) * A * (B / 100000))
    r2 <- r1 - ((Nr1 - 1) / Dr1)
    if (abs(r2 - r1) <= .0000001) {
      break
    }
    r1 <- r2
```

```
  }
  return(r2)
}

## crude rates:
RF4.1(A = dgeg4_1$Lx[1:9], B = dgeg4_1$mx[1:9])
RF4.1(A = dgeg4_1$Lx[10:18], B = dgeg4_1$mx[10:18])

## decomposition:
dgnpop(dgeg4_1,
       pop = "pop",
       factors = c("Lx", "mx"),
       id_vars = "age_group",
       ratefunction = "RF4.1(Lx,mx)") |>
  dg_table()
```

`dgeg4_3` *Das Gupta Example 4.3*

Description

Crude Birth Rate per 1,000 as a Function of Three Vector-Factors: Taiwan, 1960 and 1970

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg4_3
```

Format

A data frame with 14 rows and 5 variables:

agegroup Age group: 1 = `[15,19]`, 2 = `[20,24]`, ..., 6 = `[40,44]`, 7 = `[45,49]`

pop Population indicator

bm Age-specific marital fertility rates per 1000 women

mw Proportion of married women among total women in age group i

wp Total women in age group i as a proportion of the total population

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg4_3)
dgnpop(dgeg4_3,
       pop = "pop",
       factors = c("bm", "mw", "wp"),
       ratefunction = "sum(bm*mw*wp)") |>
  dg_table()
```

---

dgeg4_4 *Das Gupta Example 4.4*

Description

Illegitimacy Ratio as a Function of Four Vector-Factors: United States, White Population, 1963 and 1983

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg4_4
```

Format

A data frame with 12 rows and 6 variables:

pop Population indicator

agegroup Age group: 1 = `[15,19]`, 2 = `[20,24]`, ..., 5 = `[35,39]`, 6 = `[40,44]`

A Age-structure of women in childbearing ages (number of women in age group as a proportion of total women age 15-44)

B Marital status structure within childbearing age groups (proportion of women in age group i who are unmarried)

C Age-specific nonmarital fertility rate (proportion of unmarried women in age group i giving birth)

D Age-specific marital fertility rate (proportion of married women in age group i giving birth)

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg4_4)

## Option 1:
dgnpop(dgeg4_4,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "sum(A*B*C) / (sum(A*B*C) + sum(A*(1-B)*D))") |>
  dg_table()

## alternatively, define a separate rate function:
myratef <- function(a, b, c, d) {
  return(sum(a * b * c) / (sum(a * b * c) + sum(a * (1 - b) * d)))
}

dgnpop(dgeg4_4,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "myratef(A,B,C,D)") |>
  dg_table()
```

---

dgeg5_1 *Das Gupta Example 5.1*

Description

Population Sizes (Percents) and Household Headship Rates per 100 by Age Groups: United States, 1970 and 1985

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg5_1
```

Format

A data frame with 26 rows and 4 variables:

age_group Age group: 1 = `[15,19]`, 2 = `[20,24]`, ..., 11 = `[65,69]`, 12 = `[70,74]`, 13 = `[75,Inf)`

pop Population indicator

size Percentage of population in age group i

rate Age-specific household headship rate

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg5_1)

## Option 1: create separate factor of sub-group size as a proportion:
dgeg5_1$age_str <- dgeg5_1$size / 100

dgnpop(dgeg5_1,
       pop = "pop",
       factors = c("age_str", "rate"),
       id_vars = "age_group",
       ratefunction = "sum(age_str*rate)") |>
  dg_table()

## Option 2: sub-group proportion created within rate function
dgnpop(dgeg5_1,
       pop = "pop",
       factors = c("size", "rate"),
       id_vars = "age_group",
       ratefunction = "sum( (size/sum(size)) * rate )") |>
  dg_table()

## Option 3: use crossclassified to indicate sub-groups
dgnpop(dgeg5_1,
       pop = "pop",
       factors = c("rate"),
       id_vars = "age_group",
       crossclassified = "size") |>
  dg_table()
```

---

dgeg5_3 *Das Gupta Example 5.3*

Description

Population (in thousands) and Death Rates (per 1,000 Population) by Age and Race: United States, 1970 and 1985

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg5_3
```

