Package ‘SDAResources’

October 22, 2021

Type Package

Title Datasets and Functions for 'Sampling: Design and Analysis, 3rd Edition'

Version 0.1.1

Maintainer Yan Lu <yanlu@unm.edu>

Description Includes all the datasets of 'Sampling: Design and Analysis' (3rd edition by Sharon Lohr) in R format and additional functions for analyzing and graphing probability samples.

License GPL-2 | GPL-3

Encoding UTF-8

LazyData true

NeedsCompilation no

Depends R (>= 3.5.0)

RoxygenNote 7.1.1

Suggests rmarkdown, knitr

VignetteBuilder knitr

Author Yan Lu [aut, cre], Sharon Lohr [aut]

Repository CRAN

Date/Publication 2021-10-22 09:20:13 UTC

R topics documented:

agpop .......................................................... 3
agpps .......................................................... 4
agsrs .......................................................... 6
agstrat ........................................................ 6
algebra .......................................................... 7
anthrop ......................................................... 8
anthsrs ....................................................... 8
anthuneq ..................................................... 9
R topics documented:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>artifratio</td>
<td>9</td>
</tr>
<tr>
<td>asafellow</td>
<td>10</td>
</tr>
<tr>
<td>auditresult</td>
<td>11</td>
</tr>
<tr>
<td>auditselect</td>
<td>12</td>
</tr>
<tr>
<td>azcounties</td>
<td>12</td>
</tr>
<tr>
<td>baseball</td>
<td>13</td>
</tr>
<tr>
<td>books</td>
<td>14</td>
</tr>
<tr>
<td>captureci</td>
<td>15</td>
</tr>
<tr>
<td>census1920</td>
<td>16</td>
</tr>
<tr>
<td>census2010</td>
<td>16</td>
</tr>
<tr>
<td>cherry</td>
<td>17</td>
</tr>
<tr>
<td>classes</td>
<td>18</td>
</tr>
<tr>
<td>classpps</td>
<td>18</td>
</tr>
<tr>
<td>classppspjp</td>
<td>19</td>
</tr>
<tr>
<td>college</td>
<td>20</td>
</tr>
<tr>
<td>collegeorg</td>
<td>23</td>
</tr>
<tr>
<td>collshr</td>
<td>25</td>
</tr>
<tr>
<td>coots</td>
<td>28</td>
</tr>
<tr>
<td>counties</td>
<td>29</td>
</tr>
<tr>
<td>crimes</td>
<td>30</td>
</tr>
<tr>
<td>deatdtrees</td>
<td>31</td>
</tr>
<tr>
<td>divorce</td>
<td>32</td>
</tr>
<tr>
<td>emppmf</td>
<td>33</td>
</tr>
<tr>
<td>gini</td>
<td>33</td>
</tr>
<tr>
<td>golfrsr</td>
<td>34</td>
</tr>
<tr>
<td>gpa</td>
<td>35</td>
</tr>
<tr>
<td>healthjournals</td>
<td>36</td>
</tr>
<tr>
<td>htcdf</td>
<td>37</td>
</tr>
<tr>
<td>htop</td>
<td>38</td>
</tr>
<tr>
<td>htsrs</td>
<td>38</td>
</tr>
<tr>
<td>htsrt</td>
<td>39</td>
</tr>
<tr>
<td>hunting</td>
<td>39</td>
</tr>
<tr>
<td>impute</td>
<td>40</td>
</tr>
<tr>
<td>integerwt</td>
<td>41</td>
</tr>
<tr>
<td>intellonline</td>
<td>41</td>
</tr>
<tr>
<td>intelltel</td>
<td>43</td>
</tr>
<tr>
<td>intellwts</td>
<td>43</td>
</tr>
<tr>
<td>intervals_ex40</td>
<td>44</td>
</tr>
<tr>
<td>ipums</td>
<td>45</td>
</tr>
<tr>
<td>journal</td>
<td>47</td>
</tr>
<tr>
<td>measles</td>
<td>48</td>
</tr>
<tr>
<td>mysteries</td>
<td>49</td>
</tr>
<tr>
<td>nhanes</td>
<td>50</td>
</tr>
<tr>
<td>nybight</td>
<td>53</td>
</tr>
<tr>
<td>otters</td>
<td>54</td>
</tr>
<tr>
<td>ozone</td>
<td>54</td>
</tr>
<tr>
<td>pitcount</td>
<td>56</td>
</tr>
<tr>
<td>profresp</td>
<td>56</td>
</tr>
</tbody>
</table>
Description

Data from the 1992 U.S. Census of Agriculture.

Usage

data(agpop)

Format

This data frame contains the following columns:

- **county**: county name (character variable)
- **state**: state abbreviation (character variable)
- **acres92**: number of acres devoted to farms, 1992
agpps data

Description

Data from a without-replacement probability-proportional-to-size sample from agpop data.

Usage

data(agpps)

Format

This data frame contains the following columns:

- **county**: county name (character variable)
- **state**: state abbreviation (character variable)
- **acres92**: number of acres devoted to farms, 1992
- **acres87**: number of acres devoted to farms, 1987
- **acres82**: number of acres devoted to farms, 1982
- **farms92**: number of farms, 1992
- **farms87**: number of farms, 1987
- **farms82**: number of farms, 1982
- **largef92**: number of farms with 1,000 acres or more, 1992
- **largef87**: number of farms with 1,000 acres or more, 1987
- **largef82**: number of farms with 1,000 acres or more, 1982
- **smallf92**: number of farms with 9 acres or fewer, 1992
- **smallf87**: number of farms with 9 acres or fewer, 1987
- **smallf82**: number of farms with 9 acres or fewer, 1982
- **region**: S = south; W = west; NC = north central; NE = northeast

References

largef92: number of farms with 1,000 acres or more, 1992
largef87: number of farms with 1,000 acres or more, 1987
largef82: number of farms with 1,000 acres or more, 1982
smallf92: number of farms with 9 acres or fewer, 1992
smallf87: number of farms with 9 acres or fewer, 1987
smallf82: number of farms with 9 acres or fewer, 1982
region: S = south; W = west; NC = north central; NE = northeast
sizemeas: size measure used to select the pps sample
SelectionProb: inclusion probability for county
SamplingWeight: sampling weight for county
Unit: unit number for indexing joint inclusion probabilities
JtProb_1: columns of joint inclusion probabilities
JtProb_2: columns of joint inclusion probabilities
JtProb_3: columns of joint inclusion probabilities
JtProb_4: columns of joint inclusion probabilities
JtProb_5: columns of joint inclusion probabilities
JtProb_6: columns of joint inclusion probabilities
JtProb_7: columns of joint inclusion probabilities
JtProb_8: columns of joint inclusion probabilities
JtProb_9: columns of joint inclusion probabilities
JtProb_10: columns of joint inclusion probabilities
JtProb_11: columns of joint inclusion probabilities
JtProb_12: columns of joint inclusion probabilities
JtProb_13: columns of joint inclusion probabilities
JtProb_14: columns of joint inclusion probabilities
JtProb_15: columns of joint inclusion probabilities

References

Data from an SRS of size 300 from the 1992 U.S. Census of Agriculture *agpop* data.

Usage
data(agsrs)

Format

Variables are the same as in *agpop* data.

References


Data from a stratified random sample of size 300 from the 1992 U.S. Census of Agriculture *agpop* data.

Usage
data(agstrat)

Format

This data frame contains the following columns:

- **county**: county name (character variable)
- **state**: state abbreviation (character variable)
- **acres92**: number of acres devoted to farms, 1992
- **acres87**: number of acres devoted to farms, 1987
- **acres82**: number of acres devoted to farms, 1982
- **farms92**: number of farms, 1992
- **farms87**: number of farms, 1987
**farms82**: number of farms, 1982

**largef92**: number of farms with 1,000 acres or more, 1992

**largef87**: number of farms with 1,000 acres or more, 1987

**largef82**: number of farms with 1,000 acres or more, 1982

**smallf92**: number of farms with 9 acres or fewer, 1992

**smallf87**: number of farms with 9 acres or fewer, 1987

**smallf82**: number of farms with 9 acres or fewer, 1982

**region**: S = south; W = west; NC = north central; NE = northeast

**rn**: random numbers used to select sample in each stratum

**strwt**: sampling weight for each county in sample

**References**


---

**algebra**

**algebra data**

---

**Description**

Fictional data for an SRS of 12 algebra classes in a city, from a population of 187 classes.

**Usage**

data(algebra)

**Format**

This data frame contains the following columns:

- **class**: class number
- **Mi**: number of students $M_i$ in class
- **score**: score of student on test

**References**


Description
Finger length and height for 3,000 criminals. This data set contains information for the entire population.

Usage
data(anthrop)

Format
This data frame contains the following columns:

- **finger**: length of left middle finger (cm)
- **height**: height (inches)

References

Description
Length of left middle finger and height for an SRS of size 200 from *anthrop* data.

Usage
data(anthsrs)

Format
This data frame contains the following columns:

- **finger**: length of left middle finger (cm)
- **height**: height (inches)
- **wt**: sampling weight
anthuneq

References
Boca Raton, FL: CRC Press.

---

anthuneq anthuneq data

Description
Finger length and height for a with replacement unequal probability sample of size 200 from data anthrop. The probability of selection, \( \psi_i \), was proportional to 24 for \( y < 65 \), 12 for \( y = 65 \), 2 for \( y = 66 \) or 67, and 1 for \( y > 67 \).

Usage
data(anthuneq)

Format
This data frame contains the following columns:

- **finger**: length of left middle finger (cm)
- **height**: height (inches)
- **wt**: sampling weight

References
Boca Raton, FL: CRC Press.

---

artifratio artifratio data

Description
Values from all possible SRSs for an artificial population in Chapter 4 of SDA.

Usage
data(artifratio)
Format

This data frame contains the following columns:

sample: sample number
i1: first unit in sample
i2: second unit in sample
i3: third unit in sample
i4: fourth unit in sample
xbars: \( \bar{x}_s \)
ybars: \( \bar{y}_s \)
bhat: \( \hat{B} \)
tSRS: \( \hat{t}_{y,srs} = N \bar{y}_s \)
thatr: \( \hat{t}_{yr} \)

References


Description

Information from a stratified random sample of Fellows of the American Statistical Association elected between 2000 and 2018. The list of Fellows serving as the population was downloaded from amstat on March 18, 2019. All other information was obtained from public sources.

Usage

data(asafellow)

Format

This data frame contains the following columns:

awardyr: year of award
gender: gender of Fellow (character variable, M = male, F = female)
popsizes: population size in stratum ( = \( N_h \))
sampsizes: sample size in stratum ( = \( n_h \))
**field**: field of employment (character variable)
  - acad = academia
  - ind = industry
  - govt = government

**degreeyr**: year in which Fellow received terminal degree (year of Ph.D. if applicable, otherwise year of Master’s or Bachelor’s degree)

**math**: = 1 if majored in mathematics as undergraduate
  - = 0 if did not major in math
  - = NA if missing

References

Description

Selection of accounts for audit data used in Chapter 6 of SDA.

Usage

data(auditselect)

Format

This data frame contains the following columns:

- **account**: audit unit
- **bookval**: book value of account
- **cumbv**: cumulative book value
- **rn1**: random number 1 selecting account
- **rn2**: random number 2 selecting account
- **rn3**: random number 3 selecting account

References


Description

Population and housing unit estimates for Arizona counties, excluding Maricopa and Pima counties, from the American Community Survey 2018 5-year estimates.

