Package ‘ddi’

October 13, 2022

Type Package
Title The Data Defect Index for Samples that May not be IID
Version 0.1.0
Description Implements Meng’s data defect index (ddi), which represents
the degree of sample bias relative to an iid sample. The data defect
correlation (ddc) represents the correlation between the outcome of interest
and the selection into the sample; when the sample selection is independent
across the population, the ddc is zero. Details are in Meng (2018)
Law of Large Populations, Big Data Paradox, and the 2016 US Presidential
Election." Survey estimates from the Cooperative Congressional Election Study
(CCES) is included to replicate the article's results.

Encoding UTF-8
URL https://github.com/kuriwaki/ddi
BugReports http://github.com/kuriwaki/ddi/issues
License GPL (>= 2)
LazyData true
RoxygenNote 7.0.2
Depends R (>= 2.10)
Suggests testthat (>= 2.1.0), dplyr, tibble
NeedsCompilation no
Author Shiro Kuriwaki [aut, cre] (<https://orcid.org/0000-0002-5687-2647>)
Maintainer Shiro Kuriwaki <shirokuriwaki@gmail.com>
Repository CRAN
Date/Publication 2020-01-26 10:50:02 UTC

R topics documented:

  ddc ................................................................. 2
g2016 ............................................................... 3

Index 5
Data Defect Correlation (ddc) is the correlation between response and group membership. It quantifies the correlation between the outcome of interest and the selection into the sample; when the sample selection is independent across members of the population, the ddc is zero. Currently both variables are binary. The data defect index (ddi) is the square of ddc. Squaring the ddc is more useful for characterizing the asymptotics of MSE.

Usage

\[
\text{ddc}(\mu, \muhat, N, n, \text{cv} = \text{NULL})
\]

Arguments

- \(\mu\): Vector of population quantity of interest
- \(\muhat\): Vector for sample estimate
- \(N\): Vector of population size
- \(n\): Vector of sample size
- \(\text{cv}\): Coefficient of variation of the weights, if survey weights exist and \(\muhat\) is the weighted proportion. The coefficient of variation is a summary statistic computed by \(\text{sd}(\text{weights}) / \text{mean}(\text{weights})\).

Value

A vector of ddc of the same length of the input, or a scalar if all input variables are scalars.

References


Examples

```r
library(tibble)
library(dplyr)
data(g2016)

# 1. scalar input
select(g2016, cces_pct_djt_vv, cces_n_vv, tot_votes, votes_djt) %>%
  summarize_all(sum)

## plug those numbers in
```
# Description

Donald Trump’s voteshare in each U.S. state, with survey estimates from the Cooperative Congressional Election Study (pre-election wave). See Meng (2018) referenced below for more details. We focus on unweighted estimates to capture the response patterns, before correcting for any imbalances through weights.

## Usage

g2016

## Format

A data frame with 51 rows (all U.S. states and D.C.)

- **state** state (full name)
- **st** state (abbreviation).
- **pct_djt_voters** Donald J. Trump’s voteshare, the estimand.
- **cces_pct_djt_vv** CCES unweighted proportion of Trump support, one estimate.
- **cces_pct_djtrund_vv** CCES unweighted proportion counting Republican undecideds as Trump voters.
- **votes_djt** Total number of votes by Trump.
- **tot_votes** Turnout in Presidential as total number of votes cast.
- **cces_totdjt_vv** Validated voters intending to vote for Trump. Used as the numerator for the above CCES estimates.
- **cces_n_vv** Validated voters in survey sample. Used as the denominator for the above CCES estimates.
- **vap** Voting Age Population in the state.
- **vep** Voting Eligible Population in the state (estimate from the US Election Project).
Source


References

For an explanation in the context of d.d.i., see Meng (2018) <doi:10.1214/18-AOAS1161SF>

Examples

```r
library(dplyr)
data(g2016)

transmute(g2016, 
  st, 
  ddc = ddc(mu = pct_djt_voters, 
            muhat = cces_pct_djt_vv, 
            N = tot_votes, 
            n = cces_n_vv))
```
Index

* datasets
  g2016, 3

ddc, 2

g2016, 3