Format

A data frame with 44 rows and 5 variables:

race race group: 1 = White, 2 = non-White

age Age group: 1 = `[0,1)`, 2 = `[1,4]`, 3 = `[5,14]`, 4 = `[15,24]`, ..., 9 = `[65,74]`, 10 = `[75,84]`, 11 = `[85,Inf)`

pop Population indicator

size Number of people in age-race-group i

rate Age-race-specific death rate (deaths per 1000 people)

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg5_3)

dgnpop(dgeg5_3,
       pop = "pop",
       factors = c("rate"),
       id_vars = c("race", "age"),
       crossclassified = "size") |>
  dg_table()
```

`dgeg6_5` *Das Gupta Example 6.5*

Description

Illegitimacy Ratio as a Function of Four Vector-Factors: United States, White Population, 1963, 1968, 1973, 1978, and 1983.

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg6_5
```

Format

A data frame with 30 rows and 6 variables:

pop Population indicator

agegroup Age group:`[15,19]`, `[20,24]`, ..., `[35,39]`, `[40,44]`

A Age-structure of women in childbearing ages (number of women in age group as a proportion of total women age 15-44)

B Marital status structure within childbearing age groups (proportion of women in age group i who are unmarried)

C Age-specific nonmarital fertility rate (proportion of unmarried women in age group i giving birth)

D Age-specific marital fertility rate (proportion of married women in age group i giving birth)

Source

Das Gupta, Prithwis (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports, Series P23-186. U.S. Government Printing Office, Washington, DC.

Examples

```
data(dgeg6_5)

dgnpop(dgeg6_5,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "1000*sum(A*B*C) / (sum(A*B*C) + sum(A*(1-B)*D))") |>
  dg_table()

dgnpop(dgeg6_5,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "1000*sum(A*B*C) / (sum(A*B*C) + sum(A*(1-B)*D))") |>
  dg_plot()
```

---

dgeg6_6 *Das Gupta Example 6.6: US Population Birth Rates*

Description

Population and Birth Rates by Nine Age-Sex Groups: United States, 1940 to 1990

To allow for direct comparison with Das Gupta's original publication we have preserved these datasets as they were originally presented

Usage

```
dgeg6_6
```

Format

A data frame with 459 rows and 4 variables:

year Population indicator

agebin Age-sex group: `F [10,14]`, `F [15,19]`, `F [20,24]`, `F [25,29]`, `F [30,34]`, `F [35,39]`, `F [40,44]`, `F [45,49]`, `Remainder (F [0,10), F [50, Inf), M [0,Inf))`

thous Population (in thousands) in age-sex group i

birthrate Age-sex-specific birth rates (per 1000 population)

Source

"Standardization and decomposition of rates: A user's manual", Das Gupta (1993)

Examples

```
data(dgeg6_6)
# for speed, take every 5th year
dgeg6_6 <- dgeg6_6[dgeg6_6$year %in% seq(1940,1990,5), ]

dgo_us <- dgnpop(dgeg6_6,
                 pop = "year",
                 factors = c("birthrate"),
                 id_vars = "agebin",
                 crossclassified = "thous")

dg_plot(dgo_us)

# Recreating Figure 1 (p.112) from Das Gupta 1993
# crude rates
plot(x = seq(1940,1990,5),
     y = by(dgeg6_6, dgeg6_6$year,
            \(x) sum(x$birthrate * (x$thous/sum(x$thous)))),
     type = "l", ylim = c(13,33))
# rates with 1940 as standard
baseline = dgeg6_6[dgeg6_6$year == 1940, ]
lines(x = seq(1940,1990,5),
      y = by(dgeg6_6, dgeg6_6$year,
             \(x) sum(x$birthrate * (baseline$thous/sum(baseline$thous)))),
      lty = "dotted")
# rates with 1990 as standard
baseline = dgeg6_6[dgeg6_6$year == 1990, ]
lines(x = seq(1940,1990,5),
      y = by(dgeg6_6, dgeg6_6$year,
             \(x) sum(x$birthrate * (baseline$thous/sum(baseline$thous)))),
      lty = "dashed")
# rates standardised with Das Gupta's method
lines(x = seq(1940,1990,5),
```

```
y = dgo_us$rate[dgo_us$factor=="birthrate"],
lty = "longdash", col = "red")
```

---

dgnpop *Prithwis Das Gupta's 1993 standardisation and decomposition of rates over K rate-factors and N populations.*

Description

Prithwis Das Gupta's 1993 standardisation and decomposition of rates over K rate-factors and N populations.