Usage

data(azcounties)
baseball

Format

This data frame contains the following columns:

name: county name (character variable, length 15)
number: county number
population: population estimate for county
housing: housing unit estimate for county
ownerocc: number of owner-occupied housing units for county

References


baseball baseball data

Description

Statistics on 797 baseball players, compiled by Jenifer Boshes from the rosters of all major league teams in November 2004. Missing values (for variables pball, intwalk, hbp, sacrfly; all other variables have complete data) are coded as NA.

Usage

data(baseball)

Format

This data frame contains the following columns:

team: team played for at the beginning of the season
leagueid: AL or NL
player: a unique identifier for each baseball player
salary: player salary in 2004
pos: primary position coded as P, C, 1B, 2B, 3B, SS, RF, LF, or CF
gplay: games played
gstart: games started
inning: number of innings
putout: number of putouts
assist: number of assists
error: errors
dplay: number of double plays
pball: number of passed balls (only applies to catchers)
gbat: number of games that player appeared at bat
atbat: number of at bats
run: number of runs scored
hit: number of hits
secbase: number of doubles
thirdbase: number of triples
homerun: number of home runs
rbi: number of runs batted in
stolenb: number of stolen bases
csteal: number of times caught stealing
walk: number of times walked
strikeout: number of strikeouts
intwalk: number of times intentionally walked
hbp: number of times hit by pitch
sacrhit: number of sacrifice hits
sacrfly: number of sacrifice flies
gidplay: grounded into double play

References

<table>
<thead>
<tr>
<th>books</th>
<th>books data</th>
</tr>
</thead>
</table>

Description
Data from homeowner’s survey to estimate total number of books, used in Chapter 5.

Usage
data(books)
Format

This data frame contains the following columns:

- **shelf**: shelf number
- **Mi**: number of books on shelf
- **booknumber**: number of the book selected
- **purchase**: purchase cost of book
- **replace**: replacement cost of book

References


captureci  

Capture-recapture confidence interval function

Description


Usage

captureci(xmat, y, alpha)

Arguments

- **xmat**: Define 1 = in sample and 0 = not in sample. For example, if there are two samples, xmat has two columns; the row (1,0) represents the category of being in sample 1 but not in sample 2.
- **y**: Number of units corresponding to xmat.
- **alpha**: Confidence level with a default value of 0.05.

Value

- **cell**: estimated cell value for the missing count of category (0, 0)
- **N**: the estimated total counts
- **CI_cell**: the estimated confidence interval for the missing category count
- **CI_N**: the estimated confidence interval for total counts

Examples

```r
xmat <- cbind(c(1,1,0),c(1,0,1))
y <- c(20,180,80)
captureci(xmat, y, alpha = 0.1)
```
Description
Population sizes for each state, from the 1920 U.S. census. The data set contains only the 48 states and excludes Washington D.C., Puerto Rico, and U.S. territories (these areas were not allowed to have voting representatives in Congress).

Usage
data(census1920)

Format
This data frame contains the following columns:

- **state**: state name
- **population**: state population in 1920 census

References
Source: U.S. Bureau of the Census (1921).

Description
Population sizes for each state, from the 2010 U.S. census. The data set contains only the 50 states and excludes the areas that, as of 2020, are not allowed to have voting representatives in Congress: Washington D.C., Puerto Rico, and U.S. territories.

Usage
data(census2010)

Format
This data frame contains the following columns:

- **state**: state name
- **population**: state population in 2010 census
cherry

References

Source: U.S. Census Bureau (2019).

---

cherry data

Description

Data for a sample of 31 cherry trees.

Usage

data(cherry)

Format

This data frame contains the following columns:

- **diameter**: diameter of tree (inches)
- **height**: height of tree (feet)
- **volume**: timber volume of tree (cubic feet)

References

Description

Population sizes for 15 classes, used in Chapter 6 of SDA to illustrate unequal-probability sampling.

Usage

data(classes)

Format

This data frame contains the following columns:

- **class**: class ID number
- **class_size**: number of students in class

References


---

Description

Two-stage unequal-probability sample without replacement from the population of classes in classes data.

Usage

data(classpps)

Format

This data frame contains the following columns:

- **class**: class ID number
- **class_size**: number of students in class
- **finalweight**: sampling weight for student
- **hours**: number of hours spent studying statistics
classppsjp

References


classppsjp

classppsjp data

Description

Joint inclusion probabilities for unequal probability sample without replacement from the population of classes in data classes.

Usage

data(classppsjp)

Format

This data frame contains the following columns:

- **class**: class ID number
- **class_size**: number of students in class
- **SelectionProb**: probability of being included in sample, $\pi_i$
- **SamplingWeight**: sampling weight $w_i = 1/(\pi_i)$
- **JtProb_1**: columns of joint inclusion probabilities, $\pi_{1k}$
- **JtProb_2**: columns of joint inclusion probabilities, $\pi_{2k}$
- **JtProb_3**: columns of joint inclusion probabilities, $\pi_{3k}$
- **JtProb_4**: columns of joint inclusion probabilities, $\pi_{4k}$
- **JtProb_5**: columns of joint inclusion probabilities, $\pi_{5k}$

References


college

Description

Selected variables from the U.S. Department of Education College Scorecard Data (version updated on June 1, 2020). Some of the variables in the book data have been calculated from other variables in the original source; these have been given new variable names that are not found in the data dictionary.

Usage

data(college)

Format

This data frame contains the following columns:

- **unitid**: unit identification number
- **instnm**: institution name (character, length 81)
- **city**: city (character, length 24)
- **stabbr**: state abbreviation (character, length 2)
- **highdeg**: highest degree awarded
  - 3 = Bachelor’s degree
  - 4 = Graduate degree
- **control**: control (ownership) of institution
  - 1 = public
  - 2 = private nonprofit
- **region**: region where institution is located
  - 1 New England (CT, ME, MA, NH, RI, VT)
  - 2 Mid East (DE, DC, MD, NJ, NY, PA)
  - 3 Great Lakes (IL, IN, MI, OH, WI)
  - 4 Plains (IA, KS, MN, MO, NE, ND, SD)
  - 5 Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
  - 6 Southwest (AZ, NM, OK, TX)
  - 7 Rocky Mountains (CO, ID, MT, UT, WY)
  - 8 Far West (AK, CA, HI, NV, OR, WA)
- **locale**: locale of institution
  - 11 City: Large (population of 250,000 or more)
  - 12 City: Midsize (population of at least 100,000 but less than 250,000)
  - 13 City: Small (population less than 100,000)
  - 21 Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more)
22 Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000)
23 Suburb: Small (outside principal city, in urbanized area with population less than 100,000)
31 Town: Fringe (in urban cluster up to 10 miles from an urbanized area)
32 Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area)
33 Town: Remote (in urban cluster more than 35 miles from an urbanized area)
41 Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster)
42 Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster)
43 Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster)

ccbasic: carnegie basic classification
15 Doctoral Universities: Very High Research Activity
16 Doctoral Universities: High Research Activity
17 Doctoral/Professional Universities
18 Master’s Colleges & Universities: Larger Programs
19 Master’s Colleges & Universities: Medium Programs
20 Master’s Colleges & Universities: Small Programs
21 Baccalaureate Colleges: Arts & Sciences Focus
22 Baccalaureate Colleges: Diverse Fields

ccsizset: carnegie classification, size and setting
6 Four-year, very small, primarily nonresidential
7 Four-year, very small, primarily residential
8 Four-year, very small, highly residential
9 Four-year, small, primarily nonresidential
10 Four-year, small, primarily residential
11 Four-year, small, highly residential
12 Four-year, medium, primarily nonresidential
13 Four-year, medium, primarily residential
14 Four-year, medium, highly residential
15 Four-year, large, primarily nonresidential
16 Four-year, large, primarily residential
17 Four-year, large, highly residential

hbcu: historically black college or university
1 = yes, 0 = no

openadmp: does the college have an open admissions policy, that is, does it accept any students that apply or have minimal requirements for admission?
1 = yes, 0 = no

adm_rate: fall admissions rate, defined as the number of admitted undergraduates divided by the number of undergraduates who applied
**sat_avg**: average SAT score (or equivalent) for admitted students

**ugds**: number of degree-seeking undergraduate students enrolled in the fall term

**ugds_men**: proportion of ugds who are men

**ugds_women**: proportion of ugds who are women

**ugds_white**: proportion of ugds who are white (based on self-reports)

**ugds_black**: proportion of ugds who are black/African American (based on self-reports)

**ugds_hisp**: proportion of ugds who are Hispanic (based on self-reports)

**ugds_asian**: proportion of ugds who are Asian (based on self-reports)

**ugds_other**: proportion of ugds who have other race/ethnicity (created from other categories on original data file; race/ethnicity proportions sum to 1)

**npt4**: average net price of attendance, derived from the full cost of attendance, including tuition and fees, books and supplies, and living expenses, minus federal, state, and institutional grant scholarship aid, for full time, first time undergraduate Title IV receiving students. NPT4 created from scorecard data variables NPT4_PUB if public institution and NPT4_PRIV if private

**tuitionfee_in**: in-state tuition and fees

**tuitionfee_out**: out-of-state tuition and fees

**avgfacsal**: average faculty salary per month

**pftfac**: proportion of faculty that is full-time

**c150_4**: proportion of first-year, full-time students who complete their degree within 150% of the expected time to complete; for most institutions, this is the proportion of students who receive a degree within 6 years

**grads**: number of graduate students

**Details**

This data set is made available for pedagogical purposes only. Anyone wishing to draw conclusions from College Scorecard data should obtain the full data set from the Department of Education. The original data set has 1,925 variables and includes institutions (such as those that do not grant undergraduate degrees) that are not in the data college.

The college data includes institutions in the original data set that: (1) are located in the 50 states plus District of Columbia, (2) contain information on average net price (NPT4), (3) are predominantly Bachelor’s degree-granting, (4) were currently operating as of June 2020, (5) are not private for-profit institutions or "global" campuses, (6) have Carnegie size classification (variable ccsizset) between 6 and 17 and Carnegie basic classification(variable ccbasic) between 14 and 22 (these offer Bachelor’s degrees), (7) enrolls first-time students, and (8) are not U.S. Service Academies.

For all variables, missing data are coded as NA.

**References**


Description

Five replicate SRSs from the set of public colleges and universities (having control = 1) in colllege data. Columns 1-29 are as in college data, with additional columns 30-32 listed below. Note that the selection probabilities and sampling weights are for the separate replicate samples, so that the weights for each replicate sample sum to the population size 500.

Usage

data(collegerg)

Format

This data frame contains the following columns:

- **unitid**: unit identification number
- **instnm**: institution name (character, length 81)
- **city**: city (character, length 24)
- **stabbr**: state abbreviation (character, length 2)
- **highdeg**: highest degree awarded
  - 3 = Bachelor’s degree
  - 4 = Graduate degree
- **control**: control (ownership) of institution
  - 1 = public
  - 2 = private nonprofit
- **region**: region where institution is located
  - 1 New England (CT, ME, MA, NH, RI, VT)
  - 2 Mid East (DE, DC, MD, NJ, NY, PA)
  - 3 Great Lakes (IL, IN, MI, OH, WI)
  - 4 Plains (IA, KS, MN, MO, NE, ND, SD)
  - 5 Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
  - 6 Southwest (AZ, NM, OK, TX)
  - 7 Rocky Mountains (CO, ID, MT, UT, WY)
  - 8 Far West (AK, CA, HI, NV, OR, WA)
- **locale**: locale of institution
  - 11 City: Large (population of 250,000 or more)
  - 12 City: Midsize (population of at least 100,000 but less than 250,000)
  - 13 City: Small (population less than 100,000)
  - 21 Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more)
22 Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000)
23 Suburb: Small (outside principal city, in urbanized area with population less than 100,000)
31 Town: Fringe (in urban cluster up to 10 miles from an urbanized area)
32 Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area)
33 Town: Remote (in urban cluster more than 35 miles from an urbanized area)
41 Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster)
42 Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster)
43 Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster)

**ccbasic:** carnegie basic classification
15 Doctoral Universities: Very High Research Activity
16 Doctoral Universities: High Research Activity
17 Doctoral/Professional Universities
18 Master’s Colleges & Universities: Larger Programs
19 Master’s Colleges & Universities: Medium Programs
20 Master’s Colleges & Universities: Small Programs
21 Baccalaureate Colleges: Arts & Sciences Focus
22 Baccalaureate Colleges: Diverse Fields

**ccsizset:** carnegie classification, size and setting
6 Four-year, very small, primarily nonresidential
7 Four-year, very small, primarily residential
8 Four-year, very small, highly residential
9 Four-year, small, primarily nonresidential
10 Four-year, small, primarily residential
11 Four-year, small, highly residential
12 Four-year, medium, primarily nonresidential
13 Four-year, medium, primarily residential
14 Four-year, medium, highly residential
15 Four-year, large, primarily nonresidential
16 Four-year, large, primarily residential
17 Four-year, large, highly residential

**hbcu:** historically black college or university,
1 = yes, 0 = no

**openadmp:** does the college have an open admissions policy, that is, does it accept any students that apply or have minimal requirements for admission?
1 = yes, 0 = no

**adm_rate:** fall admissions rate, defined as the number of admitted undergraduates divided by the number of undergraduates who applied
**sat_avg**: average SAT score (or equivalent) for admitted students

**ugds**: number of degree-seeking undergraduate students enrolled in the fall term

**ugds_men**: proportion of ugds who are men

**ugds_women**: proportion of ugds who are women

**ugds_white**: proportion of ugds who are white (based on self-reports)

**ugds_black**: proportion of ugds who are black/African American (based on self-reports)

**ugds_hisp**: proportion of ugds who are Hispanic (based on self-reports)

**ugds_asian**: proportion of ugds who are Asian (based on self-reports)

**ugds_other**: proportion of ugds who have other race/ethnicity (created from other categories on original data file; race/ethnicity proportions sum to 1)

**npt4**: average net price of attendance, derived from the full cost of attendance, including tuition and fees, books and supplies, and living expenses, minus federal, state, and institutional grant scholarship aid, for full time, first time undergraduate Title IV receiving students. NPT4 created from scorecard data variables NPT4_PUB if public institution and NPT4_PRIV if private

**tuitionfee_in**: in-state tuition and fees

**tuitionfee_out**: out-of-state tuition and fees

**avgfacsal**: average faculty salary per month

**pftfac**: proportion of faculty that is full-time

**c150_4**: proportion of first-year, full-time students who complete their degree within 150% of the expected time to complete; for most institutions, this is the proportion of students who receive a degree within 6 years

**grads**: number of graduate students

**selectionprob**: selection probability for each replicate sample

**samplingweight**: sampling weight for each replicate sample

**repgroup**: replicate group number

**References**


---

collshr

collshr data

**Description**

Probability-proportional-to-size sample of size 10 from the stratum of small, highly residential colleges (having `ccsizeset = 11`) in data `college`. Columns 1-29 are as in `college` data, with additional columns 30-34 listed below.
Usage

data(collshr)

Format

This data frame contains the following columns:

- **unitid**: unit identification number
- **instnm**: institution name (character, length 81)
- **city**: city (character, length 24)
- **stabbr**: state abbreviation (character, length 2)
- **highdeg**: highest degree awarded
  - 3 = Bachelor's degree
  - 4 = Graduate degree
- **control**: control (ownership) of institution
  - 1 = public
  - 2 = private nonprofit
- **region**: region where institution is located
  - 1 New England (CT, ME, MA, NH, RI, VT)
  - 2 Mid East (DE, DC, MD, NJ, NY, PA)
  - 3 Great Lakes (IL, IN, MI, OH, WI)
  - 4 Plains (IA, KS, MN, MO, NE, ND, SD)
  - 5 Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
  - 6 Southwest (AZ, NM, OK, TX)
  - 7 Rocky Mountains (CO, ID, MT, UT, WY)
  - 8 Far West (AK, CA, HI, NV, OR, WA)
- **locale**: locale of institution
  - 11 City: Large (population of 250,000 or more)
  - 12 City: Midsize (population of at least 100,000 but less than 250,000)
  - 13 City: Small (population less than 100,000)
  - 21 Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more)
  - 22 Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000)
  - 23 Suburb: Small (outside principal city, in urbanized area with population less than 100,000)
  - 31 Town: Fringe (in urban cluster up to 10 miles from an urbanized area)
  - 32 Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area)
  - 33 Town: Remote (in urban cluster more than 35 miles from an urbanized area)
  - 41 Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster)
  - 42 Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster)
  - 43 Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster)
**ccbasic**: carnegie basic classification
- 15 Doctoral Universities: Very High Research Activity
- 16 Doctoral Universities: High Research Activity
- 17 Doctoral/Professional Universities
- 18 Master’s Colleges & Universities: Larger Programs
- 19 Master’s Colleges & Universities: Medium Programs
- 20 Master’s Colleges & Universities: Small Programs
- 21 Baccalaureate Colleges: Arts & Sciences Focus
- 22 Baccalaureate Colleges: Diverse Fields

**ccsizset**: carnegie classification, size and setting
- 6 Four-year, very small, primarily nonresidential
- 7 Four-year, very small, primarily residential
- 8 Four-year, very small, highly residential
- 9 Four-year, small, primarily nonresidential
- 10 Four-year, small, primarily residential
- 11 Four-year, small, highly residential
- 12 Four-year, medium, primarily nonresidential
- 13 Four-year, medium, primarily residential
- 14 Four-year, medium, highly residential
- 15 Four-year, large, primarily nonresidential
- 16 Four-year, large, primarily residential
- 17 Four-year, large, highly residential

**hbcu**: historically black college or university,
- 1 = yes, 0 = no

**openadmp**: does the college have an open admissions policy, that is, does it accept any students that apply or have minimal requirements for admission?
- 1 = yes, 0 = no

**adm_rate**: fall admissions rate, defined as the number of admitted undergraduates divided by the number of undergraduates who applied

**sat_avg**: average SAT score (or equivalent) for admitted students

**ugds**: number of degree-seeking undergraduate students enrolled in the fall term

**ugds_men**: proportion of ugds who are men

**ugds_women**: proportion of ugds who are women

**ugds_white**: proportion of ugds who are white (based on self-reports)

**ugds_black**: proportion of ugds who are black/African American (based on self-reports)

**ugds_hisp**: proportion of ugds who are Hispanic (based on self-reports)

**ugds_asian**: proportion of ugds who are Asian (based on self-reports)

**ugds_other**: proportion of ugds who have other race/ethnicity (created from other categories on original data file; race/ethnicity proportions sum to 1)
npt4: average net price of attendance, derived from the full cost of attendance, including tuition and fees, books and supplies, and living expenses, minus federal, state, and institutional grant scholarship aid, for full time, first time undergraduate Title IV receiving students. NPT4 created from scorecard data variables NPT4_PUB if public institution and NPT4_PRIV if private

tuitionfee_in: in-state tuition and fees

tuitionfee_out: out-of-state tuition and fees

avgfacsal: average faculty salary per month

pftfac: proportion of faculty that is full-time

c150_4: proportion of first-year, full-time students who complete their degree within 150% of the expected time to complete; for most institutions, this is the proportion of students who receive a degree within 6 years

grads: number of graduate students

mathfac: number of mathematics faculty

psychfac: number of psychology faculty

biolfac: number of biology faculty

psii: selection probability = ugds / (sum of ugds for stratum)

wt: sampling weight = 1/(10*ψ_i)

References


---

coots                coots data

Description

Selected information on egg size, from a larger study by Arnold (1991). Data provided courtesy of Todd Arnold. Not all observations are used for this data set, so results may not agree with those in Arnold (1991).

Usage

data(coots)
Format

This data frame contains the following columns:

- **clutch**: clutch number from which eggs were subsampled
- **csize**: number of eggs in clutch ($M_i$)
- **length**: length of egg (mm)
- **breadth**: maximum breadth of egg (mm)
- **volume**: calculated as $0.000507 \times \text{length} \times \text{breadth}^2$ (mm$^3$)
- **tmt**: = 1 if received supplemental feeding
  = 0 otherwise

References

Boca Raton, FL: CRC Press.

<table>
<thead>
<tr>
<th>counties</th>
<th>counties data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description

Data (from 1990) from an SRS of 100 of the 3141 counties in the United States. Missing values are coded as NA.

Usage

data(counties)

Format

This data frame contains the following columns:

- **RN**: random number used to select the county
- **state**: state abbreviation
- **county**: county name
- **landarea**: land area, 1990 (square miles)
- **totpop**: total number of persons, 1992
- **physician**: active non-Federal physicians on Jan. 1, 1990
- **enroll**: school enrollment in elementary or high school, 1990
- **percpub**: percent of school enrollment in public schools
- **civlabor**: civilian labor force, 1991
unemp: number unemployed, 1991
farmpop: farm population, 1990
numfarm: number of farms, 1987
farmacre: acreage in farms, 1987
fedgrant: total expenditures in federal funds and grants, 1992 (millions of dollars)
fedciv: civilians employed by federal government, 1990
milit: military personnel, 1990
veterans: number of veterans, 1990
percviet: percent of veterans from Vietnam era, 1990

References
Source: U.S. Census Bureau (1994).
Boca Raton, FL: CRC Press.

Description
Data from selected variables in a simple random sample of 5,000 records from the 7,048,107 records with dates between 2001 and 2019 in the City of Chicago database "Crimes-2001 to Present". This file was downloaded on August 11, 2020 from https://data.cityofchicago.org/. These data are provided for pedagogical purposes only. Anyone wishing to publish analyses of Chicago crime data should obtain the most recent data from the website. For a list and map of Community Areas, see https://www.chicago.gov/city/en/depts/dgs/supp_info/citywide_maps.html.

Usage
data(crimes)

Format
This data frame contains the following columns:

year: year in which crime occurred (between 2001 and 2019)
crimetype: type of crime, determined from detailed crime description in database
homicide = homicide
sexualasslt = sexual assault
robbery = robbery
aggasslt = aggravated assault
burglary = burglary
deadtrees = motor vehicle theft
idtheft = identity theft
theft = other type of theft
arson = arson
simpleasslt = simple assault (assaults that are not aggravated)
threat = threat or harassment
fraud = fraud
weapon = weapons violation
trespass = trespassing
vandalism = vandalism
narcotics = narcotics or liquor law violation
other = other

violent:  = 1 if violent crime
          = 0 otherwise

arrest:  = 1 if an arrest was made
          = 0 otherwise

domestic: = 1 if crime was domestic-related as defined by the Illinois Domestic Violence Act
          = 0 otherwise

commarea: number of the Community Area in Chicago where the crime occurred

location: type of location where crime occurred (e.g. street, apartment)

References


---

deadtrees data

Description

Number of dead trees recorded by photograph and field count for a (fictional) SRS of 25 plots taken from a population of 100 plots.