Usage

```
dgnpop(
  x,
  pop,
  factors,
  id_vars = NULL,
  crossclassified = NULL,
  ratefunction = NULL,
  agg = TRUE,
  baseline = NULL,
  quietly = TRUE,
  diffs = FALSE
)
```

Arguments

x	dataframe or tibble object, with columns specifying 1) population, 2) each rate-factor to be considered, and (optionally) 3) variables indicating underlying sub-populations
pop	name (character string) of variable indicating population
factors	names (character vector) of variables indicating compositional factors
id_vars	character vector of variables indicating sub-populations
crossclassified	character string of variable indicating size of sub-population. If specified, the proportion of each population in a given sub-population (e.g. each age-sex combination) is re-expressed as a product of symmetrical expressions representing the different variables (age, sex) constituting the sub-populations. These expressions are then used as compositional factors in the standardisation. If NULL, then providing a single variable as a compositional factor that represents the proportion of the population in each given sub-population will combine the contribution of all sub-population variables.

`ratefunction`	user defined character string in R syntax that when evaluated specifies the function defining the rate as a function of factors. if NULL then will assume rate is the product of all factors. When sub-populations are provided, this should aggregate to a summary value (e.g., for the simple product rate this should be provided as `"sum(A*B*C*)"`.). User-defined functions can also be provided, as whatever string is given here will be parsed and evaluated as any other R code (see example eg4.4).
`agg`	logical indicating whether, when cross-classified data is used, to output should be aggregated up to the population level
`baseline`	baseline population to standardise against. if NULL then will do Das Gupta's full N-population standardisation.
`quietly`	logical indicating whether interim messages should be outputted indicating progress through the K factors and N populations
`diffs`	logical indicating whether to return list of standardised rates and rate-differences, or just the standardised rates.

Details

Population rates are often composed of various different compositional factors. Standardisation techniques calculate the rate were a set of factors to be held constant (either with a specific population as standard, or at the average of the populations). Decomposition methods quantify the amount of the difference between two population crude rate that is due to differences in population characteristics.

Das Gupta's general solution for the decomposition of two rates can be written as:

$$\Delta\text{crude-}r = \sum_{\vec{\alpha} \in K} Q(\vec{\alpha}^p) - Q(\vec{\alpha}^{p'})$$

Where K is the set of factors $\alpha, \beta, ..., \kappa$, which may take the form of vectors over sub-populations i. $Q(\vec{\alpha}^p)$ denotes the rate in population p holding all factors other than α — $K \setminus \alpha$ — equal (standardised across populations p and p'). The total crude rate difference is the sum of all standardised-rate differences, and the standardisation Q is expressed as:

$$Q(\vec{\alpha}^p) = \sum_{j=1}^{\lfloor \frac{|K|}{2} \rfloor} \frac{\sum_{L \in \binom{K \setminus \{\alpha\}}{j-1}} f(\{L^p, (K \setminus L)^{p'}, \vec{\alpha}^p\}) + f(\{L^{p'}, (K \setminus L)^p, \vec{\alpha}^p\})}{|K| \binom{|K|-1}{j-1}}$$

Where $f(K)$ is the function that defines the calculation of the rate

Value

data.frame containing K-a standardised rates (or differences) for each population.