Usage

data(deadtrees)

Format

This data frame contains the following columns:

photo: number of dead trees in plot from photograph
field: number of dead trees in plot from field observation
References


---

divorce data

Description

Data from a sample of divorce records for states in the Divorce Registration Area.

Usage

data(divorce)

Format

This data frame contains the following columns:

- **state**: state name (character variable)
- **abbrev**: state abbreviation (character variable)
- **samprate**: sampling rate for state
- **numrecs**: number of records sampled in state
- **hsblt20**: number of records in sample with husband’s age < 20
- **hsb20to24**: number of records with 20 <= husband’s age <= 24
- **hsb25to29**: number of records with 25 <= husband’s age <= 29
- **hsb30to34**: number of records with 30 <= husband’s age <= 34
- **hsb35to39**: number of records with 35 <= husband’s age <= 39
- **hsb40to44**: number of records with 40 <= husband’s age <= 44
- **hsb45to49**: number of records with 45 <= husband’s age <= 49
- **hsbge50**: number of records with husband’s age => 50
- **wflt20**: number of records with wife’s age < 20
- **wf20to24**: number of records with 20 <= wife’s age <= 24
- **wf25to29**: number of records with 25 <= wife’s age <= 29
- **wf30to34**: number of records with 30 <= wife’s age <= 34
- **wf35to39**: number of records with 35 <= wife’s age <= 39
- **wf40to44**: number of records with 40 <= wife’s age <= 44
- **wf45to49**: number of records with 45 <= wife’s age <= 49
- **wfge50**: number of records with wife’s age => 50
emppmf

**Empirical mass function**

**Description**
Calculates the empirical probability mass function for a variable with associated weights.

**Usage**
```r
emppmf(y, weight)
```

**Arguments**
- `y` : Numerical variable
- `weight` : Associated weights of the variable of interest, default weight is `rep(1,length(y))`

**Value**
- `vals`: the distinct values of `y`
- `epmf`: empirical probability mass function corresponding to each `y` in `vals`

**Examples**
```r
emppmf(seq(1:10))
emppmf(htsrs$height, rep(2000/200,200))
```

---

gini

**Gini data**

**Description**
Data from the population of districts for the 1921 Italian general census.

**Usage**
```r
data(gini)
```
Format

This data frame contains the following columns:

- **id**: ID number
- **district**: district name
- **birth_rate**: births per 1,000 population
- **death_rate**: deaths per 1,000 population
- **marriage_rate**: marriages per 1,000 population
- **agricultural_pop**: percentage of males over 10 years old who work in agriculture
- **urban_population**: percentage of population in urban areas
- **income**: average income
- **altitude**: average altitude above sea level (meters)
- **pop_density**: number of inhabitants per square kilometer
- **natural_growth**: rate of average increase of the population
- **population**: population of area
- **area**: land area (square kilometers)
- **in_GG_sample**: = 1 if in the purposive sample selected by Gini and Galvani
  = 0 otherwise

References


Boca Raton, FL: CRC Press.

data(golfsrs)

Description

A simple random sample of 120 golf courses, taken from the population on the website ww2.golfcourse.com on August 5, 1998. Missing data are denoted by NA.

Usage

data(golfsrs)
**Format**

This data frame contains the following columns:

- **RN**: random number used to select golf course for sample
- **state**: state name
- **holes**: number of holes
- **type**: type of course:
  - priv = private
  - semi = semi-private
  - pub = public
  - mili = military
  - resort = resort
- **yearblt**: year course was built
- **wkday18**: greens fee for 18 holes during week
- **wkday9**: greens fee for 9 holes during week
- **wkend18**: greens fee for 18 holes on weekend
- **wkend9**: greens fee for 9 holes on weekend
- **backtee**: back tee yardage
- **rating**: course rating
- **par**: par for course
- **cart18**: golf cart rental fee for 18 holes
- **cart9**: golf cart rental fee for 9 holes
- **caddy**: are caddies available? (y or n)
- **pro**: is a golf pro available? (y or n)

**References**


data(gpa)

gpa data

**Description**

GPA data from Chapter 5 of SDA.

**Usage**

data(gpa)
### healthjournals

#### Format

This data frame contains the following columns:

- **suite**: suite (psu) identifier
- **gpa**: grade point average of person in suite
- **wt**: sampling weight, = 20 for every observation

#### References


### healthjournals data

#### Description

Randomization and statistical inference practices in a stratified random sample of 196 public health articles. The data, provided courtesy of Dr. Matt Hayat, are discussed in Hayat and Knapp (2017). The variables provided in `healthjournals` are a subset of the variables collected by the authors.

#### Usage

```r
data(healthjournals)
```

#### Format

This data frame contains the following columns:

- **journal**: journal that published the article
  - AJPH = American Journal of Public Health
  - AJPM = American Journal of Preventive Medicine
  - PM = Preventive Medicine
- **NumAuthors**: number of authors
- **RandomSel**: "Yes" if data in the article were from a randomly selected (probability) sample
  - "No" otherwise
- **RandomAssn**: "Yes" if study subjects for the article were randomly assigned to treatment groups
  - "No" otherwise
- **ConfInt**: "Yes" if a confidence interval appeared in the article’s main text, tables, or figures
  - "No" otherwise
- **HypTest**: "Yes" if a p-value or significance test appeared in the article’s main text, tables, or figures
  - "No" otherwise
- **Asterisks**: "Yes" if asterisks were used to represent p-value ranges
  - "No" otherwise
htcdf

References


---

### htcdf

**htcdf data**

| height: | height value, cm |
| frequency: | number of times height value in column 1 occurs in population |
| epmf: | empirical probability mass function |
| ecdf: | empirical distribution function |

**References**


**htpop**

**htpop data**

**Description**

Height and gender of 2,000 persons in an artificial population.

**Usage**

data(htpop)

**Format**

This data frame contains the following columns:

- **height**: height of person, cm
- **gender**: M = male
  - F = female

**References**


---

**htsrs**

**htsrs data**

**Description**

Height and gender for an SRS of 200 persons, taken from *htpop* data

**Usage**

data(htsrs)

**Format**

This data frame contains the following columns:

- **rn**: random number used to select unit
- **height**: height of person, cm
- **gender**: M = male
  - F = female
htstrat

References

htstrat  
htstrat data

Description
Height and gender for a stratified random sample of 160 women and 40 men, taken from htpop data.

Usage
data(htstrat)

Format
The columns and names are as in htsrs data.

References

hunting  
hunting data

Description
Population and sample sizes for the poststrata used for the Sunday hunting survey.

Usage
data(hunting)

Format
This data frame contains the following columns:

region: region of state (East, Central, West)
gender: gender (female, male)
age: age group (16-24, 25-34, 35-44, 45-54, 55-64, 65+)
popsize: population size in poststratum from the 2000 U.S. census
tsampsise: sample size in poststratum
impute
data

Description
Small artificial data set used to illustrate imputation methods. Missing values are denoted by NA.

Usage
data(impute)

Format
This data frame contains the following columns:

person: identification number for person
age: age in years
gender: M = male
F = female
education: number of years of education
crime: = 1 if victim of any crime
= 0 otherwise
violcrime: = 1 if victim of violent crime
= 0 otherwise

References
integerwt

integerwt data

Description

Artificial population of 2000 observations.

Usage

data(integerwt)

Format

This data frame contains the following columns:

* **stratum**: stratum number
* **y**: y value of observation

References


intellonline

intellonline data

Description

Data from the online (Mechanical Turk) survey. The data were downloaded from PLOS ONE on February 8, 2020; the variables extracted from the full data set are provided here for educational purposes only.

Usage

data(intellonline)

Format

This data frame contains the following columns:
int: response to question about agreement with the statement "I am more intelligent than the average person”
   1 = Strongly Agree
   2 = Mostly Agree
   3 = Mostly Disagree
   4 = Strongly Disagree
   5 = Don’t Know or Not Sure

region: census region of respondent (character variable, length 10):
   Northeast
   South
   Midwest
   West

sex: sex (character variable, length 8):
   Male
   Female

race: race (character variable, length 18):
   White
   African American
   Asian American
   Hispanic American
   Another origin

age: age, years

income: household income level (character variable, length 8):
   < $40k,
   $40-80k,
   > $80k

education: highest education level attained (character variable, length 12):
   No College
   Some College
   College Grad
   Grad School
   MISSING

postwt: relative weight, obtained by poststratifying to demographic proportions in the 2010 U.S. Census. The weights are normed so that they sum to 750.

References


Description

Data from the telephone survey studied by Heck et al. (2018). The data were downloaded from here and are provided for educational purposes only.

Usage

data(intelltel)

Format

The variables are the same as in intellonline.

References


Description

Relative weights for demographic groups in intellonline and intelltel (Heck et al., 2018). Each sample was weighted using the 2010 U.S. Census demographics for sex (male, female), age ( < 44, => 44), and race/ethnicity (white, nonwhite). The table entries give the weights for each of these eight demographic groups.

Usage

data(intellwts)
Format

This data frame contains the following columns:

sex: Female and Male
agegroup: Young = (age less than 44)
Old = (age greater than or equal to 44)
race: White or Nonwhite
tel_n: number of telephone survey respondents in the sex/age group/race class
online_n: number of online survey respondents in the sex/agegroup/race class
tel_wgt: relative weight for each respondent to the telephone survey in this sex/agegroup/race class
online_wgt: relative weight for each respondent to the telephone survey in this sex/agegroup/race class

References


intervals_ex40

| intervals_ex40 | Interval estimates using SRS formulae and formulae appropriate for cluster samples |

Description

Simulate a population of clusters, then draw a simple random sample of clusters and construct interval estimates using incorrect SRS formulae and formulae appropriate for cluster samples.

Usage

`intervals_ex40(groupcorr, numintervals, groupsize, sampgroups, popgroups, mu, sigma)`

Arguments

- `groupcorr`: The intracluster correlation coefficient rho
- `numintervals`: Number of samples to be taken from population
- `groupsize`: Number of elements in each population cluster
- `sampgroups`: Number of clusters to be sampled
- `popgroups`: Number of clusters in population
- `mu`: Mean for generating population
- `sigma`: Standard deviation for generating population
Value

SRS_cover_prob: proportion of intervals using SRS formulae that include the true population mean mu

cl_cover_prob: proportion of intervals using cluster sampling formulae that include the true population mean mu

SRS_mean_CI_width: the average width of the interval estimates from SRS

Cluster_mean_CI_width: the average width of the interval estimates from cluster sampling

Replicate: Simulation replicate

srs_lci: lower limit of CI from SRS

srs_uci: upper limit of CI from SRS

clus_lci: lower limit of CI from cluster sampling

clus_uci: upper limit of CI from cluster sampling

scatter plot: first graph shows scatter plot of the last simulated sample

CI plots: second graph shows interval estimates produced for each sample if analyzed as an SRS (with red interval not containing the true parameter), and the third shows the interval estimates produced for each sample when analyzed as a cluster sample.

Examples

# default setting
intervals_ex40(groupcorr = 0, numintervals = 100, groups = 5, sampgroups = 10, popgroups = 5000, mu = 0, sigma = 1)

# change groupcorr and leave others as default setting
intervals_ex40(groupcorr = 0.3)

intervals_ex40(groupcorr = 0.7, numintervals = 100, groups = 5, sampgroups = 10, popgroups = 5000, mu = 0, sigma = 1)

Description

Data extracted from the 1980 Census Integrated Public Use Microdata Series, using the "Small Sample Density" option in the data extract tool, on September 17, 2008. The stratum and psu variables were constructed for use in the book exercises. Data analyses on this file do NOT give valid results for inference to the 1980 U.S. population.

Usage

data(ipums)
Format

This data frame contains the following columns:

stratum: stratum number (1-9)
psu: psu number (1-90)
inctot: total personal income (dollars), topcoded at $75,000
age: age, with range 15-90
sex: 1 = Male
       2 = Female
race: 1 = White
       2 = Black
       3 = American Indian or Alaska Native
       4 = Asian or Pacific Islander
       5 = Other Race
hispanic: 0 = Not Hispanic
          1 = Hispanic
marstat: marital Status:
         1 = Married
         2 = Separated
         3 = Divorced
         4 = Widowed
         5 = Never married/single
ownersh: ownership of housing unit:
         0 = Not Applicable (N/A)
         1 = Owned or being bought
         2 = Rents
yrsusa: number of years a foreign-born person has lived in the U.S.:
        0 = N/A
        1 = 0-5 years
        2 = 6-10 years
        3 = 11-15 years
        4 = 16-20 years
        5 = 21+ years
school: is person in school?
        1 = No, not in school
        2 = Yes, in school
educrec: educational attainment:
        1 = None or preschool
        2 = Grade 1, 2, 3, or 4
        3 = Grade 5, 6, 7, or 8
        4 = Grade 9
Types of sampling used for articles in a sample of journals. Note that columns 2 and 3 do not always sum to column 1; for some articles, the investigators could not determine which type of sampling was used. When working with these data, you may wish to create a fourth column, "indeterminate", which equals column1 - (column2 + column3).

Usage

data(journal)
Format

This data frame contains the following columns:

- **numemp**: number of articles in 1988 that used sampling
- **prob**: number of articles that used probability sampling
- **nonprob**: number of articles that used non-probability sampling

References


---

measles

Description

Roberts et al. (1995) reported on the results of a survey of parents whose children had not been immunized against measles during a recent campaign to immunize all children in the first five years of secondary school. The original data were unavailable; univariate and multivariate summary statistics from these artificial data, however, are consistent with those in the paper. All variables are coded as 1 for yes, 0 for no, and NA for no answer. A parent who refused consent (variable 4) was asked why, with responses in variables 5 through 10. If a response in variables 5 through 10 was checked, it was assigned value 1; otherwise, it was assigned value 0. A parent could give more than one reason for not having the child immunized.

Usage

data(measles)

Format

This data frame contains the following columns:

- **form**: parent received consent form
- **returnf**: parent returned consent form
- **consent**: parent gave consent for measles immunization
- **hadmeas**: child had already had measles
- **previmm**: child had been immunized against measles
- **sideeff**: parent concerned about side effects
- **gp**: parent wanted general practitioner (GP) to give vaccine
mysteries


- **noshot**: child did not want injection
- **notser**: parent thought measles not a serious illness
- **gpadv**: GP advised that vaccine was not needed
- **school**: school attended by child
- **Mitotal**: population size in school
- **mi**: sample size in school

References


---

Data from a stratified random sample of books nominated for the Edgar awards for Best Novel and Best First Novel. The sample was drawn from the population listing of 655 books at [http://theedgars.com/awards/](http://theedgars.com/awards/) on August 14, 2020.

Usage

data(mysteries)

Format

This data frame contains the following columns:

- **stratum**: stratum number, from 1 to 12, computed from the stratification variables in columns 2-4
- **time**: time period in which award was given:
  1 = 1946-1980
  2 = 1981-2000
  3 = 2001-2020
- **category**: award category (character variable, length 16): Best Novel, or Best First Novel
- **winner**: = 1 if book won the award that year
  = 0 if book was nominated but did not win award
- **popsize**: number of population books in stratum (= \( N_h \))
- **sampsize**: number of sampled books in stratum (= \( n_h \))
obtained: = 1 if book was obtained (responded) in original sample
= 2 if book was obtained in phase II subsample of nonrespondents
= 0 if not obtained

p1weight: weight for phase I sample, calculated as \( N_h / n_h \); use for exercises in Chapters 1-11 of SDA

p2weight: final weight for phase II sample; use for exercises in Chapter 12 of SDA and analyses involving variables victims and firearm

genre: genre of book (character variable, length 11).
"private eye" (protagonist is a private detective)
"procedural" (a detailed, step-by-step analysis of how the crime is solved, using the skills of the detective)
"suspense" (the protagonist is at the center of action or is involved in espionage, but is not a professional detective)

historical: = 1 if the main action in the book takes place at least 20 years before the book’s publication date
= 0 if book action is within 20 years of the publication date

urban: = 1 if the main action in the book takes place primarily in urban areas
= 0 otherwise

authorgender: gender of author (character variable, length 1)
= "F" if author is female
= "M" if author is male

fdetect: number of female detectives (or protagonists, if book has no detective) in book

mdetect: number of male detectives (or protagonists, if book has no detective) in book

victims: number of murder victims in book (missing value set to NA if obtained = 0)

firearm: number of murders committed with firearms in book (missing value set to NA if obtained = 0)

References


Description

Selected variables from the 2015-2016 National Health and Nutrition Examination Survey (NHANES). This data set is provided for educational purposes only. Anyone wishing to publish or use results from analyses of NHANES data should obtain the data files directly from the source: Centers for Disease Control and Prevention (2017).
Usage
data(nhanes)

Format
This data frame contains the following columns:

- **sdmvstra**: Pseudo stratum. These are groups of secondary sampling units used for variance estimation on the publicly available data. Pseudo strata and pseudo psus are released instead of the actual strata and psus to protect the confidentiality of respondents’ information. Use sdmvstra as the variable defining the strata.

- **sdmvpsu**: Pseudo-psu. Use sdmvpsu as the primary sampling unit (psu). (There are two pseudo-psus per pseudo-stratum, numbered 1 and 2.)

- **wtint2yr**: Interview weight (use as weight for variables 5-12)

- **wtmec2yr**: Mobile Examination Center weight (use as weight for any analysis involving variables 13-25)

- **ridstatr**: Interview/examination status
  - = 1 if interviewed only
  - = 2 if interviewed and had medical examination

- **ridageyr**: Age in years at screening, from 0 to 80. Anyone with age > 80 years is recorded (top-coded) as 80. No values are missing for this variable.

- **ridagemn**: Age in months at screening (reported only for persons with age 24 months or younger at the time of exam, otherwise missing)

- **riagendr**: = 1 if male
  - = 2 if female (no missing values)

- **ridreth3**: Race/ethnicity code (no missing values)
  - 1 = Mexican American
  - 2 = Other Hispanic
  - 3 = Non-Hispanic White
  - 4 = Non-Hispanic Black
  - 6 = Non-Hispanic Asian
  - 7 = Other Race, Including Multi-Racial

- **dmdeduc2**: Education level of person interviewed (given for adults age 20+only)
  - 1 = Less than 9th grade
  - 2 = 9th to 11th grade (including 12th grade with no diploma)
  - 3 = High school graduate (including GED)
  - 4 = Some college or associate’s degree
  - 5 = College graduate or above
  - 7 = Refused
  - 9 = Don’t know

- **dmdfmsiz**: Total number of people in the family. Values 1-6 indicate the number of people is that number; value 7 indicates 7 or more people in family. No missing values.
**indfmpir**: Ratio of family income to poverty guideline. A value less than 1 indicates the family is below the poverty threshold. Variable indfmpir is a continuous variable where values between 0 and 4.99 indicate the actual poverty ratio. A value of 5 indicates that the ratio of family income to the poverty guideline for that family is 5 or more.

**bmxwt**: Weight (kg)

**bmxht**: Standing height (cm)

**bmbxmi**: Body mass index (kg/m\(^2\)), calculated as \(\frac{\text{bmxwt}}{(\text{bmxht}/100)^2}\)

**bmxwaist**: Waist circumference (cm)

**bmxleg**: Upper leg length (cm)

**bmxarml**: Upper arm length (cm)

**bmxarmc**: Upper arm circumference (cm)

**bmdavsad**: Average sagittal abdominal diameter (SAD, the distance from the small of the back to the upper abdomen), in cm. Calculated by averaging the SAD readings on the person (up to four)

**lbxtc**: Serum total cholesterol (mg/dL)

**bpxpls**: 60-second pulse

**sbp**: Average systolic blood pressure (mm Hg)

**dbp**: Average diastolic blood pressure (mm Hg)

**bpread**: Number of blood pressure readings

**Details**

The data files merged to create *nhanes* can be read directly from the SAS transport files DEMO_I.XPT,BMX_I.XPT,TCHOL_I.XPT,and BPX_I.XPT from the NHANES website. Variables 1-23 have the same names as in the SAS transport files.

The blood pressure variables sbp and dbp were created as follows. In the medical examination, three consecutive blood pressure readings were obtained after participants sat quietly for 5 minutes, and the maximum inflation level was determined. A fourth measurement was conducted for some persons who had an incomplete or interrupted blood pressure reading.

The variables sbp and dbp were calculated by discarding the first blood pressure reading and calculating the average of the remaining valid readings. Note that some of the diastolic blood pressure readings are 0.

In the comma-delimited file nhanes.csv, missing values are denoted by -9. In the R data file, missing values are denoted by NA. Note that some of the codes for variables in the table below also denote missing values; for example, the value 7 for dmdeduc2 indicates "Refused", and these codes for special types of missing values remain in the R data files.

**References**


Data collected in the New York Bight for June 1974 and June 1975. Two of the original strata were combined because of insufficient sample sizes. For variable "catchwt", weights less than 0.5 were recorded as 0.5 kg.

Usage

data(nybight)

Format

This data frame contains the following columns:

- **year**: year of data collection, 1974 or 1975
- **stratum**: stratum membership, based on depth
- **catchnum**: number of fish caught during trawl
- **catchwt**: total weight (kg) of fish caught during trawl
- **numspp**: number of species of fish caught during trawl
- **depth**: depth of station (m)
- **temp**: surface temperature (degrees °C)

Details

Missing values are coded as NA.

References


**otters**

*otters data*

**Description**

Data on number of holts (dens) in Shetland, U.K., used in Kruuk et al. (1989). Data courtesy of Hans Kruuk.

**Usage**

`data(otters)`

**Format**

This data frame contains the following columns:

- **section**: section of coastline
- **habitat**: type of habitat (stratum)
- **holts**: number of holts (dens)

**References**


---

**ozone**

*ozone data*

**Description**

Hourly ozone readings (parts per billion, ppb) from a site in Monterey County, California, for 2018 and 2019. Source: accessed November 19, 2020. Missing values are denoted by NA.

**Usage**

`data(ozone)`
Format

This data frame contains the following columns:

- **year**: year of reading (2018 or 2019)
- **month**: month of reading (1-12)
- **day**: day of reading (1-31)
- **hr0**: ozone reading (ppb) at 0:00 local time
- **hr1**: ozone reading (ppb) at 1:00 local time
- **hr2**: ozone reading (ppb) at 2:00 local time
- **hr3**: ozone reading (ppb) at 3:00 local time
- **hr4**: ozone reading (ppb) at 4:00 local time
- **hr5**: ozone reading (ppb) at 5:00 local time
- **hr6**: ozone reading (ppb) at 6:00 local time
- **hr7**: ozone reading (ppb) at 7:00 local time
- **hr8**: ozone reading (ppb) at 8:00 local time
- **hr9**: ozone reading (ppb) at 9:00 local time
- **hr10**: ozone reading (ppb) at 10:00 local time
- **hr11**: ozone reading (ppb) at 11:00 local time
- **hr12**: ozone reading (ppb) at 12:00 local time
- **hr13**: ozone reading (ppb) at 13:00 local time
- **hr14**: ozone reading (ppb) at 14:00 local time
- **hr15**: ozone reading (ppb) at 15:00 local time
- **hr16**: ozone reading (ppb) at 16:00 local time
- **hr17**: ozone reading (ppb) at 17:00 local time
- **hr18**: ozone reading (ppb) at 18:00 local time
- **hr19**: ozone reading (ppb) at 19:00 local time
- **hr20**: ozone reading (ppb) at 20:00 local time
- **hr21**: ozone reading (ppb) at 21:00 local time
- **hr22**: ozone reading (ppb) at 22:00 local time
- **hr23**: ozone reading (ppb) at 23:00 local time

References


Description

Fictional data from a fictional point-in-time (PIT) survey taken to estimate the number of persons experiencing homelessness.