- `rate`: standardised rate such that factor a is from population p and all other factors are averaged across populations, f(a^p,...)
- `pop`: population p for which factor a is taken from
- `std.set`: set of N populations (minus p) across which the standardisation has been performed

- `factor`: name of factor a that is being considered, such that for the set of factors K, the {K-a}-standardised rate is returned

Examples

```
## 2 populations, R=ab
data(dgeg2_1)

dgnpop(dgeg2_1,
       pop = "pop",
       factors = c("avg_earnings", "earner_prop")) |>
  dg_table()

## 2 populations, R=abc
data(dgeg2_2)

dgnpop(dgeg2_2,
       pop = "pop",
       factors = c("birthsw1549", "propw1549", "propw")) |>
  dg_table()

## 2 populations, R=abcd
data(dgeg2_3)

dgnpop(dgeg2_3,
   pop = "pop",
   factors = c("birth_preg", "preg_actw", "actw_prop", "w_prop" )) |>
  dg_table()

## 2 populations, R=abcde
data(dgeg2_4)

dgnpop(dgeg2_4,
       pop = "pop",
       factors = c("prop_m", "noncontr", "abort", "lact", "fecund")) |>
  dg_table()

## 2 populations, vector factors, R=sum(abc)
data(dgeg4_3)

dgnpop(dgeg4_3,
       pop = "pop",
       factors = c("bm", "mw", "wp"),
       ratefunction = "sum(bm*mw*wp)") |>
 dg_table()

## 2 populations, R=f(ab)
data(dgeg3_1)

dgnpop(dgeg3_1,
       pop = "pop",
       factors = c("crude_birth", "crude_death"),
       ratefunction = "crude_birth-crude_death") |>
```

```
  dg_table()

## 2 populations, vector factors, R=f(abcd)
data(dgeg4_4)

dgnpop(dgeg4_4,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "sum(A*B*C) / (sum(A*B*C) + sum(A*(1-B)*D))") |>
  dg_table()

#### alternatively:
myratef <- function(a, b, c, d) {
 return(sum(a * b * c) / (sum(a * b * c) + sum(a * (1 - b) * d)))
}

dgnpop(dgeg4_4,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "myratef(A,B,C,D)") |>
  dg_table()

## 2 populations, 1 sub-population variable
data(dgeg5_1)

### Option 1:
dgnpop(dgeg5_1,
       pop = "pop",
       factors = c("size","rate"),
       ratefunction = "sum( (size/sum(size)) * rate )") |>
  dg_table()

### Option 2:
dgnpop(dgeg5_1,
       pop = "pop",
       factors = c("rate"),
       id_vars = "age_group",
       crossclassified = "size") |>
  dg_table()

## 2 cross-classified variables, 2 populations, R=sum(w*r)
data(dgeg5_3)

dgnpop(dgeg5_3,
       pop = "pop",
       factors = c("rate"),
       id_vars = c("race", "age"),
       crossclassified = "size") |>
  dg_table()

## 5 populations, R = f(abcd)
```

```
data(dgeg6_5)

dgres <- dgnpop(dgeg6_5,
       pop = "pop",
       factors = c("A", "B", "C", "D"),
       id_vars = "agegroup",
       ratefunction = "1000*sum(A*B*C) / (sum(A*B*C) + sum(A*(1-B)*D))")

dg_table(dgres)
dg_plot(dgres)
```

---

dg_check *Performs some initial checks of the input data for dgnpop(), providing messages where issues are flagged*

Description

Performs some initial checks of the input data for dgnpop(), providing messages where issues are flagged

Usage

```
dg_check(x, pop, factors, id_vars = NULL, crossclassified = NULL)
```

Arguments

x	dataframe or tibble object, with columns specifying 1) population, 2) each rate-factor to be considered, and (optionally) 3) variables indicating underlying sub-populations
pop	name (character string) of variable indicating population
factors	names (character vector) of variables indicating compositional factors
id_vars	character vector of variables indicating sub-populations
crossclassified	character string of variable indicating size of sub-population. If specified, the proportion of each population in a given sub-population (e.g. each age-sex combination) is re-expressed as a product of symmetrical expressions representing the different variables (age, sex) constituting the sub-populations. These expressions are then used as compositional factors in the standardisation. If NULL, then providing a single variable as a compositional factor that represents the proportion of the population in each given sub-population will combine the contribution of all sub-population variables.