Usage

data(pitcount)

Format

This data frame contains the following columns:

- **strat**: stratum number (from 1 to 8)
- **division**: geographic division, used to form strata
- **density**: expected density of persons experiencing homelessness (character variable, with values High or Low)
- **popsize**: \(N_h\), the number of areas in the population for stratum \(h\)
- **sampsise**: \(n_h\), the number of areas in the sample for stratum \(h\)
- **areawt**: \(N_h/n_h\), the sampling weight for the area
- **y**: number of persons experiencing unsheltered homelessness found in the area during the PIT count

References


Description

The data described in Zhang et al. (2020) were downloaded from https://www.openicpsr.org/openicpsr/project/109021/version/V1/view on January 22, 2020, from file survey4.rds.

Usage

data(profresp)
Format

This data frame contains the following columns:

**prof_cat:** Level of professionalism
- 1 = novice
- 2 = average
- 3 = professional

**panelnum:** Number of panels respondent has belonged to. A response between 1 and 6 means that the person has belonged to that number of panels; 7 means 7 or more.

**survnum_cat:** How many Internet surveys have you completed before this one?
- 1 = This is my first one
- 2 = 1-5
- 3 = 6-10
- 4 = 11-15
- 5 = 16-20
- 6 = 21-30
- 7 = More than 30

**panelq1:** Are you a member of any online survey panels besides this one?
- 1 = yes
- 2 = no

**panelq2:** To how many other online panels do you belong?
- 1 = None
- 2 = 1 other panel
- 3 = 2 others
- 4 = 3 others
- 5 = 4 others
- 6 = 5 others
- 7 = 6 others or more.

This question has a missing value if panelq1 = 2. If you want to estimate how many panels a respondent belongs to, create a new variable numpanel that equals panelq2 if panelq2 is not missing and equals 1 if panelq1 = 2.

**age4cat:** Age category
- 1 = 18 to 34
- 2 = 35 to 49
- 3 = 50 to 64
- 4 = 65 and over

**edu3cat:** Education category
- 1 = high school or less
- 2 = some college or associates’ degree
- 3 = college graduate or higher

**gender:** 1 = male
- 2 = female
non_white: 1 = race is non-white
          0 = race is white

motive: Which best describes your main reason for joining on-line survey panels?
        1 = I want my voice to be heard
        2 = Completing surveys is fun
        3 = To earn money
        4 = Other (Please specify)

freq_q1: During the PAST 12 MONTHS, how many times have you seen a doctor or other health
         care professional about your own health? Response is number between 0 and 999.

freq_q2: During the PAST MONTH, how many days have you felt you did not get enough rest or
         sleep?

freq_q3: During the PAST MONTH, how many times have you eaten in restaurants? Please in-
         clude both full-service and fast food restaurants.

freq_q4: During the PAST MONTH, how many times have you shopped in a grocery store? If you
         shopped at more than one grocery store on a single trip, please count them separately.

freq_q5: During the PAST 2 YEARS, how many overnight trips have you taken?

Details

The data set profresp contains selected variables from the set of 2,407 respondents who completed
the survey and provided information on the demographic variables and the information needed to
calculate "professional respondent" status. The full data set survey4.rds contains numerous addi-
tional questions about behavior that are not included here, as well as the data from the partially
completed surveys. The website also contains data for three other online panel surveys. Because
profresp is a subset of the full data, statistics calculated from it may differ from those in Zhang et
al. (2020).

Missing values are denoted by NA.

References

Zhang et al. (2020). Professional respondents in opt-in online panels: What do we really know?
Social Science Computer Review 38 (6), 703–719.


Boca Raton, FL: CRC Press.

Description

Population estimates from the 2011 American Community Survey (ACS) for age/gender/education
categories measured in profresp (Zhang et al., 2020). Note that age3cat has 3 categories, while the
age variable in profresp has 4 categories.
Usage

```
data(profrespacs)
```

Format

This data frame contains the following columns:

- **gender**: 1 = male  
  2 = female
- **age3cat**: age category  
  1 = 18 to 34  
  2 = 35 to 64  
  3 = 65 and over
- **edu3cat**: education category  
  1 = high school or less  
  2 = some college or associates’ degree  
  3 = college graduate or higher
- **count**: population size from ACS for the gender/age/education level combination

References

Zhang et al. (2020). Professional respondents in opt-in online panels: What do we really know?  
*Social Science Computer Review* 38 (6), 703–719.


---

**radon**

Description

Radon readings for a stratified sample of 1003 homes in Minnesota. The data were downloaded in April 2008 from an earlier version of the website now located at [https://www.stat.berkeley.edu/users/statlabs/labs.html](https://www.stat.berkeley.edu/users/statlabs/labs.html).

Usage

```
data(radon)
```
Format

This data frame contains the following columns:

- **countyname**: county name
- **countynum**: county number
- **sampsize**: sample size in county
- **popsize**: population size in county
- **radon**: radon concentration (picocuries per liter)

References


---

**rectlength**

**rectlength data**

---

Description

Lengths of rectangles.

Usage

data(rectlength)

Format

This data frame contains the following columns:

- **rectangle**: rectangle number
- **length**: rectangle length

References


Page from a random number table.

data(rnt)

This data frame contains the following columns:

- **col1**: column of 5-digit random numbers
- **col2**: column of 5-digit random numbers
- **col3**: column of 5-digit random numbers
- **col4**: column of 5-digit random numbers
- **col5**: column of 5-digit random numbers
- **col6**: column of 5-digit random numbers

All possible simple random samples that can be generated from the population in Example 2.2 of SDA.

data(sample70)
Format

This data frame contains the following columns:

- **sampnum**: sample number
- **u1**: sampled units in $S$
- **u2**: sampled units in $S$
- **u3**: sampled units in $S$
- **u4**: sampled units in $S$
- **y1**: values of $y_1$ in sample $S$
- **y2**: values of $y_2$ in sample $S$
- **y3**: values of $y_3$ in sample $S$
- **y4**: values of $y_4$ in sample $S$
- **total**: estimated population total

References


---

santacruz data

Description

The number of seedlings in the sampled psus on Santa Cruz Island, California, in 1992 and 1994.

Usage

data(santacruz)

Format

This data frame contains the following columns:

- **tree**: tree number
- **seed92**: number of seedlings in 1992
- **seed94**: number of seedlings in 1994

References

Description

Math and reading test results from a two-stage cluster sample of tenth-grade students. An SRS of 10 schools was selected from the 75 schools in the population, and then 20 students were sampled from each school. These data are fictional but the summary statistics are consistent with those seen in educational studies.

Usage

data(schools)

Format

This data frame contains the following columns:

- **schoolid**: school number (use as cluster variable)
- **gender**: gender of student (character variable, F = female, M = male)
- **math**: score on math test
- **reading**: score on reading test
- **mathlevel**: category level for math test score:
  - 1 if 1 <= math <= 40
  - 2 if 41 <= math
- **readlevel**: category level for reading test score:
  - 1 if 1 <= read <= 32
  - 2 if 33 <= read <= 50
- **Mi**: number of students in school, $M_i$
- **finalwt**: weight for student in sample

References


seals

Description
Data on number of breathing holes found in sampled areas of Svalbard fjords, reconstructed from summary statistics given in Lydersen and Ryg (1991).

Usage
data(seals)

Format
This data frame contains the following columns:
- **zone**: zone number for sampled area
- **holes**: number of breathing holes Imjak found in area

References

shapespop

Description
Population of black and gray squares and circles.

Usage
data(shapespop)

Format
This data frame contains the following columns:
- **ID**: identification number for object
- **shape**: shape of object (square or circle)
- **color**: color of object (gray or black)
- **area**: area of object ($cm^2$)
- **conv**: = 1 if object can be reached through convenience sample
  = 0 otherwise
Description

Two-phase sample of shorebird nests. These are artificial data constructed from summary statistics given in Bart and Earnst (2002).

Usage

data(shorebirds)

Format

This data frame contains the following columns:

- **plot**: plot number
- **rapid**: rapid-method count of number of birds in plot
- **intense**: intensive-method count of number of nests in plot
  
  = NA if the plot is not in the phase II sample

References


Description

Usage
data(sp500)

Format
This data frame contains the following columns:

- **Company**: company name (character variable, length 37)
- **Symbol**: stock symbol (character variable, length 5)
- **MarketCap**: market capitalization, in billions of U.S. dollars
- **StockPrice**: price per share of stock
- **PE_Ratio**: price-to-earnings ratio
- **EPS**: earnings per share

References

Description
Fictional cluster sample of introductory Spanish students.

Usage
data(spanish)
Format

This data frame contains the following columns:

- **class**: class number
- **score**: score on vocabulary test (out of 100)
- **trip**: = 1 if plan a trip to a Spanish-speaking country
  = 0 otherwise

References


---

srs30    srs30 data

Description

An SRS of size 30 taken from an artificial population of size 100.

Usage

data(srs30)

Format

This data frame contains the following columns:

- **y**: value of observation

References


statepop

ssc

Description

SRS of 150 members of the Statistical Society of Canada, downloaded from ssc.ca in August, 2006.

Usage

data(ssc)

Format

This data frame contains the following columns:

- gender: m = male
  f = female
- occupation: a = academic
  i = industry
  n = not determined
- ASA: = 1 if person is member of American Statistical Association
  = 0 otherwise

References


statepop

Description

Data from an unequal-probability sample of 100 counties from the 1994 County and City Data Book (U.S. Census Bureau, 1994). The sample was selected with probability proportional to population.

Usage

data(statepop)
Format

This data frame contains the following columns:

- **county**: county name (character variable, length 14)
- **state**: state name (character variable)
- **landarea**: land area of county, 1990 (square miles)
- **popn**: population of county, 1992
- **phys**: number of physicians, 1990
- **farmpop**: farm population, 1990
- **numfarm**: number of farms, 1987
- **farmacre**: number of acres devoted to farming, 1987
- **veterans**: number of veterans, 1990
- **percviet**: percent of veterans from Vietnam era, 1990
- **psi**: probability of selection
- **wt**: sampling weight, \( = 1/(100\psi_i) \)

References


---

**statepps**

*statepps data*

Description

Number of counties (or county equivalents; Alaska has boroughs, Louisiana has parishes, and some states have independent cities), population estimates for 2019, land area, and water area for the 50 states plus the District of Columbia. Total area for a state can be calculated by summing land area and water area. Source: Population estimates are from U.S. Census Bureau (2019). Land and water areas are from U.S. Census Bureau (2012).

Usage

data(statepps)

Format

This data frame contains the following columns:

- **state**: state name (character variable, length 20)
- **counties**: number of counties or county equivalents
- **pop2019**: population of state, 2019
- **landarea**: land area of state (square kilometers)
- **waterarea**: water area of state (square kilometers)
### swedishlcs

#### References


<table>
<thead>
<tr>
<th>swedishlcs</th>
<th>swedishlcs data</th>
</tr>
</thead>
</table>

#### Description

Data on call attempts from the Swedish Survey of Living Conditions.

#### Usage

```r
data(swedishlcs)
```

#### Format

This data frame contains the following columns:

- **attempt**: call attempt number
- **resprate**: response rate at call attempt (percent)
- **benefits**: relative bias for variable benefits
- **income**: relative bias for variable income
- **employed**: relative bias for variable employed
- **note**: character variable, length 25: notes about data collection

#### Details

The variable attempt takes on values 1-25 for the initial fieldwork period. Values 31-40 denote the follow-up period, and value 45 gives the final estimates. The gaps in the attempt variable allow one to see the separation of the periods on the graph.

#### References


Description

Selected variables from the Survey of Youth in Custody (Beck et al., 1988).

Usage

data(syc)

Format

This data frame contains the following columns:

- **stratum**: stratum number
- **psu**: psu number
  - = facility number for residents in strata 1-5
  - = person number for residents in strata 6-16
- **facility**: facility number
- **facsize**: number of eligible residents in psu
- **finalwt**: final weight
- **randgrp**: random group number
- **age**: age of resident (NA = missing)
- **race**: race of resident
  - 1 = White
  - 2 = Black
  - 3 = Asian/Pacific Islander
  - 4 = American Indian, Alaska Native
  - 5 = Other
  - NA = Missing
- **ethnicity**: 1 = Hispanic
  - 2 = not Hispanic
  - NA = missing
- **educ**: highest grade attended before sent to correctional institution
  - 0 = never attended school
  - 1 - 12 = highest grade attended
  - 13 = GED
  - 14 = other
- **gender**: 1 = male
  - 2 = female
livewith: who did you live with most of the time you were growing up?
   1 = mother only
   2 = father only
   3 = both mother and father
   4 = grandparents
   5 = other relatives
   6 = friends
   7 = foster home
   8 = agency or institution
   9 = someone else
   MA = blank

famtime: has anyone in your family, such as your mother, father, brother, sister, ever served time in jail or prison?
   1 = yes
   2 = no
   NA = don’t know

 crimtype: most serious crime in current offense
   1 = violent (e.g., murder, rape, robbery, assault)
   2 = property (e.g. burglary, larceny, arson, fraud, motor vehicle theft)
   3 = drug (drug possession or trafficking)
   4 = public order (weapons violation, perjury, failure to appear in court)
   5 = juvenile status offense (truancy, running away, incorrigible behavior)
   NA = missing

everviol: ever put on probation or sent to correctional inst for violent offense
   1 = yes
   0 = no

numarr: number of times arrested (NA = missing)

probtn: number of times on probation (NA = missing)

corrinst: number of times previously committed to correctional institution (NA = missing)

evertime: prior to being sent here did you ever serve time in a correctional institution?
   1 = yes
   2 = no
   NA = missing

prviol: 1 if previously arrested for violent offense, 0 otherwise

prprop: 1 if previously arrested for property offense, 0 otherwise

prdrug: 1 if previously arrested for drug offense, 0 otherwise

prpub: 1 if previously arrested for public order offense, 0 otherwise

prjuv: 1 if previously arrested for juvenile status offense, 0 otherwise

agefirst: age first arrested (NA = missing)

usewepn: did you use a weapon . . . for this incident? (1 = yes, 2 = no, NA = blank)
alcuse: did you drink alcohol at all during the year before being sent here this time?
   1 = yes
   2 = no, didn’t drink during year before
   3 = no, don’t drink at all
   NA = missing

everdrug: ever used illegal drugs;
   0 = no
   1 = yes
   NA = missing

Details
Source: U.S. Department of Justice (1989). Strata 6-16 each contain one facility; the psus in those
strata are residents. In strata 1-5, the psus are facilities. The number of facilities in the population
($N_h$) for those five facilities are: $N_1 = 99$, $N_2 = 39$, $N_3 = 30$, $N_4 = 13$, $N_5 = 14$. Eleven
facilities are sampled from stratum 1 and seven facilities are sampled from each of strata 2 through
5.

References
NCJ-113365, Bureau of Justice Statistics, Washington, DC.


Boca Raton, FL: CRC Press.

teachers data

Description
Selected variables from a study on elementary school teacher workload in Maricopa County, Ari-
izona. Data courtesy of Rita Gnap (Gnap, 1995). The psu sizes are given in data teachmi. The large
stratum had 245 schools; the small/medium stratum had 66 schools. Missing values are coded as
NA.

Usage
data(teachers)
Format

This data frame contains the following columns:

- **dist**: school district size, character variable:
  - large
  - sm/me
- **hrwork**: number of hours required to work at school per week
- **size**: class size
- **preprmin**: minutes spent per week in school on preparation
- **assist**: minutes per week that a teacher’s aide works with the teacher in the classroom
- **school**: school identifier

References


teachnr

---

**teachnr data**

**Description**

Data from a follow-up study of nonrespondents from Gnap (1995).

**Usage**

data(teachnr)

**Format**

This data frame contains the following columns:

- **hrwork**: number of hours required to work at school per week
- **size**: class size
- **preprmin**: minutes spent per week in school on preparation
- **assist**: minutes per week that a teacher’s aide works with the teacher in the classroom

**References**


---

uneqvar

---

**uneqvar data**

**Description**

Artificial data used in exercises of Chapter 11.

**Usage**

data(uneqvar)

**Format**

This data frame contains the following columns:

- **x**: x value
- **y**: y value
References


---

vietnam data

Description


Usage

data(vietnam)

Format

This data frame contains the following columns:

- **apc**: APC stratum. character variable with options "Yes," "No," "NotAvail"
- **p2sample**: indicator variable for phase II sample
  - = 1 if in phase II sample
  - = 0 otherwise
- **vietnam**: = 1 if service in Vietnam
  - = 0 if service not in Vietnam
  - = NA if not in phase II sample
- **phase1wt**: weight for phase I sample
- **phase2wt**: conditional weight for phase II sample
  - = (phase I sample size in stratum) / (phase II sample size in stratum)
  - = NA for observations not in phase 2 sample
- **finalwt**: final weight for phase II sample
  - = phase1wt*phase2wt
  - = NA for observations not in phase II sample
- **p1apcsie**: number of observations in the observation’s APC stratum that are in the phase I sample ($n_h$)
- **p2apcsie**: number of observations in the observation’s APC stratum that are in the phase II sample ($m_h$)
References


---

vius data

Description

Selected variables from the 2002 U.S. Vehicle Inventory and Use Survey (VIUS).

Usage

data(vius)

Format

This data frame contains the following columns:

- **stratum**: stratum number (contains all 255 strata)
- **adm_state**: state number
- **state**: state name
- **trucktype**: type of truck, used in stratification
  1. pickups
  2. minivans, other light vans, and sport utility vehicles
  3. light single-unit trucks with gross vehicle weight less than 26,000 pounds
  4. heavy single-unit trucks with gross vehicle weight greater than or equal to 26,000 pounds
  5. truck-tractors
- **tabtrucks**: column of sampling weights
- **bodytype**: body type of vehicle
  01. Pickup
  02. Minivan
  03. Light van other than minivan
  04. Sport utility
  05. Armored
  06. Beverage
  07. Concrete mixer
  08. Concrete pumper
  09. Crane
10. Curtainside
11. Dump
12. Flatbed, stake, platform, etc.
13. Low boy
14. Pole, logging, pulpwood, or pipe
15. Service, utility
16. Service, other
17. Street sweeper
18. Tank, dry bulk
19. Tank, liquids or gases
20. Tow/Wrecker
21. Trash, garbage, or recycling
22. Vacuum
23. Van, basic enclosed
24. Van, insulated non-refrigerated
25. Van, insulated refrigerated
26. Van, open top
27. Van, step, walk-in, or multistop
28. Van, other
99. Other not elsewhere classified

adm_modelyear: model year
01. 2003, 2002
02. 2001
03. 2000
04. 1999
05. 1998
06. 1997
07. 1996
08. 1995
09. 1994
10. 1993
11. 1992
12. 1991
13. 1990
14. 1989
15. 1988
16. 1987
17. Pre-1987

vius_gvw: Gross vehicle weight based on average reported weight
01. Less than 6,001 lbs
02. 6,001 to 8,500 lbs
03. 8,501 to 10,000 lbs
04. 10,001 to 14,000 lbs
05. 14,001 to 16,000 lbs
06. 16,001 to 19,500 lbs
07. 19,501 to 26,000 lbs
08. 26,001 to 33,000 lbs
09. 33,001 to 40,000 lbs
10. 40,001 to 50,000 lbs
11. 50,001 to 60,000 lbs
12. 60,001 to 80,000 lbs
13. 80,001 to 100,000 lbs
14. 100,001 to 130,000 lbs
15. 130,001 lbs. or more

**miles_annl:** number of miles driven during 2002

**miles_life:** number of miles driven since manufactured

**mpg:** miles per gallon averaged during 2002, range from 0.3 to 35, NA denotes not reported or not applicable

**opclass:** operator classification with highest percent

1. Private
2. Motor carrier
3. Owner operator
4. Rental
5. Personal transportation
6. Not applicable (Vehicle not in use)

**opclass_mtr:** percent of miles driven as a motor carrier, NA denotes vehicle not in use

**opclass_own:** percent of miles driven as an owner operator, NA denotes vehicle not in use

**opclass_psl:** percent of miles driven for personal transportation, NA denotes vehicle not in use

**opclass_pvt:** percent of miles driven as private (carry own goods or internal company business only), NA denotes vehicle not in use

**opclass_rnt:** percent of miles driven as rental, NA denotes vehicle not in use

**transmssn:** type of transmission

1. Automatic

**trip_primary:** primary range of operation

1. Off-the-road
2. Less than 50 miles
3. 51 to 100 miles
4. 101 to 200 miles
5. 201 to 500 miles
6. 501 miles or more
7. Not reported
8. Not applicable (Vehicle not in use)
trip0_50: percent of annual miles accounted for with trips 50 miles or less from the home base
trip051_100: percent of annual miles accounted for with trips 51 to 100 miles from the home base
trip101_200: percent of annual miles accounted for with trips 101 to 200 miles from the home base
trip201_500: percent of annual miles accounted for with trips 201 to 500 miles from the home base
trip500more: percent of annual miles accounted for with trips 501 or more miles from home base
adm_make: make of vehicle
01. Chevrolet
02. Chrysler
03. Dodge
04. Ford
05. Freightliner
06. GMC
07. Honda
08. International
09. Isuzu
10. Jeep
11. Kenworth
12. Mack
13. Mazda
14. Mitsubishi
15. Nissan
16. Peterbilt
17. Plymouth
18. Toyota
19. Volvo
20. White
21. Western Star
22. White GMC
23. Other (domestic)
24. Other (foreign)
business: Business in which vehicle was most often used during 2002
01. For-hire transportation or warehousing
02. Vehicle leasing or rental
03. Agriculture, forestry, fishing, or hunting
04. Mining
05. Utilities
06. Construction
07. Manufacturing
08. Wholesale trade
09. Retail trade
10. Information services
11. Waste management, landscaping, or administrative/support services
12. Arts, entertainment, or recreation services
13. Accommodation or food services
14. Other services
NA. Not reported or not applicable

Details

Source: Census:VIUS:2006. The data were downloaded from https://www.census.gov/svsd/www/vius in May, 2006. The website from which the data were downloaded no longer exists, and online information about VIUS may now be found at https://www.bts.gov/vius, which provides a link to the archived 2002 data. The missing value of state for records with adm_state = 42 was recoded to "PA", the state that has code 42. This data set has 98,682 records, which may be too large for some software packages to handle; the file viusca is a smaller data set, with the same columns described below, containing only vehicles from California. The variable descriptions below are taken from the VIUS Data Dictionary. Missing values are coded as NA. For some variables, the value is missing because the question is not applicable or the vehicle is not in use; see the individual variable descriptions. Note that a new VIUS is planned for 2022, with data to be released in 2023; see https://www.bts.gov/vius.