Value

error message processing

dg_plot — *Creates a plot of Das Gupta standardised rates across the set of populations*

Description

Creates a plot of Das Gupta standardised rates across the set of populations

Usage

```
dg_plot(
  dgo,
  xlab = "population",
  ylab = "rate",
  legend.position = "topright",
  ...
)
```

Arguments

dgo	output from dgnpop()
xlab	x axis label
ylab	y axis label
legend.position	legend position, passed to graphics::legend(), choose from "bottomright", "bottom", "bottomleft", "left", "topleft", "top", "topright", "right" and "center", or provide xy.coords
...	other arguments passed to methods, such as graphics::par()

Value

A plot of each of the set of K-a standardised rates across populations

dg_table — *Creates a small table of Das Gupta standardised rates. If no populations are specified, rates will be shown for all available populations. If only two populations (or if two particular populations are specified), then rate-differences and 'decomposition effects' are calculated and presented.*

Description

Creates a small table of Das Gupta standardised rates. If no populations are specified, rates will be shown for all available populations. If only two populations (or if two particular populations are specified), then rate-differences and 'decomposition effects' are calculated and presented.

Usage

```
dg_table(dgo, pop1 = NULL, pop2 = NULL)
```

Arguments

dgo	output from dgnpop()
pop1	optional name of first population for decomposition (character/numeric)
pop2	optional name of second population for decomposition (character/numeric)

Value

data.frame object with rows for each of the K-a standardised rates and the crude rates, and columns for each of the N populations. When only two populations are included, or if two populations are explicitly specified, standardised rate differences are provided, and are also expressed as a percentage of the crude rate differences (typically referred to as 'decomposition effects').

---

reconv *Scottish Reconvictions data 2004-2016*

Description

Official statistics on reconvictions in Scotland, 2004 to 2016 inclusive, split by sex and 6 age groups

Usage

```
reconv
```

Format

A data frame with 130 rows and 8 variables:

year Population indicator

Sex Sex group (Female, Male)

Age Age group: `(0,20]`, `[21,25]`, `[26,30]`, `[31,40]`, `(40,Inf)`

convicted_population Total number of offenders who either received a non-custodial conviction or were released from a custodial sentence in a given year

offenders Number of offenders in a given age-sex group

reconvicted Number of people in a given age-sex group who were reconvicted one or more times

reconvictions Number of reconvictions in a given age-sex group

prev_rate age-sex specific reconviction rate: percentage of offenders in a given age-sex group who were reconvicted one or more times

Source

Scottish Government (2019) Reconviction rates in Scotland: 2016-2017 offender cohort, Main publication tables

Examples

```
data(reconv)

# dgnpop(reconv,
#        pop="year",
#        factors=c("prev_rate"),
#        id_vars=c("Sex","Age"),
#        crossclassified="offenders") |>
#   dg_plot()
```

`split_popstr` *Das Gupta equation 5.36 for a single population: Decomposes cross-classified population structures into a set of symmetric proportions indicating contribution of individual structural variables.*

Description

Das Gupta equation 5.36 for a single population: Decomposes cross-classified population structures into a set of symmetric proportions indicating contribution of individual structural variables.

Usage

```
split_popstr(x, id_vars, nvar)
```

Arguments

`x`	dataframe consisting of one population, including variables indicating cross-classified structure, and a variable indicating size of each cell
`id_vars`	character vector of variables indicating cross-classified structure.
`nvar`	variable name (character string) containing cell sizes

Value

inputted data.frame is returned with the addition of variables for each of the the cross-classified variables representing the contribution to the population size.

Index

∗ **datasets**
  dgeg2_1, 7
  dgeg2_2, 8
  dgeg2_3, 9
  dgeg2_4, 10
  dgeg3_1, 11
  dgeg4_1, 12
  dgeg4_3, 13
  dgeg4_4, 14
  dgeg5_1, 15
  dgeg5_3, 16
  dgeg6_5, 17
  dgeg6_6, 18
  reconv, 26

ccwrap, 2

dg2pop, 3
dg2pop(), *5*, *6*
dg354, 4
dg611, 5
dg612, 6
dg_check, 24
dg_plot, 25
dg_table, 25
dgcc, 6
dgeg2_1, 7
dgeg2_2, 8
dgeg2_3, 9
dgeg2_4, 10
dgeg3_1, 11
dgeg4_1, 12
dgeg4_3, 13
dgeg4_4, 14
dgeg5_1, 15
dgeg5_3, 16
dgeg6_5, 17
dgeg6_6, 18
dgnpop, 20
dgnpop(), *25*, *26*

graphics::legend(), *25*
graphics::par(), *25*

reconv, 26

split_popstr, 27