References


viusca

viusca data

Description

The data viusca is a smaller data set from vius with the same columns described below, containing only vehicles from California. The variable descriptions below are taken from the VIUS Data Dictionary.

Usage

data(viusca)
Format

This data frame contains the following columns:

**stratum**: stratum number (contains all 255 strata)

**adm_state**: state number

**state**: state name

**trucktype**: type of truck, used in stratification

1. pickups
2. minivans, other light vans, and sport utility vehicles
3. light single-unit trucks with gross vehicle weight less than 26,000 pounds
4. heavy single-unit trucks with gross vehicle weight greater than or equal to 26,000 pounds
5. truck-tractors

**tabtrucks**: column of sampling weights

**bodytype**: body type of vehicle

01. Pickup
02. Minivan
03. Light van other than minivan
04. Sport utility
05. Armored
06. Beverage
07. Concrete mixer
08. Concrete pumper
09. Crane
10. Curtainside
11. Dump
12. Flatbed, stake, platform, etc.
13. Low boy
14. Pole, logging, pulpwood, or pipe
15. Service, utility
16. Service, other
17. Street sweeper
18. Tank, dry bulk
19. Tank, liquids or gases
20. Tow/Wrecker
21. Trash, garbage, or recycling
22. Vacuum
23. Van, basic enclosed
24. Van, insulated non-refrigerated
25. Van, insulated refrigerated
26. Van, open top
27. Van, step, walk-in, or multistop
28. Van, other
99. Other not elsewhere classified
**adm_modelyear**: model year

01. 2003, 2002
02. 2001
03. 2000
04. 1999
05. 1998
06. 1997
07. 1996
08. 1995
09. 1994
10. 1993
11. 1992
12. 1991
13. 1990
14. 1989
15. 1988
16. 1987
17. Pre-1987

**vius_gvw**: Gross vehicle weight based on average reported weight

01. Less than 6,001 lbs
02. 6,001 to 8,500 lbs
03. 8,501 to 10,000 lbs
04. 10,001 to 14,000 lbs
05. 14,001 to 16,000 lbs
06. 16,001 to 19,500 lbs
07. 19,501 to 26,000 lbs
08. 26,001 to 33,000 lbs
09. 33,001 to 40,000 lbs
10. 40,001 to 50,000 lbs
11. 50,001 to 60,000 lbs
12. 60,001 to 80,000 lbs
13. 80,001 to 100,000 lbs
14. 100,001 to 130,000 lbs
15. 130,001 lbs. or more

**miles_annl**: number of miles driven during 2002

**miles_life**: number of miles driven since manufactured

**mpg**: miles per gallon averaged during 2002, range from 0.3 to 35, NA denotes not reported or not applicable

**opclass**: operator classification with highest percent

1. Private
2. Motor carrier
3. Owner operator
4. Rental
5. Personal transportation
6. Not applicable (Vehicle not in use)

**opclass_mtr**: percent of miles driven as a motor carrier, NA denotes vehicle not in use

**opclass_own**: percent of miles driven as an owner operator, NA denotes vehicle not in use

**opclass_psl**: percent of miles driven for personal transportation, NA denotes vehicle not in use

**opclass_pvt**: percent of miles driven as private (carry own goods or internal company business only), NA denotes vehicle not in use

**opclass_rnt**: percent of miles driven as rental, NA denotes vehicle not in use

**transmssn**: type of transmission
1. Automatic

**trip_primary**: primary range of operation
1. Off-the-road
2. Less than 50 miles
3. 51 to 100 miles
4. 101 to 200 miles
5. 201 to 500 miles
6. 501 miles or more
7. Not reported
8. Not applicable (Vehicle not in use)

**trip0_50**: percent of annual miles accounted for with trips 50 miles or less from the home base

**trip051_100**: percent of annual miles accounted for with trips 51 to 100 miles from the home base

**trip101_200**: percent of annual miles accounted for with trips 101 to 200 miles from the home base

**trip201_500**: percent of annual miles accounted for with trips 201 to 500 miles from the home base

**trip500more**: percent of annual miles accounted for with trips 501 or more miles from home base

**adm_make**: make of vehicle
01. Chevrolet
02. Chrysler
03. Dodge
04. Ford
05. Freightliner
06. GMC
07. Honda
08. International
09. Isuzu
10. Jeep
11. Kenworth
12. Mack
13. Mazda
14. Mitsubishi
15. Nissan
16. Peterbilt
17. Plymouth
18. Toyota
19. Volvo
20. White
21. Western Star
22. White GMC
23. Other (domestic)
24. Other (foreign)

**business:** Business in which vehicle was most often used during 2002

01. For-hire transportation or warehousing
02. Vehicle leasing or rental
03. Agriculture, forestry, fishing, or hunting
04. Mining
05. Utilities
06. Construction
07. Manufacturing
08. Wholesale trade
09. Retail trade
10. Information services
11. Waste management, landscaping, or administrative/support services
12. Arts, entertainment, or recreation services
13. Accommodation or food services
14. Other services
NA. Not reported or not applicable

**References**


Description

Selected variables from the Arizona State University Winter Closure Survey, taken in January 1995 (provided courtesy of the ASU office of University Evaluation). This survey was taken to investigate the attitudes and opinions of university employees towards the closing of the university (for budgetary reasons) between December 25 and January 1. For the yes/no questions, the responses are coded as 1 = No, 2 = Yes. The variables treatsta and treatme are coded as 1 = strongly agree, 2 = agree, 3 = undecided, 4 = disagree, 5 = strongly disagree. The variables process and satbreak are coded as 1 = very satisfied, 2 = satisfied, 3 = undecided, 4 = dissatisfied, 5 = very dissatisfied. Variables ownsupp through offclose are coded 1 if the person checked that the statement applied to him/her, and 2 if the statement was not checked.

Usage

data(winter)

Format

This data frame contains the following columns:

- **class**: Stratum number
  - 1 = faculty
  - 2 = classified staff
  - 3 = administrative staff
  - 4 = academic professional

- **yearasu**: Number of years worked at ASU
  - 1 = 1-2 years
  - 2 = 3-4 years
  - 3 = 5-9 years
  - 4 = 10-14 years
  - 5 = 15 or more years

- **vacation**: In the past, have you usually taken vacation days the entire period between December 25 and January 1?

- **work**: Did you work on campus during Winter Break Closure?

- **havediff**: Did the Winter Break Closure cause you any difficulty/concerns?

- **negaeffe**: Did the Winter Break Closure negatively affect your work productivity?

- **ownsupp**: I was unable to obtain staff support in my department/office

- **othersup**: I was unable to obtain staff support in other departments/offices

- **utility**: I was unable to access computers, copy machine, etc. in my department/office

- **environ**: I was unable to endure environmental conditions, e.g., not properly climatized
uniserve: I was unable to access university services necessary to my work
workelse: I was unable to work on my assignments because I work in another department/office
offclose: I was unable to work on my assignments because my office was closed
treatsta: Compared to other departments/offices, I feel staff in my department/office were treated fairly
treatme: Compared to other people working in my department/office, I feel I was treated fairly
process: How satisfied are you with the process used to inform staff about Winter Break Closure?
satbreak: How satisfied are you with the fact that ASU had a Winter Break Closure this year?
brakeaga: Would you want to have Winter Break Closure again?

Details
Missing values are coded as NA.

References

Description
Hypothetical sample of size 114, with indirect sampling. The data set has multiple records for adults with more than one child; if adult 254 has 3 children, adult 254 is listed 3 times in the data set. Note that to obtain $L_k$, you need to take numadult +1.

Usage
data(wtshare)

Format
This data frame contains the following columns:

id: identification number of adult in sample
child: = 1 if record is for a child
            = 0 if adult has no children
preschool: = 1 if child is in preschool
            = 0 otherwise
numadult: number of other adults in population who link to that child
References


Index

* Source:
ipums, 45
*datasets*
agpop, 3
agpps, 4
agsrs, 6
agstrat, 6
algebra, 7
anthrop, 8
anthsrs, 8
anthuneq, 9
artifratio, 9
asafellow, 10
auditresult, 11
auditselect, 12
azcounties, 12
baseball, 13
books, 14
census1920, 16
census2010, 16
cherry, 17
classes, 18
classpps, 18
classppsjp, 19
college, 20
collegerg, 23
collshr, 25
coots, 28
counties, 29
cri mes, 30
deadtrees, 31
divorce, 32
gini, 33
golfsrs, 34
gpa, 35
healthjournals, 36
htcdf, 37
htpop, 38
htsrs, 38
htstrat, 39
hunting, 39
impute, 40
integerwt, 41
intellonline, 41
intelltel, 43
intellwts, 43
ipums, 45
journal, 47
measles, 48
mysteries, 49
nhanes, 50
nybight, 53
otters, 54
ozone, 54
pitcount, 56
profresp, 56
profrespacs, 58
radon, 59
rectlength, 60
rnt, 61
sample70, 61
santacruz, 62
schools, 63
seals, 64
shapespop, 64
shorebirds, 65
sp500, 66
spanish, 66
srs30, 67
ssc, 68
statepop, 68
statepps, 69
swedishlcs, 70
syc, 71
teachers, 73
teachmi, 74
teachnr, 75
uneqvar, 75
vietnam, 76
vius, 77
viusca, 81
winter, 86
wtshare, 87
agpop, 3
agpps, 4
agsrs, 6
agstrat, 6
algebra, 7
anthrop, 8
anthsrs, 8
anthuneq, 9
artifratio, 9
asafellow, 10
auditresult, 11
auditselect, 12
azcounties, 12
baseball, 13
books, 14
captureci, 15
census1920, 16
census2010, 16
cherry, 17
classes, 18
classpps, 18
classppsjp, 19
college, 20
collegerg, 23
collshr, 25
coots, 28
counties, 29
crimes, 30
deadtrees, 31
divorce, 32
emppmf, 33
gini, 33
golfhrs, 34
gpa, 35
healthjournals, 36
htcdf, 37
htpop, 38
htsrs, 38
htstrat, 39
hunting, 39
impute, 40
integerwt, 41
intellonline, 41
intelltel, 43
intellwts, 43
intervals_ex40, 44
ipums, 45
journal, 47
measles, 48
mysteries, 49
nhanes, 50
nybight, 53
otters, 54
ozone, 54
pitcount, 56
profresp, 56
profrespsacs, 58
radon, 59
rectlength, 60
rnt, 61
sample70, 61
santacruz, 62
schools, 63
seals, 64
shapespop, 64
shorebirds, 65
sp500, 66
spanish, 66
srs30, 67
ssc, 68
statepop, 68
statepps, 69
swedishlcs, 70
syc, 71
teachers, 73
teachmi, 74
teachnr, 75
uneqvvar, 75
INDEX

vietnam, 76
vius, 77
viusca, 81

winter, 86
wtshare, 87